Cognitive Psychology 159 (2025) 101744

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Retrieving past experiences to inform novel decisions through a process of cascading episodic sampling

Achiel Fenneman [a,*], Sarah T. Malamut [b], Alan G. Sanfey [c,d]

[a] Department of Cognition, Emotion, and Methods in Psychology, University of Vienna, Austria
[b] INVEST Research Flagship, Department of Psychology, University of Turku, Finland
[c] Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands
[d] Behavioural Science Institute, Radboud University Nijmegen, The Netherlands

ARTICLE INFO

Keywords:
Episodic memory
Decision
Sampling
Computational model

ABSTRACT

We can guide our decisions in novel situations by drawing on our past experiences (episodic memories). While at times we can retrieve relevant episodes via cued recall, other situations may require a process of memory search. But what mechanisms underlie this search? In this work we synthesize six key principles concerning the storage and retrieval of episodic memories, and build on these principles to propose a cognitive mechanism which allows for the retrieval of relevant past experiences through a process of cascading recall. In this process, observing a stimulus triggers the cued recall of a past event. If this memory does not provide sufficient information to warrant a decision, then it next reinstates all the memory's constituent features. These features then form the inputs to sample an additional memory in a subsequent recall step, which in turn reinstates its own features and so forth. This process continues until a suitable past experience is retrieved. We provide empirical support for key predictions of this cascading process through three online experiments in which participants interacted with unfamiliar stimuli. The results indicate that participants rely on cued recall of similar past experiences (experiment 1), and on indirectly related experiences when cued recall is not informative (experiment 2). Additionally, participants were substantially more likely to retrieve a predicted memory, and did so faster, when relying on cued recall versus cascadizng memory search (experiment 3). We conclude by discussing how this cascading recall process bridges several influential models of memory-based decision-making, as well as offering promising directions for future research.

Many of our decisions in everyday life are guided by our past experiences. In familiar situations this guidance is often facilitated through cued recall, whereby our observation of the present situation facilitates the retrieval of past experiences with similar elements. As an illustration, *imagine you go out to dine at an Ethiopian restaurant. Perhaps you have previously visited this particular restaurant once before, and reading the menu triggers a (cued) recall of this experience. As your memory of the previous dish at this restaurant is a positive one, you decide to order the same meal again.*

But what about unfamiliar situations, where cued recall only brings a limited number of memories to mind? Such a limited availability of past events is a common occurrence in everyday life, which is simultaneously rich in complexity yet sparse in prior experience (Gershman & Daw, 2017). If cued recall only brings a limited number of experiences to mind, then these memories may not

* Corresponding author.
E-mail address: achiel.fenneman@univie.ac.at (A. Fenneman).

https://doi.org/10.1016/j.cogpsych.2025.101744
Received 4 June 2024; Received in revised form 27 May 2025; Accepted 29 May 2025
Available online 9 June 2025

yield sufficient information to guide your behavior. Your cued memories may exclusively provide information on what *not* to do (*maybe you visited this restaurant once before, but didn't like the dish you ordered*), or they may support an action which is not currently available (*you've been to a different Ethiopian restaurant in the past, but the dish that you have ordered at the previous restaurant is not listed on the current menu*). Instead of cued recall, these situations require us to search our store of (non-cued) memories to retrieve additional relevant past experiences (Kahana, 2020).

One potential mnemonic strategy may be to draw on indirect (transitive) associations. Imagine that the observation of the current situation (event A) only results in the cued recall of a single past experience (event B), and that this past experience is not informative in the current situation. In such a situation, a second-best past retrieval strategy may be to search for a second memory (event C) which is similar to the cued memory (B) even if it has not direct relationship to your current situation (A). Returning to our example: *reading the menu at the restaurant (event A) reminds you of your previous trip to this restaurant, where you did not enjoy your meal (event B, cued recall). In turn, recalling this past meal brings to mind the dish that your partner ordered during this last visit. While you did not taste this dish, it reminds you of a similar-looking meal you recently enjoyed at a different restaurant (C). You therefore decide to order this other dish.* Such transitive associations have previously been documented in low-dimensional and/or repetitive task environments (Gilboa et al., 2019; Howard et al., 2009; Shohamy & Wagner, 2008; Wimmer & Shohamy, 2012; Zeithamova, Dominick, et al., 2012). But does this strategy also apply to the retrieval of one-shot, high-dimensional (rich) yet low-repetitive (sparse) episodic experiences? And if so – which mechanisms facilitate and guide such indirect retrieval?

In this work we address this question through a first-principles approach. We begin by summarizing previous findings on the functioning of the episodic memory system across six key principles. We then demonstrate that these principles provide a mechanism to support the retrieval of episodic memories via a cascading recall process. At the core of this recall process lies an iterative application of cue-based recall. In this iterative process, the observation of a stimulus first triggers the cued recall of a memory. If this cued memory does not provide sufficient information to guide our behavior, then its constituent features form the input for a second (cascading) recall process. If this memory also does not yield sufficient information to guide a decision, then its features provide the input for an additional recall process, and so on.

This cascading recall process yields three empirical predictions. When faced with a novel decision context, we selectively retrieve episodic memories which share an overlap to the observed context (prediction 1). If these cued memories do not yield sufficient information to warrant a decision, then they reinstate their constituent features – which in turn constitutes the input for a subsequent retrieval step. Any memories retrieved during these subsequent ('cascading') recall steps are therefore predicted to be similar to the previously sampled episode, but need not be similar to the observed stimulus itself (prediction 2). All else being equal, such additional recall steps require more time and lead to more stochastic behavior than actions based on cued recall alone (prediction 3).

We provide empirical evidence to support these predictions through three naturalistic experiments. In all experiments participants first interacted with a set of four unique fantasy animals ('*Fennimals*'). Participants were then presented with a set of four different Fennimals, each of which shared an overlapping element with exactly one previously encountered Fennimal. By systematically varying the features of each new Fennimal, we provide behavioral evidence that participants relied on cued recall if a directly similar past experience was sufficiently informative. However, if this cued memory did not yield sufficient information, then participants next drew on cascading recall – and such cascading recall required more time and yielded more stochastic outcomes than decisions based on cued recall.

We then demonstrate how this cascading recall process both bridges and extends various (currently disjointed) approaches to memory-based decision-making. The first step of this cascading recall process mimics the predictions of similarity-based evaluative models. In line with previous work on limited sampling, this cascading recall process does not evaluate all past episodes, but instead retrieves a limited number of past experiences. Just as in associative-context models, the retrieval of an event B enhances the probability of retrieving a related memory C – even if *A* and *C* are not similar to each other. Finally, the cascading recall process operates by drawing upon similar foundations, and extends the scope of, previous models centered around hippocampal generalization. We conclude by arguing that this cascading recall process offers new insights for various topics in memory-based decision-making.

1. Principles underlying the retrieval of past experiences via a process of cascading episodic sampling

As a starting point, we build on recent insights in the neuroscience of memory and synthesize this research into six key principles that describe the functioning of the episodic memory system during the sampling of past experiences. These principles pertain to: (1) the storage of singular episodic memories, (2) the integrative encoding of multiple overlapping memories, (3) the retrieval of a limited number of past episodes to guide behavior, (4) the competitive recall of memories, (5) the process of pattern-completion that guides the retrieval of non-cued experiences and (6) a temporary suppression of memories after they have been retrieved.

1.1. Principle 1: episodic memories are stored as conjunctions of features

Episodic memories are stored in the medial temporal lobe (Squire et al., 2004; Tulving, 2002). Of particular interest is the hippocampus, a sub-region which functions as a multi-modal convergence hub (Backus et al., 2016) and which stores each episodic memory as a distinct (i.e., pattern-separated) representation (Chadwick et al., 2011; Leutgeb et al., 2007; Yassa & Stark, 2011) in the form of engrams or traces (Dudai, 2012; Josselyn et al., 2015; Tonegawa et al., 2015). Within the wider medial temporal lobe, memories are stored as a set of references (Teyler & Rudy, 2007); that is, each episode is stored as a set of weighted connections between the elemental features that make up the memory, with the strength of the connections determined by the salience of that feature within the memory. This representation of memories as the conjunctions of their components allows the hippocampus to assign value to configurations of elemental cues (Duncan et al., 2018; Honey et al., 2014; Iordanova et al., 2008, 2011; Rudy & Sutherland, 1995; Sutherland & Rudy, 1989) by combining the what, where, and when of an episodic memory (Eichenbaum, 2017a; Ergorul & Eichenbaum, 2004; Horner & Doeller, 2017; Komorowski et al., 2009). In conclusion: the hippocampus stores episodic memories as the conjunction of features, which are in turn represented in the wider medial temporal lobe (Fig. 1**A**).

1.2. Principle 2: multiple overlapping episodic memories are integrated into an associative network

The next principle relates to the storage of a system of multiple episodic memories. Neuroscientific studies on the organization of such systems reveal that features which are shared between memories are consistently encoded in the same neural nodes across experiences (McKenzie et al., 2013, 2014). Consequentially, episodic memories with shared features are interconnected via their overlapping neural populations (Chadwick et al., 2011; McKenzie et al., 2014; Schlichting & Preston, 2015; Shohamy & Wagner, 2008), such that two (or more) unrelated episodes may be linked via a mutually related third episode (Garvert et al., 2017; Schlichting & Preston, 2015; Zeithamova, Dominick, et al., 2012). This interconnected coding scheme plays an important part in the recollection of memories. Events which are well-connected with other memories have a higher probability of being recalled (Horner & Burgess, 2013, 2014; Horner & Doeller, 2017), allowing the recombination of previously encountered event details when encountering new experiences (Shohamy & Daw, 2015; Zeithamova, Dominick, & Preston, 2012; Zeithamova, Schlichting, & Preston, 2012; . In turn, such flexible retrieval has been argued to allow the hippocampus to generalize knowledge across related episodes (Barron et al., 2013; Shohamy & Wagner, 2008). In conclusion: multiple episodic memories are stored as an associative network, such that memories which

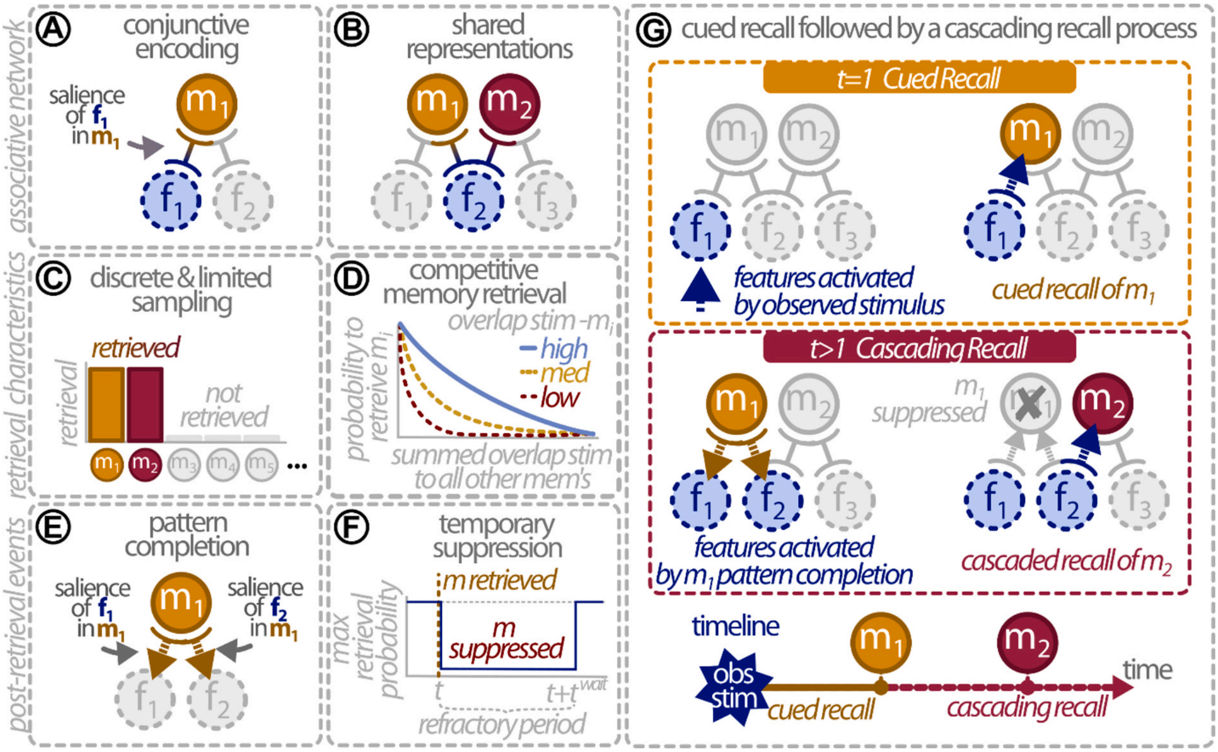

Fig. 1. The Cascading Recall of Episodic Memories. Synthetizing previous insights from memory neuroscience forms the basis for a cascading retrieval process. **(A)**: episodic memories are stored in the medial temporal lobe as conjunctions of features, such that each memory (*m*) is stored as a link of references to its constituent features (*f*). The weights between a memory and its features represent the salience of the feature within the memory. **(B):** a system of multiple memories is stored as an associative network, such that memories with overlapping elements share the same underlying (feature-based) representations. **(C)**: during memory-based decisions, we draw on a limited number of discretely recalled experiences. **(D)**: memory retrieval is a competitive process; the probability of a memory being recalled increases when it has a larger overlap to the stimulus (lines), but decreases with the summed overlap of the stimulus with all other memories (x-axis). **(E)**: the recollection of a memory triggers a subsequent process of pattern-completion, resulting in the reinstatement of all its constituent features. **(F)**: after a memory is retrieved, it is temporarily suppressed for further retrieval. **(G)**: together, these four principles form the basis for the cascading sampling of past experiences. This process starts with the observation of a stimulus, which (stochastically) results in the cued recall of a past experience. If this cued memory is not sufficiently informative to warrant a decision, then additional memories are sampled through sequential steps of cascading recall. Each step of which starts with the reinstatement of the constituent features of the previously retrieved episode, after which this previously retrieved episode is suppressed for the duration of the recall process. These activated features next trigger the retrieval of a subsequent episode, which triggers the reinstatements of its constituent features, and so forth. The resulting cascading process retrieves a consecutive series of related experiences as the sampling process unfolds.

overlapping features share the same underlying representations in the medial temporal lobe (Fig. 1**B**).

1.3. Principle 3: only a limited number of episodes are retrieved for any one decision

While the first two principles describe the organizational structure of episodic memories in the medial temporal lobe, they do not provide details on the retrieval of memories to guide decisions. Previous work suggests that our decisions are oftentimes not based on diffuse and gradually learned associations, but are instead informed by a small number of discretely sampled memories (Erev et al., 2023). Experimental observations in both value-based (Bornstein et al., 2017; Bornstein & Norman, 2017; Mack & Preston, 2016; Wimmer & Büchel, 2016) and perceptual (Gershman et al., 2012) decision-making have provided further support for such a reliance on a limited number of discretely sampled past experiences. Moreover, the limited sampling of past experiences has previously been proposed as a mechanism explaining a wide range of observed behavioral regularities (Erev et al., 2023; Stewart et al., 2006), a further discussion of which follows in the discussion. Summarizing this literature: decisions from memory are based on a limited sample of our past experiences (Fig. 1**C**).

1.4. Principle 4: memory retrieval is competitive

Cued recall, that is, the retrieval of previously stored memory following the presentation of (some of the) elements contained in memory, has been a central topic in the study of memory. Such cued recall is consistent with the functional organization of the hippocampal region, which receives its primary input from feature-based cells located in the entorhinal cortex (Squire et al., 2004), while damage to the hippocampal region impedes our ability to draw on specific past events (Eldridge et al., 2000). However, the retrieval of a memory does not occur in isolation. In the context of recalled words, a considerable body of research in cognitive psychology has demonstrated that cued recall is subject to a competitive process. Memories are more likely to be recalled if they have a greater overlap with the observed cue, but are less likely to be recalled when there are more other memories which have a (stronger) overlap to the observed stimulus (Anderson & Neely, 1996; Tulving & Pearlstone, 1966; Yassa & Reagh, 2013). Summarizing this literature: the probability of a memory being recalled increases when it shares more features with the stimulus, but decreases both when the stimulus has a stronger overlap to other memories, and/or when there are additional memories with an equally strong overlap to the stimulus (Fig. 1**D**).

1.5. Principle 5: upon the recollection of an episodic memory, a pattern-completion process reinstates all its constituent features

The last two principles describe the events that unfold after a memory has been recalled. Previous research provides ample evidence for the involvement of the medial temporal lobe in general, and the hippocampus in particular, in a mnemonic pattern completion process (Marr, 1971; Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; Rolls, 2013, 2016). Upon activation, the hippocampus reinstates the contents of the recalled memory across the medial temporal lobe via a process of scene construction (Hassabis et al., 2007; Hassabis & Maguire, 2007, 2009; Robin & Moscovitch, 2017). Such reinstatement of context has been documented across modalities, with supporting evidence for both location-based context in spatial navigation (Miller et al., 2013) as well as the reinstatement of temporal context in episodic recall (Dragoi & Buzsáki, 2006; Manning et al., 2011). Crucially, the reinstatement of memories in the hippocampus results in a holistic reinstatement of the memory's content in the perirhinal and the parahippocampal cortices (Horner et al., 2015; Mack & Preston, 2016), two regions in the medial temporal lobe which implement a predominantly feature-based coding scheme (Ballard et al., 2019). To summarize: the reinstatement of a particular memory in the hippocampus results in the activation of all its constituent features (Fig. 1**E**).

1.6. Principle 6: once recalled, episodes are temporarily suppressed for subsequent samples

Individual items are recalled at most once per 'recall run'. That is, when performing a single cued-recall effort, the erroneous repetitions of previously observed items is relatively infrequent, especially for back-to-back (immediate) repetitions (Henson, 1998b). Such suppression appears to be a binary inhibition, instead of gradual decay (Duncan & Lewandowsky, 2005). Notably, such repetition-inhibition occurs even when the task demands repeated recall (Henson, 1998a), suggesting that such suppression is automatic (Farrell & Lewandowsky, 2002). To summarize: when drawing on episodic memories, we sample a limited number of unique (that is, non-repeating) and discretely recalled past experiences (Fig. 1**F**).

2. The combination of the above principles results in a process of cascading episodic sampling

The interplay between these six guiding principles provides a ready-made mechanism to retrieve (indirect) past experiences via a cascading recall process. Fig. 1**G** provides a graph-based representation of this proposed mechanism, whilst the following text contains a matrix-based description.

2.1. The formal structure of the model

The memory-based decision process starts when the decision-maker encounters a stimulus. We will refer to this observed stimulus as ω, which consists of present or absent features: $\omega = [f_1, \ldots, f_F]$. The decision-maker next draws on her memories to select one action from a set of potential actions. We describe this sampling process in discrete time steps, and represent the episodic memory system as consisting of three layers: features, memories and action-outcomes.

Each episode is represented as a single node in the memory layer. Based on the conjunctive encoding of episodes (principle 1), each memory m is a vector representing the memory's constituent features: $m = [f_1, \ldots, f_F]$. In this vector, each value $f_i \in [0,1]$ represents the relative salience of a feature within the memory. Any feature f_i may either be entirely absent from the memory (corresponding to $f_i = 0$), fully determine the memory (corresponding to $f_i = 1$), or can have any value in-between. Moreover, we assume that the relative salience values are normalized, such that the total relative salience of each memory's features sum to one.

No restrictions are placed on the contents or modality of features, and each feature may represent any modality, object, place, or time. The only constraint is the consisting encoding following from principle 2: if two or more memories contain the same features, then they share a connection to the same nodes on the feature-level. Consequentially, we can represent a network of multiple episodic memories as a matrix $\boldsymbol{M}$. This matrix has a unique row for each episodic memory, a unique column for each previously observed feature, and its value $\boldsymbol{M}_{m,f}$ denotes the salience of feature f in memory m.

In addition, each memory may (but does not have to) be associated to an action. We represent these action-outcomes with a matrix $\boldsymbol{A}$, which has one row for each possible action (including those actions which may not be presently available), one column for each individual memory. If the action is available for the current decision, then $\boldsymbol{A}_{a,m}$ denotes the remembered outcome of taking action a in memory m. If this action is not available in the current situation, then $\boldsymbol{A}_{a,m} = 0$. As the aim of the this work is to describe the cascading retrieval of memories, we do not make claims regarding the processing, evaluation and integration of ambiguous or noisy signals. In the remainder of this work we therefore assume that these outcomes are all equal and positive (all values in A have either a value of 0 or 1), and the reward of any past action is sufficient to warrant selection of this action in the current situation. The implications of relaxing this assumption are discussed at the end of this work.

The dynamic state of the network is represented as two vectors: φ and μ. The level of activation for all features is denoted as vector φ, which contains one element for each feature such that $\varphi_f^t \in [0,1]$ denotes the degree of activation of feature f at time t. Analogously, a vector μ denotes the state of all memories. In line with principle 3, the recall of an episodic memory is a binary affair; it can either be recalled or not recalled. Therefore, μ contains one element for each memory, such that $\mu_m^t \in \{0,1\}$ denotes whether or not a memory m is retrieved at time t. At the start of the decision process, both vectors are at rest: no features are active and no memories are recalled.

2.2. At first, a memory is retrieved via a process of cued recall

Each decision process starts with activation of the features observed in the stimulus ω:

$$\varphi^{t=1} = \omega$$

The activation of these features may result in the cued retrieval of a past experience. We formalize this recall process with a two-step approach. First, memories are selected to be candidates for retrieval. This is represented with a vector C, which has one element for each memory and $C_m \in \{0, 1\}$ determines whether memory m is a candidate for retrieval (corresponding to a value of 1) or not (0). Candidates are selected through a stochastic process, where the probability that a memory is a candidate for retrieval is given by equation (1):

$$p\left(C_i^{t=1} = 1\right) = \sqrt{m_i\omega} \tag{1}$$

Here $m_i\varphi^t \in [0,1]$ represents the salience-weighted activation of a memory i. Since all features in m_i have a maximum value of 1, we include the square root to re-normalize the recall probabilities. These probabilities are then converted into a vector of binary values (containing at most as single 1, with the rest being 0) through a random draw.

In line with principle 4, the retrieval of memories is competitive: a memory is more likely to be retrieved when it receives a stronger (salience-weighted) input signal from the features-layer, and a decreased probability of recall when other memories receive a (stronger) salience-weighted input from their constituent features. Thus, if there is at least one candidate, then each candidate's probability of being selected depends on the relative strength of its activation as compared to the strength of activation of all candidates. This competition is represented with a softmax function, in which each candidate's probability of being selected for recall is based on the strength of its own activation, divided by the summed activation received by all candidates:

$$P\left(\mu_m^t = 1\right) = \frac{\boldsymbol{M}\varphi C}{sum(\boldsymbol{M}\varphi C)}$$

For notational simplicity, we will refer to this two-step retrieval process (candidate selection and cross-candidate competition) as a single retrieval function α:

$$\mu^t = \alpha(\boldsymbol{M}\varphi^t)$$

This retrieval function returns a vector with a single element for each memory. If a memory m is recalled at time t, then the corresponding value of μ_m^t is equal to 1. If it is not recalled, then the corresponding value is equal to zero. Thus, $\alpha(\mathbf{M}\varphi^t)$ yields the activation of at most a single, discretely recalled, memory.

This formalization allows us to characterize the first step of the sampling process, which is triggered by observing the stimulus ω. At first, this stimulus triggers the cued recall of past episodes, that is, memories with overlapping features as captured by equation (2):

$$\textit{Cued recall}: \mu^{t=1} = \alpha(\boldsymbol{M}\varphi^{t=1}) = \alpha(\boldsymbol{M}\omega) \tag{2}$$

2.3. If a recalled memory supports an action, take it. If it does not, reinstate features and draw a different memory

The recall of a memory results in the retrieval of the action-outcome previously associated to this experience. This is obtained by multiplying the action-outcome matrix **A** with the vector representing which memory is retrieved during this step: $\boldsymbol{A}\mu^t$. If the recalled memory was previously positively associated to an action, then this action is selected and the retrieval process stops. Alternatively, if this recalled memory does not support any of the available actions (that is, if all the values in $\boldsymbol{A}\mu^t$ are equal to zero), then the sampling process continues with one or multiple cascading recall steps.

At the start of each cascading recall step, the pattern-completing process outlined in principle 5 results in the reinstatement of all the constituent features of the memory which was retrieved in the previous recall step. The degree to which each feature is activated is proportional to its salience within the previously recalled memory, and therefore the proportional activation of each feature can be obtained by the transposition of matrix **M** (denoted as $\mathbf{M}^{\mathrm{T}}$):

$$\varphi^{t>1} = \boldsymbol{M}^T\mu^{t-1}$$

These newly activated features then form the input for the next memory retrieval step. These subsequent retrieval steps are similar to the first (cued-)recall step, save for one exception: in line with principle 6, we impose a refractory period on the recall of memories.

In particular, we assume that this refractory period is sufficiently long such that a memory can be sampled at most once during a single search process. This refractory period prevents the cascading process from becoming 'stuck' in a cyclical recall loop, whereby the retrieval of a memory A triggers the retrieval of B, which in turn triggers the retrieval of A, and so forth. Such cyclical recall is both computationally inefficient (it re-samples memories which have already been deemed uninformative to the current situation), and stands in contrast to observed behavior in recall tasks (experimental participants do not appear to frequently re-sample the same sequence of memories). We capture this refractory suppression by assuming that once a memory has been retrieved during the recall process, it can no longer become a candidate for retrieval in equation (3):

$$p\left(C_i^{t>1} = 1\right) = \begin{cases} 0, & \textit{if } m_i \textit{ was previously sampled}; \\ \sqrt{m_i\varphi^t}, & \textit{otherwise} \end{cases} \tag{3}$$

These probabilities are then transformed into a vector of Boolean values through a random draw. After candidates are selected, a competitive soft-max process determines the selection of, at most, a single memory for retrieval. The retrieval of a memory in each cascading step therefore depends on the features of the previously recalled memory. We capture such cascading recall with equation (4):

$$\textit{Cascading recall}: \mu^{t>1} = \alpha(\boldsymbol{M}\varphi^t) = \alpha\left(\boldsymbol{M}\boldsymbol{M}^T\mu^{t-1}\right) \tag{4}$$

As with cued recall, any memories reinstated through the cascading recall process result in the retrieval of the action (if any) which was previously associated to this memory (obtained by $\boldsymbol{A}\mu^t$). As before, if the memory positively supports this action, then the action is selected and the retrieval process stops.

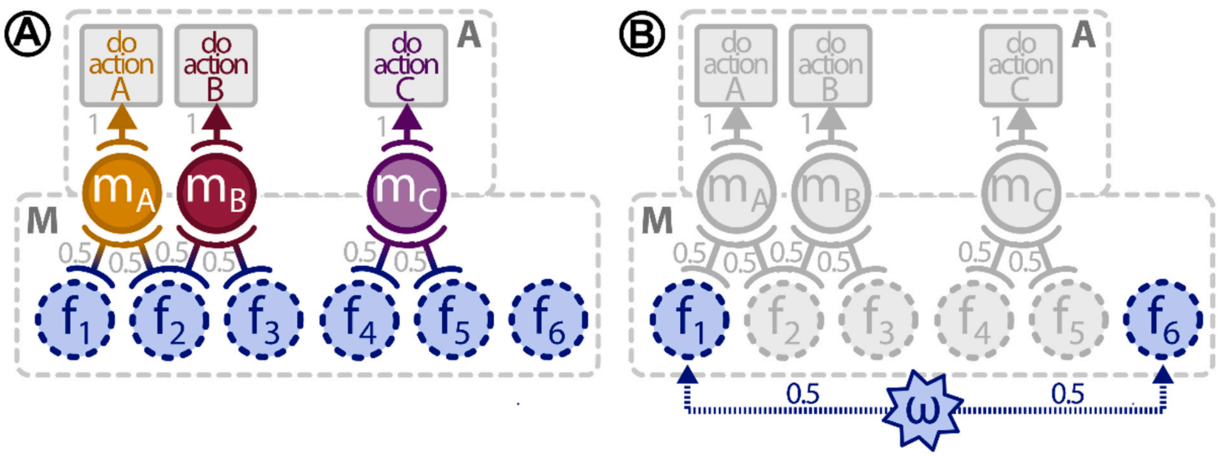

Fig. 2. An illustrative example of the cascading retrieval process. . This is a graph-based representation of the illustrative example detailed in the main text. **(A)**: In this example, matrix **M** consists of three memories: m_A, m_B and m_C, and matrix **A** consists of three actions (each supported by different memories). Memory A (m_A) consists of features f_1 and f_2 (both with a salience of 0.5) and supports action A. Memory B (m_B) consists of features f_2 and f_3 (both with a salience of 0.5) and supports action B. Memory C (m_C) consists of features f_4 and f_5 (again with a salience of 0.5 each), and supports action C. **(B)**: The observed stimulus consists of features f_1 and f_6, with an equal salience (0.5).

2.4. The sampling process stops once there are no memories retrieved during any one step

This sampling of episodes continues until a memory is retrieved which provides sufficient information to warrant a decision, or when there are no activated memories at the end of a step – that is, when there no candidates in C. If this occurs, then the decision-maker can decide to *abandon the search* and randomly select one the available actions. Alternatively, the decision-maker can *restart the search* process by halting for a brief period to wait out the refractory period of all sampled episodes, and then continuing with re-observation of the stimulus. Because the memory recall function α has an inherent randomness (both in determining which memories are candidates for recall, as well as in the competition between candidates), this restart is likely to yield a different set of recalled episodes.

The cognitive processes underlying the decision-maker's choice to either abandon the search or to restart the search process is beyond the scope of the current work, and likely involves a wide range of attentional and *meta*-cognitive factors. For the remainder of this work, we make the simplifying assumption that the decision-maker has limited patience: if the retrieval process terminates, the decision-maker will restart the search process with some probability, or otherwise abandon the search.

2.5. The cascading recall process yields three key predictions, as illustrated by two examples

In this section we provide a step-by-step illustration of the cascading recall process. For simplification we will consider a world which consists of only 6 features (f_1 to f_6), and an episodic memory store (**M**) which only contains three memories: memories A, B and C. Memory A (henceforth abbreviated as m_A) consists of two features: f_1 and f_2, both with a salience of 0.5. Memory B (m_B) consists of features f_2 and f_3, also both with a salience of 0.5. Finally, memory C (m_C) consists of features f_4 and f_5, both with a salience of 0.5. Fig. 2**A** provides a graphical representation. Since **M** contains one row for each memory and one column for each feature, we can represent this set of memories as:

$$\boldsymbol{M} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \end{bmatrix}$$

Note that m_A and m_B share a single overlapping feature: they both have a nonzero value for f_2 (column 2). In contrast, m_C is fully independent from the other two memories: there are no columns in which m_C has a non-zero salience together with either m_A or m_B. In addition, no memories contain f_6, and correspondingly column 6 of **M** only contains zeros.

For the purpose of this illustration, we will next assume that each of these memories supports a different action: memory m_A supports *action a*, memory m_B supports *action b*, and memory m_C supports *action c*. As the action-outcome matrix contains a single row for each possible action, and one column for each memory, we can represent these relationships between memories and supports actions as a matrix **A**:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The decision process starts with the observation of a stimulus ω. For this illustration we will set this stimulus to contain two features: f_1 and f_6, both with a salience of 0.5 (Fig. 2**B**). We can represent this stimulus as a column vector with one element for each feature f. Starting with the top-most value of f_1, and ending with the bottom-most value of f_6, we can write this vector as:

$$\omega = \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix}$$

2.5.1. Example 1 cued recall

We first illustrate the process of cued recall by assuming that all actions (a, b and c) are available. At the start of the first recall step, the stimulus first determines which memories are candidates for selection. Each memory's probability of being selected as a candidate can be obtained by equation (1), that is, by multiplying the memory's features (its corresponding row in **M**) with the features of the stimulus, and then taking the square root of this product.

$$p(C^{t=1} = 1) = \sqrt{\mathbf{M}\omega} = \sqrt{\begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \end{bmatrix}\begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \end{bmatrix}} = \begin{bmatrix} \sqrt{0.5*0.5} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 0 \end{bmatrix}$$

Memory m_A is the only memory which shares a feature with the stimulus, as both contain feature f_1 with a salience of 0.5. As a consequence, only m_A has a non-zero probability of being a candidate for recall, and the associated probability of it becomming a candidate is $\sqrt{0.5*0.5} = 50$ %. In contrast, neither m_B or m_C share any features with the stimulus and the probability that either is a candidate for recall is therefore equal to zero. Thus, if m_A is selected as a candidate, then it will not have any competition from other candidates and will always be selected for recall.

There are thus two different possible outcomes from the candidate-selection step. If m_A is a candidate, then it is recalled. If this occurs, then $\mu^{t=1} = [1 \quad 0 \quad 0]$. Upon recall, m_A next triggers the activation of its associated action and outcome. We represent this by multiplying the action-outcome matrix **A** with $\mu^{t=1}$:

$$\boldsymbol{A}\mu^{t=1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

This resulting vector represents the evidence in favor for each of the three actions. Since *action a* now has a value of 1, this action is selected and the recall process is concluded.

But what if memory m_A is not selected to be a candidate for retrieval? In this case there are no candidates for recall – and thus the retrieval process is terminated. What happens next depends on the patience of the decision-maker; they can either *restart the recall process*, or *abandon the search* and select an available action at random.

Fig. 3**A** provides a graphical depiction of which actions are selected as the cascading process unfolds, based on 10,000 simulated decisions. In this example we assume that the decision-maker has limited patience: if the cascading recall process terminates without supporting any one action, then she will restart the process with 75% probability, or select a random action with a probability of 25% (these probabilities are arbitrarily chosen for illustrative purposes).

As an illustrative simplification, here we exclusively depict the number of cascading recall *processes* before an action is selected, and

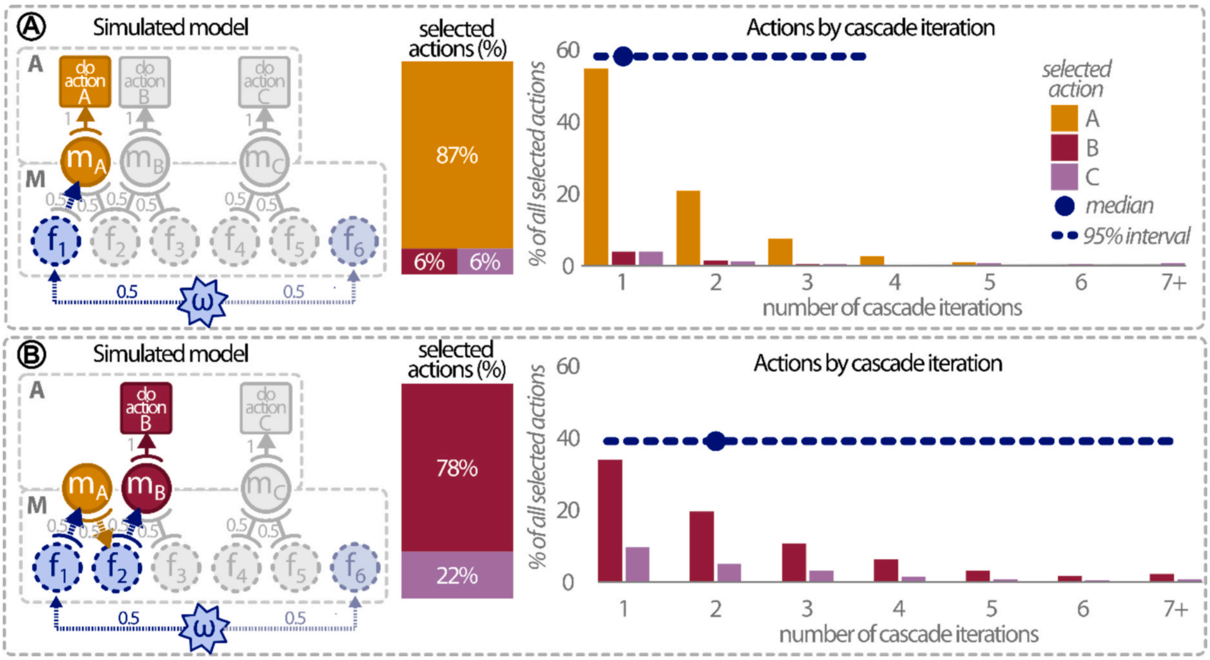

Fig. 3. Simulated results of the cascading recall process. Two sets of simulation results highlight the three key predictions of the cascading recall process. Both sets of simulations are based on 10,000 repeated decisions following the depicted network, and with a limited patience (if the recall process terminated without selecting an action, then the process was restarted with a 75% probability, or an available action was selected in 25% of cases). The stimulus always consisted of features f_1 and f_6, both with a salience of 0.5. Consequently, the observation of f_1 triggered the retrieval of m_A in 50% of simulations. **(A)**: In the first set of stimulations, the action previously associated to the cued memory (action A) was available. If this memory was retrieved, then action A was selected. Thus, a large majority of all simulated decisions (87%) resulted in the selection of action A, and the other two actions were selected in only 6% of simulations each. This illustrates the first prediction: we preferentially draw upon episodic memories which share features with the observed context (cued recall). This cued action is selected relatively fast, with 95% of all decisions being completed after the third repetition of the retrieval process. **(B)**: Action A was not available in the second set of simulations. The retrieval of m_A now reinstated its constituent features (f_1 and f_2), which in turn stochastically triggered the retrieval of m_B with 50% probability. If m_B was retrieved, action B was selected, illustrating our second prediction: if cued memories do not yield sufficient information to warrant a decision, then we next draw on additional, non-cued, memories which are sampled in subsequent steps. Notably, such cascading recall is both more stochastic (m_B was selected in 78% of simulations) and slower (95% of decisions being completed after 7 iterations of the retrieval process) than cued recall, illustrating our third prediction.

not the number of steps *within* each cascading recall process. While the execution of steps *within* a cascade is not assumed to be instantaneous, the time *between* restarts of the cascading process (to out-wait the refractory period of recalled memories) is presumably the significant driver of differences in reaction times. In this example, the cascading recall process results in the selection of action a in a large majority of all simulated decisions, whilst actions b and c are selected only a small number of times. Moreover, an action is selected relatively rapidly, with 50% of all simulations yielding a decision after the first repetition of the cascading process, and 95% after the third repetition.

These simulated results illustrate the first key prediction of the cascading recall process: when we are faced with a novel decision context, we do not recall all past episodes with equal probability. Instead, we preferentially draw upon episodic memories which share features with the observed context (cued recall).

2.5.2. Example 2. cascading recall (when cued recall is not sufficient to select an action)

But what if action a is not available? We can represent this unavailability by zeroing the row-vector for action a in **A**. We denote this new action-outcome matrix as **A'**:

$$\boldsymbol{A}' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We will now illustrate how the retrieval process unfolds if *action a* is not available. In this example, the first step (cued recall) proceeds as before. At the end of the first step, either m_A is recalled (again with 50% probability) or the recall process failed to retrieve a memory and hence terminates.

If m_A was retrieved, then it next activates its corresponding column in **A'**: $\boldsymbol{A}'\mu^{t=1} = [0 \quad 0 \quad 0]$. In contrast to the first example, the retrieval of m_A now no longer leads to a decision, as *action a* is not available. Instead, the retrieval process now continues to a second step. At the start of this step, memory m_A reinstates all of its constituent features. We can denote this reinstatement with the transposition of matrix **M**, such that:

$$\varphi^{t=2} = \mathbf{M}^{\mathrm{T}}\,\mu^{t=1} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The activation of m_A results in the activation of features f_1 and f_2, both with a salience of 0.5.

These activated features now form the input for an additional (cascading) recall step. Note that m_A was previously sampled, and its activation is now suppressed for the remainder of this sampling process. In line with equation (3), we represent this suppression by fixing m_A's recall probability to 0. For notational simplicity, we do so by modifying **M:** setting all features of memory *a (top row)* to zero. This results in a probability of zero that this memory is considered as a candidate for retrieval:

$$p(C^{t=2}=1) = \begin{cases} 0, & \text{if } m_i \text{ was previously sampled;} \\ \sqrt{m_i\varphi^{t=2}}, & \text{otherwise} \end{cases}$$

$$p(C^{t=2}=1) = \sqrt{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}} = \begin{bmatrix} 0 \\ \sqrt{0.5*0.5} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.5 \\ 0 \end{bmatrix}$$

Only memory B has a non-zero probability of being selected as a candidate during the second recall step, with an associated probability of 50%. If m_B is a candidate, then it is the only candidate and does not have any competition from other candidates – and hence it is selected for recall.

If m_B is recalled, then $\mu^{t=2} = [0 \quad 1 \quad 0]$ and correspondingly $\boldsymbol{A}'\mu^{t=2} = [0 \quad 1 \quad 0]$, resulting in the selection of *action b*. If memory B is not a candidate for recall, then the cascading recall process terminates as there are no other candidates.

Fig. 3**B** provides a graphical of the results of 10,000 simulated decisions. As before, we assume that if the cascading recall process terminates without supporting any one action, then the decision-maker will restart the process with 75% probability, or select a random action with a probability of 25%.

In this example the cascading recall process results in the selection of *action b* in a large majority of all simulated decisions, whilst *action c* is selected much less frequently. These results illustrate the second key prediction of the cascading recall process: if cued memories do not yield sufficient information to warrant a decision, then in subsequent steps we next draw on additional (non-cued) memories. This additional retrieval starts with the reinstatement of the previously retrieved memory's features. These reinstated features then form the inputs for a subsequent memory retrieval process.

Additional key observations emerge when we compare the results from the first example with the results from the second example. First, the fraction of decisions in which *action a* was selected in the first example (87%) is higher than the fraction of decisions in which *action b* was selected in the second example (78%). Moreover, an action is selected more slowly in the second example than in the first example, with 50% of all simulations yielding a decision after the second repetition of the cascading process, and 95% after the sixth repetition. In comparison, in the first example 50% of all simulations yielded a decision after the first repetition of the recall process, and 95% after the third replication.

This comparison illustrates the third key prediction of the cascading recall process. All else being equal, the cascading recall process is more likely to terminate without recalling a memory if there are more steps between the stimulus and a relevant memory. Each time that the recall process terminates, there is some probability (25% in these examples) that the decision-maker abandons the search and selects an available action at random. In addition, repeating the cascading recall process requires a non-zero amount of time. As a consequence, the cascading recall process predicts that, all else being equal, actions based on memories that only require a single step of recall (cued recall; example 1) requires less time and are less stochastic than actions guided by memories which require multiple, sequential, sampling steps (cascading recall, example 2).

3. Experimental evidence supports the predictions cascasing recall process

We test all three predictions through a novel experimental paradigm (Fig. 4). In this paradigm, participants naturalistically engage with previously unencountered stimuli and are given minimal instructions. In order to create memorable stimuli while minimizing exogenous associations, the key stimuli used in all three experiments was a series of fantasy animals called *Fennimals*. Prior to data collection, this experiment design obtained ethical approval by the Ethics Assessment Committee Faculty of Law and Nijmegen School

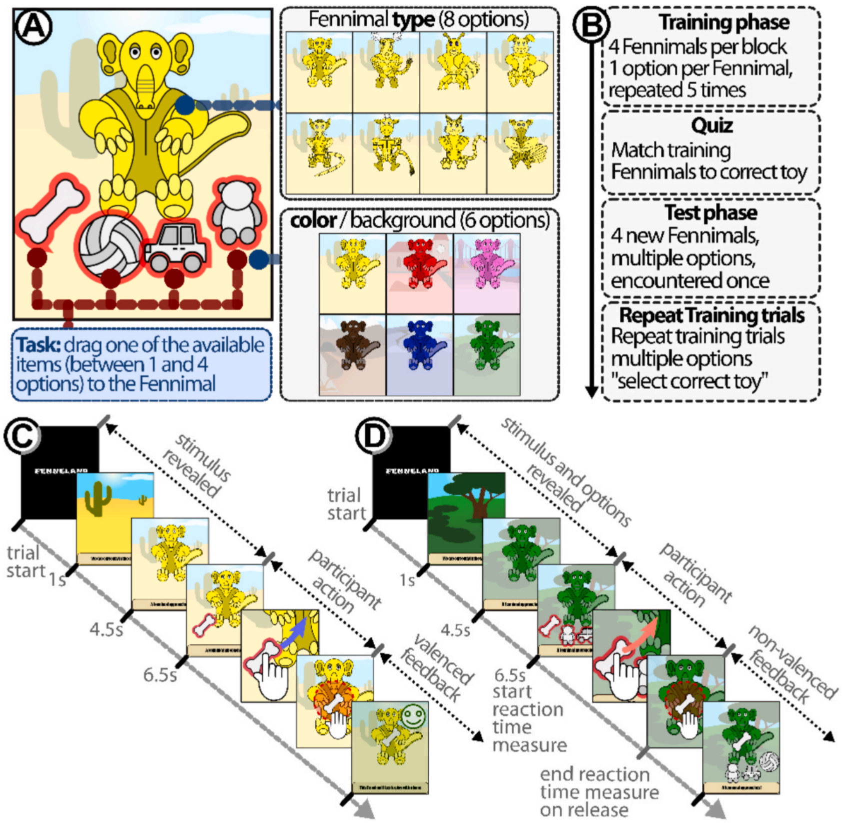

Fig. 4. Experiment Design for the Fennimals Task While the underlying structure of the stimuli was identical for all participants, the construction of Fennimals and the on-screen location of the toys were randomized on a between-subjects basis. **(A):** Each Fennimal consisted of two variable features, with each Fennimal consisting of one of eight different types and one of six different color/background combinations. **(B)**: schematic overview of the experiment procedure. The training phase consisted of 5 blocks, each containing the same 4 training-Fennimals, presented in random order. This was followed by a quiz block, where participants had to select the correct toy (out of four options) to give to the Fennimal. Failing this quiz resulted in an additional training block, followed by an additional quiz (and so on). Upon passing the quiz participants continued to the search phase and interacted with 4 new Fennimals. The experiments concluded with a final block in which participants encountered the original training-phase Fennimals and could select one of four toys. **(C)**: schematic representation of an example trial in the training phase, where participants were always forced to select a single toy per trial which was followed by positively valanced feedback. **(D)**: schematic representation of an example trial in the search phase, where participants encountered a novel Fennimal and could select one of multiple toys.

of Management (EACLM, reference number 2020.08).

General Experimental Methods

3.1. Stimuli

Each Fennimal consisted of two features, which could be varied independently of each other (Fig. 4**A**). The first feature was the Fennimal's *type*, which determined the shape of the Fennimal. The software allowed for 8 different types, which were constructed to be visually distinct from each-other. In addition, the Fennimals could differ in their color-schemes and location (as indicated by their background). Here the software allowed for six distinct options. We will henceforth refer to this color/location feature as the Fennimal's *color*. Participants interacted with these Fennimals by handing them different toys. Four different toys were included in the experiments: a bone, a ball, a car and a teddy bear. These toys were presented on the bottom part of the screen during each interaction.

While the underlying structure of the stimuli was the same for all participants within each of the three experiments, the experiment software included a large degree of randomization in order to prevent any potential confounds. At the start of each experiment, the software first randomly determined which Fennimal types and color/backgrounds were used throughout the experiment. For example, where one participant may encounter an elephant-like Fennimal in the desert, a different participant may have encountered a green rat-like Fennimal in a forest instead. The software next randomized all item-associations; one particular item may have been presented as a ball to one participant, but as a bear to a different participant. Finally, the software randomized the relative location of all four toys. As a result, even if some item was represented as a ball to two participants, this item could have been presented on different places on the screen. Due of this randomization, the exact Fennimal types, colors, item-associations and the on-screen positions of the items differed between all participants.

3.1.2. Procedure

All three experiments consisted of two phases: a training phase followed by a memory search phase. At the start of the training phase, participants were informed that they would go on a Safari to a fantasy island ("Fenneland"), where they would encounter a unique set of creatures. Participants were informed that they would be tasked with handing different toys to the Fennimals, and were instructed to pay attention to which Fennimal liked which toy. No further instructions were provided, and participants were not informed about the need to draw on these experiences during the later portion of the experiment.

The training phase consisted of multiple blocks of trials. During each trial, participants were given a single toy to hand to the Fennimal. Handing this toy to the Fennimal always resulted in a positive feedback signal: "*The Fennimal likes to play with the [toy]*". This training block was repeated five times in order to ensure that participants correctly stored committed these episodes to memory. The order in which the Fennimals were encountered was randomized for each block of trials.

After the fifth block of training trials, participants were next given a quiz to test whether they correctly stored these associations. This quiz contained one trial for each Fennimal encountered in the training phase. During each quiz trial, participants were shown the previously encountered Fennimal and had to select the toy which they previously gave to this Fennimal. Participants could freely select between all four toys. After selecting a toy, participants received a feedback message: "*Oops, You picked the wrong toy!*" if participants selected the incorrect toy, or "*Correct!*" if they selected the correct toy. If participants made any mistakes, then they would be given an additional block of training trials, followed by a restart of the quiz. This continued until participants correctly matched all Fennimals to their previously associated items.

Upon passing the quiz, participants continued to the search phase of the experiment. At the start of this phase, participants were informed that they would encounter more Fennimals, and that they had to select a toy which they believed the Fennimal would like. They earned a Token for each Fennimal which liked the toy, and an additional perfect-game bonus of 4 Tokens if all Fennimals liked the toy. Each Token was worth \$0.10 at the end of the experiment, and a perfect-game bonus was included to prevent participants from following a hedging strategy. To prevent intermittent learning, participants were not given any feedback on whether the Fennimals liked the toy until the end of the experiment. In order to prevent participant's from selecting toys which were easy to remember, they were informed that there would not be an additional quiz at the end of the experiment. In order to ensure that participants had to flexibly rely on episodic recall during the search phase, no further instructions were provided to participants.

The search phase consisted of two blocks of trials. The presentation of trials was randomized for each block. During the first block participants encountered four new Fennimals. Each of these new Fennimals shared either a color or a type with exactly one previously encountered Fennimal, but also had one previously un-encountered feature. Participants were tasked with selecting one toy to give to each of these new Fennimals. After this block of trials, participants once again interacted with the original four training-phase Fennimals and were asked to match these Fennimals with the toys they originally liked. Participants were then informed how many Tokens they earned during the search phase, after which the experiment was concluded.

3.1.3. Analysis

The main trials of interest for all three experiments were the first block of the search phase. In this block, participants encountered unfamiliar Fennimals, and could freely select a toy to give to this Fennimal. Although the exact setup differed between the three experiments, all trials were designed such that the cascading recall process predicted that participants would be more likely to select one toy over all others. We controlled for the clustering of trials within a participant by calculating the number of trials in which the participant selected this predicted toy. This resulted in a single value for each participant. We compared this value to a simulated test-distribution based on 10^6 simulated samples in which we assumed the null-hypothesis that participants selected a toy at random. Note that in the illustrative simulations above we simulated the *percentage* of toys selected, whilst in the experiments we measured the *mean* *number* of trials in which participants selected the predicted toy. In order to relate this modelled variable to the experiment observations, we therefore also report the percentage of trials in which participants selected the toy predicted by the cascading recall process.

3.1.4. Participants and power analyses

All three experiments were conducted online and with non-overlapping participant pools via Amazon Mechanical Turk. Based on the effect sizes observed in pilot experiments we conducted a power simulation for each experiment to determine the minimal sample size required for an 80% power. As these simulations indicated a low minimally required sample size, and for reasons of prudence as well as due to low cost per observation, we a-priori registered a tenfold multiplication of these minimal sample sizes. This resulted in the targeted collection of 100, 200 and 200 participants for the three experiments respectively.

Participants were excluded from analysis if they had any form of colorblindness (as this may have reduced their ability to differentiate between the Fennimals' colors) or if they minimized or otherwise lost focus on the experimental window during one or more search trials (as this may indicate that the participant had been distracted during these trials). Participants were excluded from further analysis if they made one or more errors during the last block of the search phase, in which they were tasked to match each training-phase Fennimal with its previously associated toy. A failure to match these Fennimals to their correct toy may indicate that participants did not properly memorize the training stimuli. All reported results are robust to the inclusion of these participants.

3.1.5. Preregistration and data-availability

All experimental designs, corresponding analyses, and the exclusion criteria were pre-registered prior to data collection. The preregistrations, experimental software, raw data and analysis scripts are freely available at https://osf.io/fcqhr/.

3.2. Experiment 1: Participants preferentially draw on episodic memories which share a feature with the novel context (cued recall)

In the first experiment (Fig. 5), we tested the first prediction of the cascading recall process. When we are faced with a novel decision context, the cascading recall process predicts that we do not recall all past episodes with equal probability. Instead, we preferentially draw upon episodic memories which share features with the observed context (cued recall).

3.2.1. Methods

In order to test this hypothesis participants interacted with four Fennimals during the training phase. Of these four Fennimals,

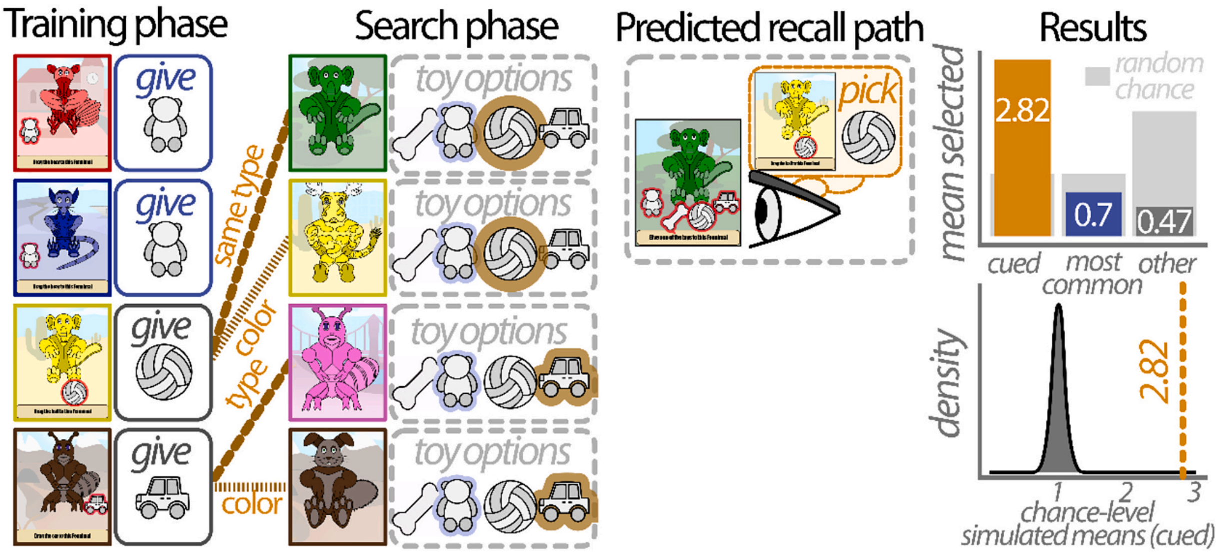

Fig. 5. Setup and results for the first experiment During the training phase participants encountered four non-overlapping Fennimals, two of which were positively associated with the same toy (the bear in the depicted example), while the other two Fennimals were associated with two different toys. During the search phase, participants encountered a novel Fennimal which shared an overlapping feature (either a type or color) with either of the latter two Fennimals. Participants then had to decide which one of four toys to give to this new Fennimal. We predicted that participants would be likely to retrieve the cued memory, and accordingly to select the toy associated to the directly similar training-phase Fennimal. Results support this prediction, with participants selecting the predicted toy in on average 2.83 out of 4 trials (71%). In contrast, the statistically most common toy (the bear in this example) was only selected in 0.7 out of 4 trials on average (18%). The gray bars indicate a chance-level performance. We next simulated 10^6 hypothetical experiments in which the same number of participants picked toys at random. The distribution of these random choices strongly suggests that our results are unlikely to be the result of random chance alone ($p < .001$). Note that the depicted Fennimals are an example only; all Fennimals and item-associations were randomized on a between-subject basis. The highlights around the toys are for illustration only, none of the options were highlighted on the participant's screen.

participants gave the same toy to two Fennimals (both of whom liked the toy). The other two Fennimal both liked a different toy. Thus, participants gave three different toys to the training-phase Fennimals, and one of these toys was given twice as often as the other two toys.

Participants encountered four new Fennimals during the search phase of the experiment. Each new Fennimal had one novel (i.e. not previously encountered) feature, and also shared one feature with one of the two training-phase Fennimals which liked different toys. Participants could next select between all four toys. If all training-phase Fennimals were equally likely to be recalled during these search-phase trials, then participants would be more likely to recall interactions in which they gave the toy that received positive feedback most often. In contrast, if our hypothesis was supported and participants were more likely to retrieve the Fennimal with overlapping features (cued recall), then they would be more likely to select the item associated to the training-phase Fennimal which shared a feature with the search-phase Fennimal.

3.2.2. Results

Of the 100 collected responses, the responses of 3 participants were lost due to software errors. Based on the exclusion criteria we excluded 23 participants from further analysis, leaving 74 valid responses. On average, participants selected the toy previously associated with the overlapping training-phase Fennimal in 2.82 of the 4 search-phase trials (71%). In contrast, participants selected the toy which was most frequently observed during the training phase was selected in on average 0.7 out of 4 trials (18%), and one of the other two toys in 0.48 out of 4 trials (12%). The rate in which participants selected the cued-recall toy (as compared to any of the other toys) was significantly larger than is expected if participants selected a toy at random (simulated one-sided $p < .001$).

3.2.3. Discussion

The results of the first experiment support the first prediction of the cascading recall process: the observation of a novel situation triggers the cued recall of similar past experiences – even when other items have been observed to result in a positive outcome at a higher frequency. Such a preferential reliance on cued memories is consistent with the proposed cascading recall process.

3.3. Experiment 2: When cued recall is insufficiently informative, participants instead rely on indirectly related experiences

What if cued recall does not yield sufficient information to warrant a decision? Here the cascading recall process predicts that the sampling process continues to draw upon additional memories tin sequential search steps. In line with the second prediction, these additional memories are retrieved by reinstating the constituent features of the memory retrieved through cued recall, and using these

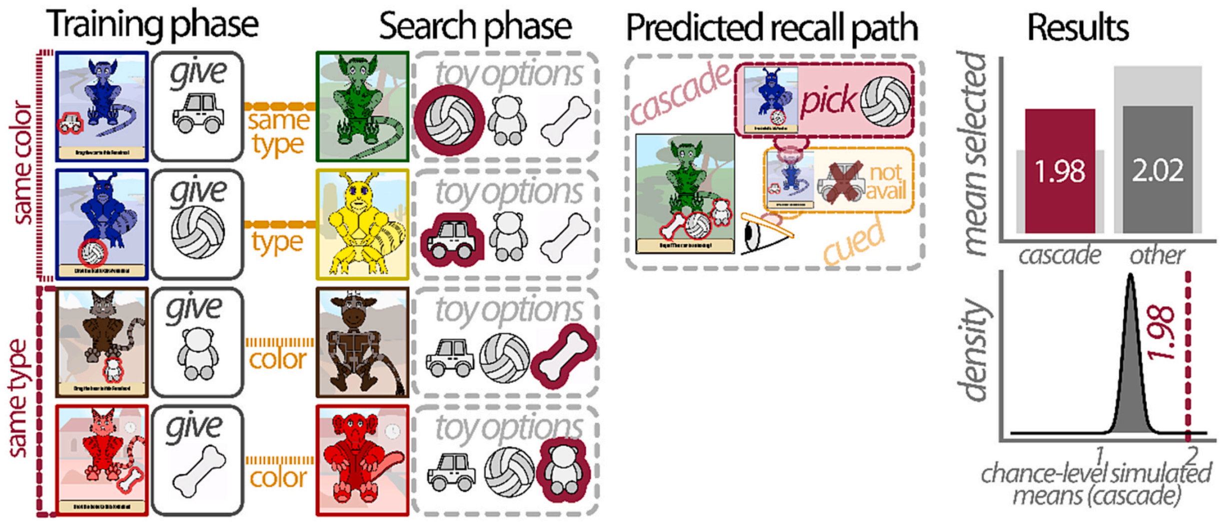

Fig. 6. Setup and results for the second experiment. During the training phase participants encountered four Fennimals, each of which was associated with a different toy. Fennimals shared overlapping features, with two Fennimals the same color and two Fennimals sharing the same type. During the search phase participants encountered a novel Fennimal which shared either a type or color with one previously encountered Fennimal. Participants could not select the toy associated with this directly overlapping Fennimal. We predicted that participants would first retrieve the directly similar Fennimal (cued recall), whose associated toy was not available. Participants were then predicted to reinstate the features of this cued Fennimal, resulting in the cascading retrieval of the training-phase Fennimal with a similar feature was the cued Fennimal – and the selection of this indirectly associated toy. The results supported this prediction: participants selected the predicted toy in on average 1.98 out of 4 trials (49%), which was substantially more than would be expected by chance alone (33% chance level per decision). We next simulated 10^6 hypothetical experiment results in which the same number of participants selected toys at random. The distribution of these random choices strongly suggests that our results are unlikely to be the result of random chance alone ($p < .001$). Note that the depicted Fennimals are an example only; all Fennimals and item-associations were randomized on a between-subject basis. The highlights around the toys are for illustration only, none of the options were highlighted on the participant's screen.

features as inputs for a subsequent memory retrieval process. We test this prediction in the second experiment (Fig. 6).

3.3.1. Methods

During the training phase, each of the four Fennimals was associated with a different item. In contrast to the first experiment, some of these training-set Fennimals shared overlapping features, such that two Fennimals shared a color scheme (but had different types) and two Fennimals shared a type (but had different color-schemes). During the search phase, participants encountered four new Fennimals. Each of these new Fennimals shared a feature with exactly one training-phase Fennimal. Thus, we predict that the observation of these novel search-phase Fennimals results in the cued recall of a single training-phase Fennimal.

In contrast to the first experiment, here we restricted which toys were available to the participants. In particular, the toy which was previously associated with the cued training-phase Fennimals was **not** available during the search trial. As a result, participants could not rely on cued recall during these trials but instead had to engage in a memory search process to guide decision-making. The cascading recall process predicts that participants then reinstate all the features of the cued training-phase Fennimal to retrieve additional memories.

Given that this cued training-phase Fennimal shared a feature with one other training-phase Fennimal, the cascading recall process predicts that participants are more likely to retrieve this indirectly-related training phase Fennimal – and to consequently select the toy which was previously associated with this Fennimal.

3.3.2. Results

Of the 200 collected responses, the responses of 3 participants were lost due to software errors. Based on the exclusion criteria we excluded 61 participants from further analysis, leaving a total of 136 valid responses. On average, these participants selected the toy previously associated to the indirectly related Fennimal in 1.94 out of 4 trials (49%). Note that participants could select one of three available items during the search phase, corresponding to a one-in-three probability that by chance alone the participant would select the singular item associated to the indirectly related Fennimal. The rate in which participants selected the predicted toy (as compared to any of the other toys) was significantly larger than is expected if participants selected a toy at random (simulated one-sided $p < .001$).

3.3.3. Discussion

The results of this experiment support the second prediction of the cascading retrieval framework: when memories retrieved through cued recall do not support the selection of an action, participants then drew on episodes which were indirectly related to the new Fennimal. Importantly, this reliance on indirectly related experiences occurred in a naturalistic choice environment where

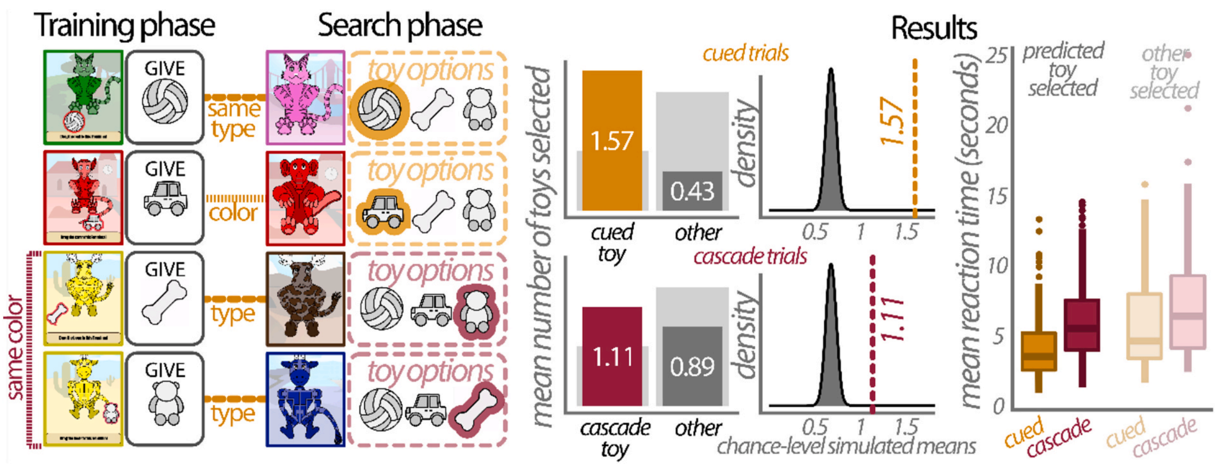

Fig. 7. Setup and results for the third experiment During the training phase participants encountered four Fennimals, each associated with a different toy. Two Fennimals shared a color. The search phase consisted of two types of trials. In cued-recall trials, the Fennimal shared a feature with the two training-phase Fennimals which did not share a color with any other Fennimal, and participants could select the toy associated with the directly overlapping Fennimal (with two other toys available). In cascading recall trials, the Fennimal shared a feature with one of the two training-phase Fennimals which shared a color and participants could either select the toy associated with the indirectly related Fennimal, or could select one of two other toys. The results replicated both previous experiments, with participants selecting the toy associated with the similar Fennimal significantly above chance levels (33%) in both the cued trials (on average selecting the predicted toy in 1.57 out of 2 trials; 79%) and in the cascading recall trials (1.11 out of 2 trials; 55%). We next simulated 10^6 hypothetical results in which the same number of participants selected 2 toys at random. The distribution of these random choices strongly suggests that our results are unlikely to be the result of random chance alone in both the cued and cascading trials (both $p < .001$). The left-most panel depicts the mean reaction times per participant, excluding 3 participants which had at least one reaction time more than four standard-deviations above the mean. Note that the depicted Fennimals are an example only; all Fennimals and item-associations were randomized on a between-subject basis. The highlights around the toys are for illustration only, none of the options were highlighted on the participant's screen.

participant's decisions in the second phase of the experiment were one-shot, where outcome feedback was not provided, where participants were given minimal instructions, and where participants had to adaptively recall indirectly relevant prior experiences without being told to do so during the initial learning phase.

3.4. Experiment 3: Decisions based on cued recall are less variable and are made faster than decisions based on cascading recall

The setup of the Experiment 2 did not allow for a within-subject comparison between trials in which participants could rely on cued recall versus trials in which they had to rely on cascading recall. Such a direct comparison is clearly key to evaluate the third prediction of the cascading recall process, which predicts that, all else being equal, decisions based on cued recall require less time and are more likely to result in the retrieval of a relevant past episode as compared to memories which require multiple (cascading) recall steps. To examine this question, we conducted a third experiment (Fig. 7) to facilitate such a prediction by including both cued-recall and cascading recall trials in the memory-search phase.

3.4.1. Methods

Participants encountered four Fennimals during the training phase, each of which was associated with a different item. Two of these Fennimals did not share any features with other Fennimals (mimicking the first experiment), while the other two Fennimals shared a color scheme with each other (mimicking the second experiment).

Participants again encountered four new Fennimals during the key trials in the search phase of the experiment. Two of the new Fennimals were *cued-recall* trials. These Fennimals shared a feature with either of the training Fennimals which did **not** share features with the other training-phase Fennimals, and the toy associated with this directly related training-phase Fennimal was available. As a result, participants could rely on cued recall in these trials. The other two search-phase trials constituted the *cascading recall* trials. In these trials, the new Fennimals shared a type with either one of the two training-phase Fennimals which shared features, the toy associated with this cued Fennimal was **not** available. As a result, participants had to rely on cascading recall on the other two trials.

In order to equalize the rates at which items would be selected if participants picked a toy at random, one of the non-cued toys was selected at random to be not available during the trials in which the participant could rely on cued recall. In addition, we measured the time between presentation of the Fennimals and their decision. This reaction-time measurement started when the Fennimal was first displayed to the participant, and stopped when the participant released the selected toy over the Fennimal. This setup therefore allowed for a direct comparison between the two types of trials to test the predicted difference in accuracy rates and reaction times.

As in the first two experiments, we hypothesized that participants would search their store of memories to retrieve a relevant training-phase Fennimal. This retrieval requires only a single recall step during cued recall trials, but multiple recall steps during the cascading-recall trials. During each recall step, there was a non-zero probability that the participant would fail to retrieve a subsequent memory. Thus, participants were predicted to require more iterations of the recall process for the cascading recall trials than for the cued-recall trials.

Participants could decide to abandon the search process after it failed to retrieve a memory, and to instead select a toy at random. As the search process requires less steps, and is therefore less likely to fail in cued recall trials, we hypothesized that participants would be more likely to select the predicted toy in the cued-recall trials than in the cascading recall trials. In addition, restarting the recall process is likely to require some time. As the cascading recall trials are likely to require more iterations of the search process before retrieving a relevant episode, we therefore hypothesized that, *if participants selected the predicted toy,* then they would on average be faster to do so during the cued recall trials as compared to the cascading recall trials.

3.4.2. Analysis of reaction times

In order to evaluate the predicted difference in reaction times when participant selected the predicted toy between cued versus cascading trials, we first calculated the mean response times on an individual basis. If participants selected the predicted toy in both trials of the same type (i.e., in both cued recall trials or in both cascading recall trials), then we calculated the mean response time of these two reaction-times. We then compared the mean reaction time between cued-recall and cascading-recall trials through a permutation test. This test was selected as it is distribution-agnostic; that is, it is robust for deviations from normality in the reaction-time measures. In this test we first calculated the observed difference in reaction times between the two trial types. The statistical significance of this observed difference was then evaluated against a simulated test distribution. This test distribution reflected the expected difference in means if the null-hypothesis was correct − that is, if there was no true difference in mean reaction times between the cued- and cascading-recall trials. Each test distribution contained 10^5 simulated mean-level differences. Each of these differences was determined by randomly shuffling all reaction-time measurements between the two trial types on a within-subjects basis. We next calculated the mean difference between cued and cascading trials in this randomized sample. These simulated means then formed a distribution of differences which would be *observed if the null-hypothesis would be correct* (no differences between the two trial types). As a robustness check, we ran a separate set of analyses in which all trials were evaluated independently, that is, without averaging them on a participant-level. All reported results are robust to these analyses.

3.4.3. Results

Of the 200 collected responses, the responses of 5 participants were lost due to a software error. Based on the exclusion criteria 55 participants were removed from further analysis, leaving a total of 140 valid responses. The observed results are consistent with the previous two experiments. In the cued-recall trials, on average participants selected the toy which was previously associated with the directly similar training-phase Fennimal in 1.57 out of 2 trials (79%). This was significantly more than would be expected if participants would have selected a toy at random (33% chance level, simulated one-sided $p < .001$). In cascading recall trials, participants selected the toy which was previously associated to the indirectly related training-phase Fennimal in 1.11 out of 2 trials (55%). This was significantly more than would be expected if participants would have selected a toy at random (33% chance level, simulated one-sided $p < .001$). The difference between these two types of trials was both substantial (24 percentage points) and significant (Fisher's exact $p < .001$). That is, participants were significantly more likely to select the predicted toy in the cued-recall trials, as compared to the cascading-recall trials.

We next explored participant's mean reaction times between the different trial types, descriptives of which are provided in Table 1. These reaction times were consistent with our hypothesis: when selecting the predicted toy, participants were faster to do so on cued recall trials (mean reaction time = 4.5 s) than on cascading recall trials (mean reaction time = 6.42 s). The results of the permutation test indicated that this difference in mean reaction times (1.9 s) was unlikely to have occurred through chance alone (permuted $p < .001$). These findings are robust to the exclusion of three participants who had a mean reaction time more than 3 standard-deviations above the mean for any of the reported reaction times.

3.4.4. Discussion

The results of the Experiment 3 supported the third prediction of the cascading recall process outlined above: participants took significantly more time to select a toy when cued recall yielded an informative past experience, as compared to trials in which a second recall step was required. Moreover, participants were significantly and substantially more likely to select the predicted toy when they could rely on cued recall than when they had to engage in a memory search process.

In addition, the design of Experiment 3 excludes a potential alternative explanation of the observed results in Experiment 2. During the training phase of Experiment 2, participants could have memorized the co-occurrence between two pairs of items, one for each set of overlapping training-phase Fennimals. During the search phase, participants could next have decided to select a toy from either of these two pairs at random, that is, *without consideration of the search-phase Fennimal's features*. Such a paired-item based strategy would have yielded a 50% chance of picking the item associated with the indirectly related Fennimal, and a 25% chance for each of the two alternative items, which is in line with the data observed in the second experiment. The setup of Experiment 3 precludes this alternative explanation by including only a single set of related training-phase Fennimals. During the cascading recall trials of Experiment 3, this strategy would have resulted in participants randomly selecting a toy from either the paired items (resulting in selection of the predicted toy), or from either of the other two Fennimals at random. Now this alternative strategy would have resulted in 33% probability to select the toy which was predicted by the cascading recall process – which was lower than was observed in the data.

4. General discussion

By building on key findings on the functioning of the episodic memory system, here we proposed a process of cascading episodic recall. This process retrieves past experiences in discrete steps, such that the features of a retrieved memory provide the input to retrieve a further memory in a subsequent step. The results of the three experiments provided empirical support for such cascading recall. When making a decision in a novel context, with a limited amount of training and minimal instructions, participants relied on the cued recall of similar previous experiences. They did so even when other alternatives were more frequently observed to result in a positive outcome (experiment 1). If these cued experiences did not provide information to warrant a decision, then, as predicted by the cascading recall model, participants relied on indirectly related experiences (experiment 2), Finally, participants were substantially faster and more likely to select a probe stimulus when they could rely on cued recall than if they had to engage in a process of memory search (experiment 3).

4.1. Cascading recall bridges several models in memory-based decision-making

While the current work is the first to propose this cascading recall process, it is far from the first model of memory-based decision-making. Below we identify four research efforts that utilize past experiences to inform behavior in novel situations: similarity-based

Table 1
Descriptive statistics for the mean reaction times per participant. If participants selected the same item type for both trials of the same trial type (e.g. selected the predicted toy in both cued recall trials), then we first calculated the mean reaction times for both trials. If this was the case, then the participant did not have a measured reaction time for the other item type of the same trial type (e.g. no measured reaction time for selecting any of the other toys on the same recall trials). The last four columns represent a robustness check, in which participants (3) were removed if their mean reaction time was more than 4 standard-deviations above the mean for any of the rows in this table.

	Full sample				3 outliers removed			
Trial types	Num obs	Mean	Median	SD	Num obs	Mean	Median	SD
All trials	140	6.3	5.07	6.85	137	5.65	5.01	2.78
Cued trials	140	5.44	3.71	8.89	137	4.54	3.71	2.61
Predicted toy selected	132	4.5	3.54	3.18	129	4.18	3.5	2.39
Other toy selected	52	9.47	4.86	25.76	50	5.77	4.63	3.31
Cascade trials	140	7.15	6.01	5.46	137	6.76	5.81	3.68
Predicted toy selected	108	6.42	5.56	3.4	106	6.34	5.48	3.36
Other toy selected	93	7.65	6.47	6.51	91	7.1	6.41	4.29

models, models of limited sampling, models based on hippocampal generalization and models in which items are retrieved through their associative context. These efforts share a conceptually similar approach to memory-based decision-making, yet are currently disjointed from one-another. We illustrate how the cascading recall process is both consistent with core elements of all approaches, and also offers a conceptual bridge between them.

In particular, the cascading recall process mimics the similarity weights of similarity-based decision-models. The cascading recall process bridges between these similarity-based models and models of limited sampling by providing a mechanism through which a limited number of past experiences can be sampled to guide our behavior. Our model adds to the literature on limited-sampling models by proposing a mechanism to adaptively retrieve a sample of past experiences. In addition, while both similarity-based models and models of limited sampling are not inconsistent with the organization of the episodic memory system, they currently lack a direct link to recent advances on this topic. Conversely, while models based on hippocampal generalization capture the neural dynamics of the episodic memory system, it remains unclear how these models align with the limited sampling of past experiences. By relaxing a previous assumption in these models of hippocampal generalization, the cascading recall process offers a mechanism to facilitate such limited sampling. Finally, the cascading recall process extends on associative-context models by including both a cognitive mechanism to support the retrieval of memories through their associative context, as well as extending these models to retrieve experiences which do not share any features with the observed stimulus.

4.1.1. As in similarity-based models, decisions are more strongly influenced by similar past experiences

Just as in the cascading recall process, similarity-based decision models are based on the assumption that not all past experiences are equally relevant for the decision at hand. Instead, similar past experiences are likely to be more informative than dissimilar experiences, and thus should carry more weight in the decision process. This similarity-weighted decision process has previously been proposed across a rich collection of models, spanning research across psychology, economics and neuroscience (Bordalo et al., 2020; Cox & Criss, 2020; Dougherty et al., 1999; Gilboa & Schmeidler, 1995; Gonzalez et al., 2003; Thomas et al., 2008).

At the core of these similarity-based models lies an evaluative process in which the decision-maker iterates through their memories and judges the similarity between the memory and the current decision context. The impact of each previous experience is then weighted by this similarity measure, resulting in a similarity-weighted expected value for each possible action. A key component of these models is therefore the relative weighting of past expereinces on their (dis)similarity to the observed stimulus. This key insight is consistent with, and replicated in, the first (cued recall) step of the cascading recall process.

To demonstrate how the cascading recall process mimics such similarity-based decision-making, we first pick a working definition for (dis)similarity. Although various metrics of similarity have been proposed (Goldstone & Son, 2012), these models have typically conceptualized similarity as being inversely proportional to the Euclidian distance between the observed stimulus (ω) and a memory (m):

$$\textit{Euclidean Distance}: \ \textit{dist}(\omega, m) = \sqrt{\sum_{f \in \{1,F\}} \left(m_f - \omega_f\right)^2}$$

Although the cascading recall process does not rely on an explicit evaluation of the similarity between the stimulus and memory, it nonetheless mimics such distance-based weighting. In the cascading recall process, a memory only influences the decision process if it

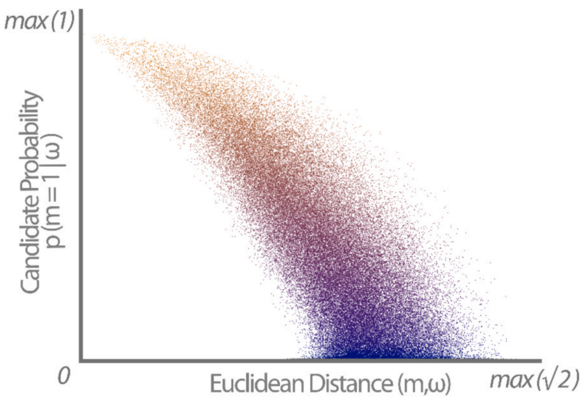

Fig. 8. The inverse relationship between Euclidean distance and the probability that a memory is a candidate for recall. Each of the 10^6 point depicts the Euclidian distance and candidate probability of a randomly sampled memory and stimulus. Each memory and stimulus consist of 10 features, five of which were always randomly selected to have a value of zero. The salience-values of the five remaining features had a total salience summed salience of one, with their values drawn from an exponential distribution (the results are robust to different set values of total features and non-zero features, as well as to the drawing of uniformly distributed salience values). The resulting scatterplot depicts the inverse relationship between Euclidian distance and the probability that a memory is selected for recall. Note that the wave-like shape reflects that there is only one way in which two events can be identical, but there are multiple ways in which two dissimilar events can differ from each-other. The code for this simulation can be freely found online at https://osf.io/fcqhr/.

is recalled. If we constrain the recall process to its initial recall step (cued recall), then memories are stochastically selected to be candidates in line with equation (1):

$$p(C_i^{t=1} = 1) = \sqrt{m_i\omega}$$

If we focus on a single memory m and drop the notation for i, then we can rewrite this equation as:

$$P(m \text{ is a candidate}|\omega) = \sqrt{m\omega} = \sqrt{\sum_{i\in\{1,F\}} (m_f{}^*\omega_f)}$$

This candidate-selection probability is inversely related to the distance between the memory and the stimulus. This becomes obvious when considering the two possible extreme scenarios for the (dis)similarity between the stimulus and m. If the stimulus and memory are entirely dissimilar, then there is no overlap in features with a non-zero salience. In this case it holds that for each feature f either $m_f = 0$ or $\omega_f = 0$, and correspondingly the candidate-probability of m is equal to zero. In contrast, consider the opposite extreme where the stimulus and memory are maximally similar, that is, when $\omega = m$. If the stimulus and memory m are identical, then the corresponding candidate-probability of m is equal to 1. As a further illustration of this inverse relationship, Fig. 8 depicts the distance and the candidate-probability for 10^6 randomly sampled pairs, where each point represents a single comparison. Just as in similarity-based models, the cued-recall candidate probability of a memory m is inversely proportional to its distance in feature-space.

Despite this overlap there are two key conceptual differences between similarity-based models and the cascading recall process. First is the continuous versus discrete nature of memory weighting. Similarity-based measures such as Euclidean distance typically yield a continuous value. Therefore, even highly dissimilar memories have a (limited) impact on the decision process. In contrast, memory retrieval is binary in the cascading sampling framework. During each recall step, only a single memory is retrieved, and memories which are not retrieved do not impact the final decision process. Consequently, a smaller fraction of memories influences the decision in the cascading process.

The second, and most substantial, differences between similarity-based evaluation and cascading retrieval lies in the consecutive nature of memory retrieval, whereby the reinstated features of a retrieved memory form the input for drawing the subsequent sample from memory. After the first memory step (cued recall), these reinstated features may not share any overlap with the observed stimulus. Thus, in contrast to similarity-based models, the cascading recall process allows for the retrieval of memories which are not directly similar to the observed stimulus.

4.1.2. As in limited sampling models, decisions in the cascading recall process are based on a limited number of sampled past experiences

Over a lifetime we collect a vast number of episodic memories, the vast majority of which are irrelevant for any one decision in particular. It would therefore be massively computationally inefficient to search through this haystack of mostly irrelevant past experiences in order to retrieve the needle of a few relevant memories. An alternative search strategy is to forgo this exhaustive search, and to instead only select a small part of the haystack to search and then generalize this limited sample of past events to behavior in novel situations.

Such limited sampling of past events is consistent with computationally tractable models for various inferential cognitive processes (Sanborn & Chater, 2016; Vul et al., 2014; Zhu et al., 2020), as well as with the literature outlined in principle 3. In previous work, the limited sampling of past experiences has been proposed as a mechanism to explain a wide range of observed behavioral regularities (Erev et al., 2023; Stewart et al., 2006). These regularities include the difference in behaviour between decisions-from-description and decisions-from-experience (Biele et al., 2009; Hertwig & Erev, 2009), our propensity for probability-matching in risky choices (Barron & Erev, 2003), the hyperbolic discounting of future rewards, the diminishing marginal utility of wealth, loss aversion, the overweighting of small probabilities and underweighting of large probabilities (Stewart, 2009) and the wavy recency effect of experiences (Plonsky et al., 2015).

The cascading recall process is consistent with such a limited-sampling strategy. Just as in these previous models, decisions in the cascading recall framework are based on a small number of retrieved episodes. The cascading recall process builds upon this literature by proposing that episodes are more likely to be retrieved if they share an overlap to the observed stimulus (cued recall), *or* those which share an overlap to the previously retrieved memories (cascading recall). As a consequence, the cascading recall process forms an adaptive approach to sampling; instead of drawing episodes at random, it tends to retrieve experiences closer in feature-space to the stimulus.

We illustrate this point by simulating a range of decisions in a ‘slice’ of reality. This slice contains 10 possible features, and each event has a non-zero salience value for 3 randomly selected features. In each simulation the decision-maker observes a stimulus and searched her memory for *a sufficiently informative* memory; that is, memories which provided sufficient evidence in favor of an action. For illustrative reasons we impose that the probability of a memory being sufficiently informative is inversely proportional to the distance between the memory and the stimulus in memory-space, according to:

$$p(\text{memory } m \text{ is informative}\,|\,\omega) = p(A_m = 1\omega) = \left(1 - \frac{dist(m,\omega)}{\sqrt{2}}\right)^2$$

The $\sqrt{2}$ represents the maximum possible distance in feature-space and is used to normalize the range of allowed probabilities to [0,1]. All these parameters are picked arbitrarily, and the results of the simulations are robust to a wide range of input parameters.

We simulated the outcomes of four decision processes. The first of these is the cascading recall process, followed by three comparison models: a cued-only recall model and two random-sampling models. The cued-only recall model retrieves one memory per step in line with equation (2). This model is therefore analogous to the above-described similarity-based models – albeit with the difference that the cued-only model retrieves at most a single memory in a binary fashion (versus the continuous weighting in similarity-based models). The search process is terminated if this retrieved memory is not sufficiently informative. Upon termination, both the cued-only and cascading recall models stochastically restarted the decision process with a 75% probability, or abandoned the search and selected a random action with a 25% probability. The cued-only and cascading decision processes were counted as a success only if an informative memory was retrieved before the search was abandoned. We next included two random-sampling models. The first of these sampled 10 episodes at random. The random strategy was counted as a success if at least one of the memories contained a memory providing sufficient information. The second strategy sequentially retrieved episodes at random (without replacement) until a sufficiently informative memory was retrieved.

We illustrate how these decision-processes unfold as the decision-maker becomes more familiar with the task environment, that is, as more and more memories become available for sampling. Fig. 9 provides the results for a total of 250 randomly generated networks and stimuli. In order to reduce noise and artifacts, each decision was repeated 100 times per strategy. All tested strategies drew on a limited sample of all available memories, and the percentage of memories sampled decreased as the number of memories in **M** increased. Nonetheless, there were significant differences between both the number of memories retrieved, their (dis)similarity to the observed stimulus, and the rates at which the strategies yielded the selection of an action.

As compared to a retrieval process exclusively based on cued recall, the cascading recall process retrieved a larger number of episodes per decision (Fig. 9**A**). These sampled episodes were on average slightly less similar to the observed stimulus (Fig. 9**B**), which increased the probability that the cascading recall process retrieved a sufficiently informative memory, and requires fewer search iterations to do so than cued recall alone (Fig. 9**C**) As the number of memories in **M** increased, this wider (more dissimilar) sampling allowed the cascading recall process to retrieve a sufficiently informative memory in all simulations, whilst the cued-only process reached a maximum success-rate of around 75 % (Fig. 9**D**).

While the cued-recall process under-sampled past experiences, the opposite holds for the two random sampling strategies. Both the 10-at-random and the sequential sampling strategies retrieved substantially more memories than either the cued model or the cascading recall process. These retrieved memories were on average substantially less similar to the stimulus, as compared to the memories which were retrieved by either the cued-only or cascading recall processes. Whereas the limited decrease in the similarity of sampled memories allowed the cascading recall process to flexibly retrieve a relevant experience, the more substantial decrease in similarity for the two random-sampling strategies backfires and results in the over-sampling of uninformative past experiences. In terms of success-rates, the 10-at-random strategy therefore performed worse than both the cued and cascading recall processes. As the sequential sampling model retrieved episodes until a sufficiently informative memory was sampled, it was thus guaranteed to (eventually) retrieve a sufficiently informative memory if one existed, and would search through all episodes if none were sufficiently informative. However, it did so at the cost of sampling a substantially larger number of memories than the cascading recall process.

4.1.3. Cascading retrieval is consistent with previous models of hippocampal generalization – And extends them to situations were only a limited number of memories are sampled

Although thus far undiscussed, neural plausibility is an additional consideration for memory-based decision-making. While the approaches outlined above have greatly contributed to our understanding of memory-based evaluative processes on a behavioral level,

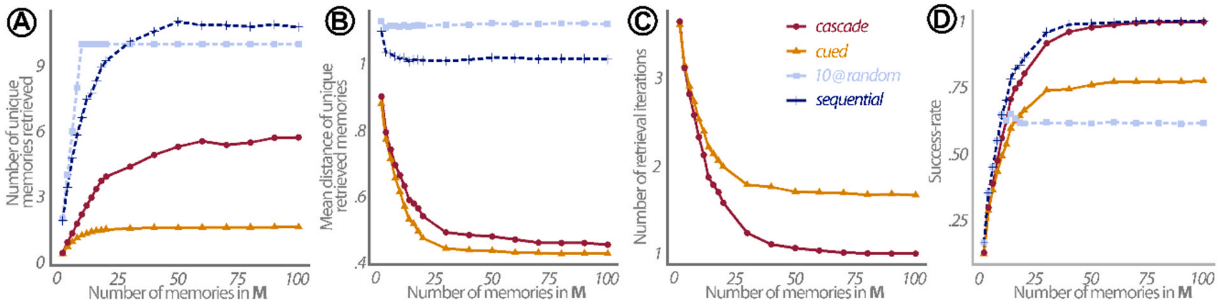

Fig. 9. Simulations comparing cued and cascading recall with limited-sampling models. Each point represents the average of 250 simulations, each of which contained 100 decisions for a total of 25,000 decisions. We varied the memory size to illustrate how the sampling processes evolves as the number of memories in **M** increases (x-axis). All simulations consisted of 10 different possible features, and each memory contained two features with a non-zero salience value. The salience values of these memories were drawn from an exponential distribution. We assumed a patience of 0.75 for both the cued and cascading retrieval process. These parameters were arbitrarily selected, and all results are robust to a wide variation of parameter values. Within a simulation, all decisions and all strategies were completed given the same stimulus. **(A):** the number of unique memories which were evaluated (either retrieved or sampled) during any of the four sampling strategies (lines). **(B)**: the mean Euclidean distance between the evaluated memories and the observed stimulus, per strategy. **(C)**: the number of recall processes undertaking by the cued and cascading recall processes before an action was selected. **(D)**: each memory had a probability of being informative (that is, a probability that it supported the selection of an action). This probability was inversely proportional to the memory's Euclidean distance to the stimulus. An iteration of the retrieval process was counted as a success if at least one sufficiently informative memory was retrieved. The code for these simulations can be freely found online at https://osf.io/fcqhr/.

these lines of research have predominantly been disconnected from the rich and rapidly growing literature on the neural underpinnings and functioning of the episodic memory system, broadly centered around the hippocampus. Insights into the functional organization of the episodic memory system have formed the basis of models of hippocampal generalization, which draw on these recent insights to propose a framework to capture the generalization of past experiences to novel events (Kumaran & McClelland, 2012a; Schapiro et al., 2016; Schlichting & Preston, 2015). A key element in these models is a process of recurrent similarity, whereby repeated interactions between the neocortex and hippocampus result in the generalization of single-shot past events to novel present contexts.

Consistent with our work, this generalization is based on a recurrent connection between conjunctions (similar to our 'memories') and features. The key difference between this recurrent activation and the cascading sampling of episodes lies in the former's inclusion of lateral inhibition. In previous work, the activation of a conjunction has been assumed to decreases the probability of activation for all other conjunctions. This inhibition prevents the model from reaching an over-excited state, that is, from reaching a runaway process in which *all* features and memories become activated. Instead, as activated conjunctions inhibit other conjunctions, the most-activated conjunctions accumulate activation as the decision-process unfolds. This accumulation allows the system to reach a steady-state, resulting in the selection of the action previously associated to the dominant conjunction.

However, the schism between models of hippocampal generalization and limited-sampling models has been two-sided; just as models of similarity-based weighting and limited-sampling models have remained agnostic on their implementation on a neural basis, so too have models of hippocampal generalization been agnostic on the limited retrieval of past experiences, or the increased impact of similar memories on the decision process. The cascading recall process bridges this gap.

In contrast to the assumption of lateral inhibition in models of hippocampal generalization, in the cascading recall process only a single memory can be recalled per step. As a result, in line with the research outlined above, only a limited number of past experiences are retrieved. This limited sampling prevents the memory retrieval process from reaching an over-excited state, and allows the cascading process to retrieve a single action without the need of the entire system to reach a steady state. Thus, by replacing lateral inhibition and the achievement of a steady state for the retrieval of a single memory per step, the cascading recall process can be broadly considered as a generalization of recurrent models to situations where a limited number of episodes have to be sampled from a large collection of (mostly) irrelevant experiences. By doing so, the cascading recall process is consistent with, and extends upon, models of hippocampal generalization by bridging them with models of limited sampling.

4.1.4. Cascading recall is consistent with models of associative recall and extends upon them by allowing the retrieval of experiences which do not overlap with the observed stimulus

Our ability to retrieve previously learned items is enhanced when we experience the same environmental context as was present when we originally observed the item (Smith & Vela, 2001). Various models have previously been proposed to describe the cognitive processes which underly such context-dependent retrieval of items following the presentation of a cue. Of particular interest are several models which represent these items as a set of memory-vectors, with each vector containing the to-be-recalled item and a set of features representing the context in which these items were observed. Below we discuss three prominent examples of these models. In order to compare these models to the cascading retrieval process, the term '*items*' can be considered to be a functional analog of our use of the term '*memory*'. In all three models, the cascading recall process builds on this literature by proposing a set of cognitively plausible mechanisms to support the retrieval of past experiences, as well as extending these models to allow for the retrieval of episodes which do not share features with the observed stimulus (in line with the results of experiments 2 & 3).

In the *Search of Associative Memory* (SAM) model (Raaijmakers & Shiffrin, 1981), the observation of a stimulus triggers the retrieval of past items in a two-step process. Items are first moved into short-term memory if their context-at-observation is sufficiently similar to the context observed in the present. Items are then sampled from short-term memory in a second step to guide recall. The SAM model is analogous to the first step of the cascading retrieval process (cued recall); the observation of the stimulus activates a limited number of memories related to the stimulus (either in 'short-term memory' in the SAM model, or via the candidate-selection function in the cascading retrieval process). Some of these memories are then selected for retrieval. The cascading retrieval process adds to this model by allowing multiple sequential retrieval steps, whereby the features of a recalled memory guide the retrieval of a subsequent memory. In contrast, the SAM model does not facilitate such sequential recall as all retrieved items share an overlap to the stimulus. Thus, by introducing the pattern-completing process, the cascading retrieval process extends upon the SAM network by allowing the cascading retrieval of experiences which do not share any features with the observed stimulus.

In the *Theory of Distributed Associative Memory model (TODAM;* Murdock, 1982) memories are retrieved through the convolution of the stimulus vector and all memory vectors – just as in the cascading recall process. However, the TODAM model does not conceptualize each memory-vector as a set of individual features, but instead as a vector of random variables. The resulting convolution between the stimulus and the memory matrix yields a composite vector. This composite vector has one element for each random variable, and the values of this vector represents the average activation of this variable across all memories which overlap with the stimulus vector. The memory with the highest similarity to this composite vector is then retrieved. The TODAM-2 model (Murdock, 1993) further includes a process of serial retrieval, whereby the retrieved composite vector is used to cue a subsequent retrieval process. The cascading retrieval process bridges the TODAM models with insights in the functional architecture of the episodic memory system which have been developed in the years since the TODAM models were originally published, as described in the six principles outlined above. In line with these findings, memories in the cascading recall process are retrieved holistically, and their reinstated features form the input for subsequent retrieval. It is this pattern-completing reinstatement which allows the cascading retrieval process to retrieve experiences which do not share any features with the observed stimulus.

As in the similarity-based models discussed in the previous section, the MINERVA (Hintzman & Ludlam, 1980) MINERVA-2

(Hintzman, 1984) and MINERVA-DM (Dougherty et al., 1999) models evaluate the similarity between the stimulus and all memory-vectors (dubbed 'traces') in parallel. Each trace is activated proportionally to this similarity, and the summed (similarity-weighted) activation of all memories yield a composite output vector, dubbed the 'echo'. This echo is the output of the retrieval process and can be used to infer the familiarity of the observed stimulus (represented by the total magnitude of the echo). As is the case with the TODAM model, the difference between the MINERVA models and the cascading retrieval process lies in the nature of the retrieved experiences. The echo in the MINERVA models and the composite vector in TODAM models yields a composite vector, which is comprised of many episodes. In contrast, the cascading retrieval process builds on recent insights in the architecture of the episodic memory system by proposing that memories are retrieved individually and holistically. Once retrieved, a memory triggers the subsequent activation of additional memories not by activating a composite vector or echo, but instead reinstating its constituent features through a pattern-completing process.

4.2. Beyond episodic memory retrieval: Cascading retrieval differs from (but is not mutually exclusive with) models of categorization and the learning of abstract knowledge structures

We have thus far discussed the retrieval of episodic memories in isolation from non-episodic knowledge. However, we do not mean to imply that memory search is independent from these alternative knowledge structures. Throughout a lifetime we collect a large number of episodic memories. The efficient retrieval of relevant memories depends on our ability to categorize these experiences into related groups, as well as the extraction of communalities across *multiple* episodes. While an exhaustive discussion of the interplay between episodic memory retrieval and additional knowledge structures is beyond the scope of the current work, below we provide a broad overview of the overlap and distinction between these organizational structures and the cascading retrieval process.

4.2.1. Exemplar models – Grouping multiple memories in different categories

Past experiences can be categorized into groups, and these groups can be accessed through the retrieval of exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986, 1988). In such exemplar models the observation of an item (for the current discussion analogous to our use of "memory") triggers the retrieval of an associated category. That is, exemplar models infer the associated category of a novel item based on the similarity between the observed stimulus and a set of pre-categorized exemplars. As in the cascading recall process, these exemplars are instances of past experiences. Consequently, exemplar models do not require abstract representations. Consistent with this retrieval of individual observations, the retrieval of exemplars has previously been associated to increased activation in the hippocampus (Bowman & Zeithamova, 2018; Mack et al., 2013; Sučević & Schapiro, 2023).

Despite this conceptual overlap to the cascading recall process, two key differences exist. First, in order to facilitate the learning of (multiple) groups of experiences, experimental tasks to validate exemplar models typically provide participants with large number of annotated observations – that is, pairing many observations with unambiguous category labels. These labels are paramount to facilitate the learning the underlying categories – after all, it is these labels which define the categories. In the current work we did not explore the role of exemplars in episodic retrieval, but instead aimed to mitigate the influence of exemplar-based learning strategies. In the current setup we therefore minimized the possibility that participants relied an exemplar-based learning strategy by providing them with a limited number of stimuli which did not contain category labels.

More broadly, exemplar models rely on the (feature-based) similarity between the stimulus and previously observed items. Exemplar models are therefore based on an analogous foundation to the above-discussed similarity-based models – albeit with a focus on the learning of categories instead of decision-making. In order to infer a category of a novel stimulus, exemplar models rely on the

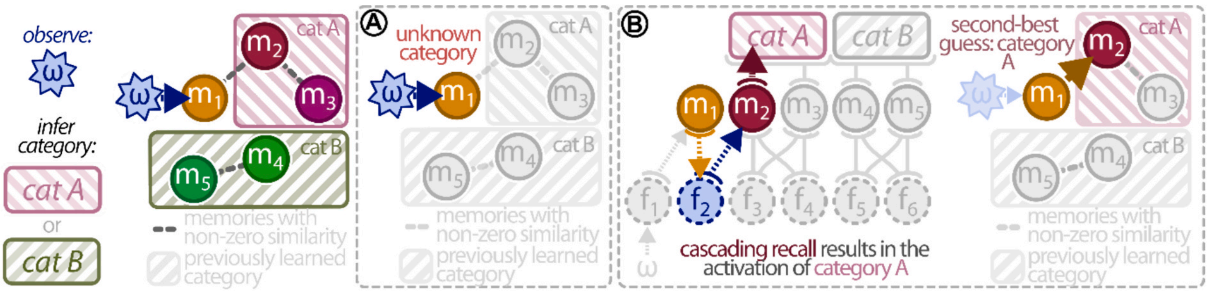

Fig. 10. Cascading recall and exemplar models. In this example, the subject observes a stimulus (ω) and is tasked with inferring its associated category: category A or B. The subject has access to five relevant memories, two of which (m_2 and m_3) are exemplars of category A, two others (m_4 and m_5) are exemplars of category B, and a single un-categorized experience (m_1), which shares overlapping features with m_2. The stimulus exclusively shares a feature with m_1. **(A)**: In exemplar models, the associated category of ω is based on the similarity between the observed stimulus and each experience. However, in this example the observed stimulus only shares overlapping features with m_1, which does not belong to any category. Consequentially, retrieval of m_1 does not facilitate the inference of a category for the stimulus **(B)**: the cascading recall process extends on exemplar models by allowing categories to be based on indirectly related experiences. In this example, the observation of ω triggers the cued recall of m_1. While this episode does not have a clear category label, it next reinstates its constituent features, resulting in the retrieval of episode m_2, which does contain a clear category label. Consequently, the cascading recall process results in a second-best guess of which category best fits the observed stimulus.

similarity between this stimulus and previously categorized exemplars – but do not provide an inference when none of these similar items belong to any category. As in similarity-based models, the cascading recall process extends on exemplar models by allowing categories to be inferred based on indirectly related experiences. In line with equation (4) (and supported by the findings of experiments 2 and 3), the observation of a stimulus may trigger the cued retrieval of a past experience. If this experience does not clearly belong to a category, then its reinstated features subsequently triggers the retrieval of a related memory, which in turn may contain a clear category label (Fig. 10). Although future research is required to further explore this prediction, cascading retrieval can broadly be considered an extension of exemplar models to situations where only a limited number of relevant observations are available. In these circumstances, cascading recall allows for an 'second-best' guess on memory categorization; inferring a category which does not strictly fit the observed stimulus or any directly similar experiences, but which is nonetheless indirectly associated to the stimulus.

4.2.2. *Temporal context model – Retrieving memories via a gradual drift through time*

Memories can also be organized based on their time of occurrence. In the Temporal Context Model (TCM; Howard & Kahana, 2002), the retrieval of an item (for this discussion analogous to our use of 'memory') triggers the reinstatement of the temporal context in which the item was originally observed. This activated temporal context subsequently enhances the recall probability of all items which were observed during this period in time. In the years since its original formulation, several findings support such temporal organization of memories: hippocampal representations have been demonstrated to store the temporal context of an experience (Manns et al., 2007), with hippocampal *time cells* (Eichenbaum, 2017b) providing the temporal signature of experiences (Eichenbaum, 2017a). While conceptually similar to the associative context models discussed in the previous section, the TCM differs from these associative frameworks by introducing a slow-moving temporal drift. In the TCM, the activation of a temporal context may gradually shift to the activation of temporally adjacent contexts. Empirical findings are also consistent with this theorized drift, as adjacent temporal contexts are represented by overlapping populations of neurons (Eichenbaum, 2014).

The cascading retrieval process can account for such a gradually drifting temporal context. Feature-nodes are modality-agnostic in the cascading retrieval process, as these features comprise the *where* and *what*, but also the *when* of the retrieved experience. Some features may therefore represent the temporal context in which events are observed, such that different periods in time (contexts) are represented by different features. Such time-features represent a gradually evolving temporal context with a non-linear resolution; the relative ordering of memories loses precision as time passes (Howard & Eichenbaum, 2013). These time-features can be readily adapted to include the mechanism at the core of the TCM, with the only adaptation required being the inclusion of bidirectional connections between adjacent temporal contexts. These connections between two adjacent features reflect their partially overlapping neural populations (Fig. 11**A**). As before, the recall of a memory reinstates all its constituent features, including the memory's temporal context. The activation of this feature increases the probability of subsequently retrieving another memory from the same temporal context (Fig. 11**B**). Moreover, the bidirectional link between adjacent contexts allows for a temporal drift as the activation of a temporal features projects activation to the features reflecting adjacent moments in time (Fig. 11**C**). As such, the cascading retrieval process is consistent with – and provides a mechanism for – the retrieval of memories following a temporal context, as well as the drift of the temporal window.

However, the cascading retrieval process differs from temporal context models by additionally including the subsequent retrieval of memories based on non-temporal features. This results in two differences between cascading retrieval and temporal context models. Subsequent memories may be retrieved even when they occurred in distinct (non-adjacent) temporal contexts, while memories within the same context are more likely to be retrieved if they also share overlapping non-temporal elements. As an example of the former, two memories may contain overlapping features but have taken place in different (non-adjacent) temporal contexts. The recall of one

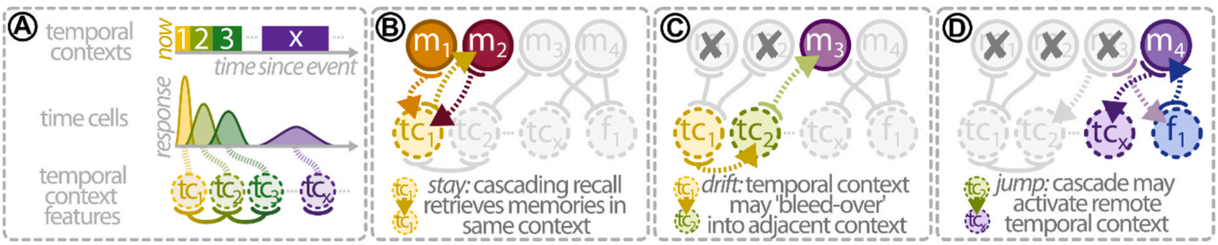

Fig. 11. Temporal context and cascading recall. The cascading retrieval process is consistent with key predictions of the temporal context model. **(A)**: Temporal contexts represent non-linear periods in time, with a decreasing temporal resolution (longer windows) as more time has passed since the event was observed. These temporal contexts are represented by hippocampal time-cells. These cells have a partially overlapping sensitivity, such that their temporal borders overlap. We represent this partial overlap in sensitivity by including a bidirectional link between temporally adjacent context-features, whereby the activation of a context increases activation for the immediately past and prior contexts. **(B)** The cascading retrieval process allows for the sequential retrieval of memories which were observed in the same temporal context ('staying in the same context'). In this example the activation of memory m_1 triggers the reinstatement of its constituent temporal context tc_1. Activation of this context triggers the retrieval of m_2, which was observed in the same temporal context. **(C):** the bidirectional connections between adjacent temporal contexts allows for a gradual drift of the active temporal context. In this example, the cascading recall process continues when tc_1 activates its temporal neighbor tc_2, which in turns triggers the retrieval of memory m_3. **(D)** The activation of non-temporal features may result in jumps between non-adjacent temporal contexts. In this example, m_3 reinstates its constituent features, which includes f_1. The activation of this non-temporal feature next triggers the activation of memory m_4, which was previously observed in a remote temporal context (tc_x).

memory may next result in the retrieval of the other memory – causing a non-adjacent jump in the activated temporal context (Fig. 11**D**). An example of the latter is provided by our experiment results. In experiments 2 and 3, we randomized the order of presentation for all training-phase Fennimals. Our observed results are therefore difficult to explain through a temporal drift alone, as the temporal context of the training phase would, on average, not predict the retrieval of one Fennimal over any other. Instead, it is indirect overlap of the non-temporal features which guided participant's decisions: the stimulus triggered the recall of a memory with an overlapping feature (cued recall), the reinstated features of which then triggered the (cascading) recall of the target Fennimal.

4.2.3. Models based on abstract representations –regularities learned across multiple episodes

Just as in the cascading recall framework, exemplar models and the TCM consists of two layers of nodes: features and memories. But what about non-episodic layers? Various models and findings support the retrieval of episodes along abstract contextual or semantic groups (Tulving & Pearlstone, 1966; Wickens, 1970). Indeed, long-standing findings support our ability to learn abstract groupings of features across multiple dimensions (Ashby & Townsend, 1986; Murphy & Ross, 1994; Nosofsky, 1986) and that statistical regularities can be extracted across multiple experiences (Ashby & Maddox, 2005). Once these abstract representations have been learned, they guide the retrieval of memories during subsequent decisions (Humphreys et al., 1989; Osth & Dennis, 2015; Polyn et al., 2009; Zhang et al., 2021). Such semantic representations have been proposed to form a complementary learning system (Kumaran & McClelland, 2012; McClelland, McNaughton, & O'Reilly, 1995; Mcclelland, Mcnaughton, & Reilly, 1995), which is neurologically centered around the prefrontal cortex (Mcclelland et al., 1995; O'Reilly et al., 1998).

A wide category of models based on abstract representations can be loosely centered around associative or contextual recall, including models based on the sampling of items through a random walk in semantic space (Hills et al., 2008, 2012; Todd & Hills, 2020) and models which extend on the temporal context model by including a semantic context (Lohnas et al., 2015; Polyn et al., 2009) Although their exact implementations differ, these associative models are based on a process whereby the recall of an item (for this discussion analogous to our use of "*feature*") activates an abstract representation associated to this item. In turn, this activated representation enhances the recall of all items which are also associated to the representation. Just as the cascading recall process, these models can account for the retrieval of indirectly related items (Howard et al., 2009); a stimulus A may trigger the activation of an item B. Item B then activates its abstract grouping, which in turn increases the probability of next recalling a item C which is also connected to this representation. Just as in cascading recall, this retrieval of C after A may occur even when there is no direct similarity between A

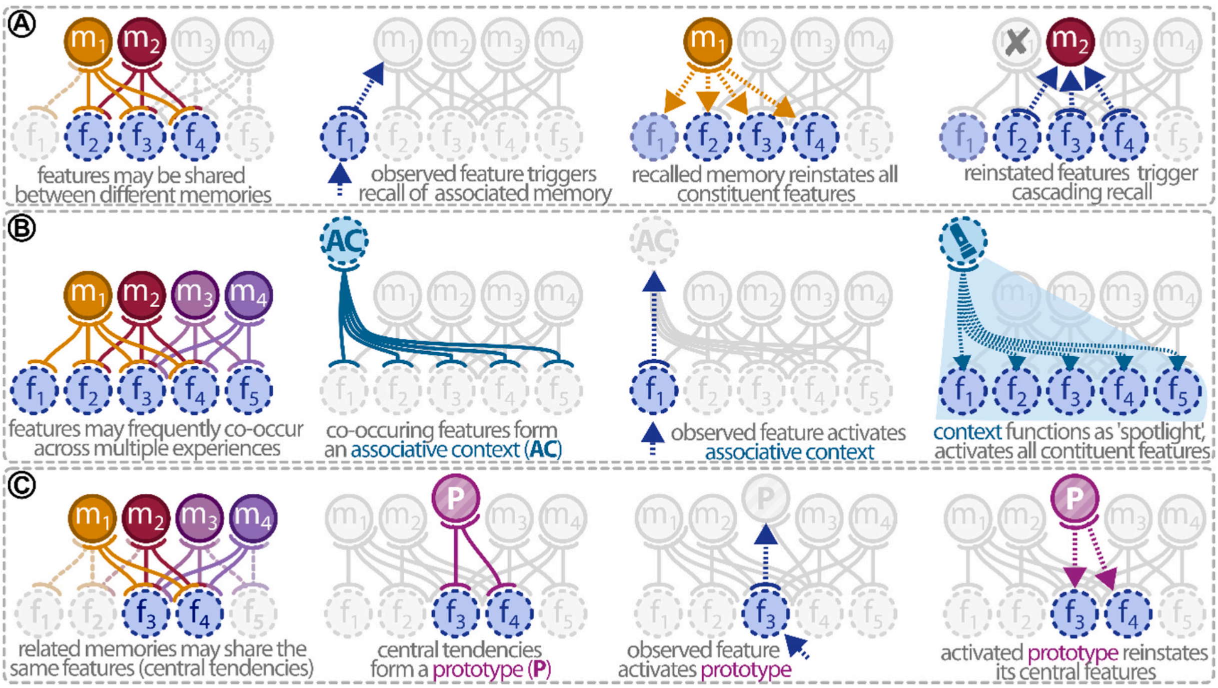

Fig. 12. Cascading recall versus abstract representations and prototype models.(A): the cascading recall process is based on the sequential recall of perceptually rich episodes. The activation of a memory reinstates its constituent features, which in turn trigger a subsequent recall process. **(B)**: a schematic overview of associative context models. Across multiple experiences, some features may frequently co-occur together. These co-occurring features are bound in the same semantic context. These contextual nodes represent a second (non-episodic) organizational system, separate from the episodic memory system. Upon activation, a feature triggers the activation of any associated context nodes. These activated contexts then function as a spotlight, increasing the activation of all constituent items. **(C)**: a schematic overview of prototype models. A set of related experiences may share the same constituent features. These shared features form the central tendency of the group, and are stored as a separate (non-episodic) prototype. The activation of a feature triggers the reinstatement of any related prototypes. Once activated, prototypes reinstate their central features.

and C.

However, there is a key conceptual difference between the cascading recall process (Fig. 12**A**) and models based on abstract representations (Fig. 12**B**). In models based on abstract representations, the recall of subsequent episodes is guided by an overarching organization between groups of features. Typically, these items are conceptualized as lists of to-be-remembered words, and predictions of this model are tested by the subsequent recall of words which have a similar semantic grouping (semantic context). Alternative implementations have been demonstrated which do not rely on the retrieval of listed words (Matute et al., 2011), but these implementations model the (statistical) learning of different contexts following multiple observations and a process of trial-and-error learning. In contrast, the cascading recall process is centered on the sequential recall of perceptually rich episodic events based on a process of feature-based reinstatement. In this cascading process, it is the reinstated features of the episode which trigger the retrieval of subsequent memories – without the activation of an abstract semantic context.

The results of Experiments 2 and 3 provide an empirical differentiation between models based on abstract representations and the cascading recall process. In these experiments, participants were more likely to select a toy for an unfamiliar Fennimal (the stimulus) by drawing on their past experience with an indirectly related training-phase Fennimal. In these experiments, the between-subjects randomization of different features (heads and colors) for the training-phase Fennimals excludes a possible semantic context effect: the indirectly retrieved Fennimal did not share a semantic context with the observed novel Fennimal. Thus, while our results are consistent with a cascading recall process, they are more difficult to explain via abstract semantic groupings.

4.2.4. *Prototype models – Memory-like abstractions, but not memories*

Related to, but conceptually distinct from, associative context models are models whereby abstract categories are learned through the formation of abstract prototypes (Homa & et al, 1973; Posner & Keele, 1968; Reed, 1972). These prototypes contain the central tendencies of a category; that is, the average features of the elements which comprise the category (Fig. 12**C**). While early work considered prototype models to be in competition with exemplar models, work in recent decades has increasingly illustrated how both prototypes and exemplars fill different cognitive niches and are based on different neural strata (Bowman et al., 2020; Zeithamova & Bowman, 2020), with prototypes being increasingly linked to the same neural regions associated to associative context models (Mack et al., 2020; Zeithamova, Dominick, et al., 2012).

As in models of abstract representations (but in contrast to exemplar models, models based on an associative context and the cascading recall process) prototypes are not instances of previously observed experiences, but instead represent commonalities which are learned across multiple experiences. In order to be able to extract these central tendencies, prototype models require that each category contains multiple experiences. In our experiments, participants' experiences with the Fennimals were too sparse to form such central tendencies. During the training phase, participants encountered four Fennimals, and had to draw on these experiences to inform their behavior in the search phase. If all training-phase Fennimals were categorized in the same group, then the prototype of this group would reflect all training-phase Fennimals – and therefore would not explain why participants selected the toy based on the Fennimal which was (in)directly similar to the observed new Fennimal in the search phase. Alternatively, in experiments 2 and 3 participants could have categorized Fennimals based on their overlapping features. However, such a category would result in groups consisting of at most two Fennimals – and their prototypical central tendency consisted of a single feature which was shared between these Fennimals. As this shared feature was not contained by the novel search-phase Fennimal, it remains unclear how a prototype model would have yielded the observed results.

This is not to say that prototype models do not shape the memory retrieval process in general. Previous literature suggests that prototypes are a consistent categorization scheme across a wide range of experimental tasks. The cascading recall process adds to this literature by proposing a mechanism through which prototypes can be retrieved even if an observed stimulus does not directly share features with the prototype itself. As in exemplar models, the cascading recall process may retrieve memories which are contained within a prototypical category – and once active, these prototypes may guide the subsequent retrieval process. This interplay between these protypes and the memory search process may prove a fertile ground for future research.

4.3. *Limitations and future directions*

While the current work is the first to take a first-principles approach to the merging episodic sampling in memory-based decisions with insight into the architecture of the episodic memory system, these experiments are subject to a number of limitations. Our focus in these experiments was to provide an empirical validation for the unique predictions of the cascading recall process in a minimalist paradigm. In experiment 1 we tested the predictions of this cascading recall process against an alternative process of random sampling. However, in order to minimize the complexity of our task, experiments 2 and 3 did not provide a clear counter-model. That is, while these experiments provide empirical support for the cascading recall process, they do not directly test the predictions of the cascading recall process *against* alternative models or decision strategies. Thus, while the present approach offers a proof-of-concept for both the experimental paradigm as well as the cascading recall process, future research efforts may be conducted to specifically test the cascading recall process against alternative predictions.

In addition, in this work we strictly explored the behavioural predictions of our proposed cascading recall process. Consequently, we did not test the implicit assumption that this process is consistent with the neural functioning and architecture of the medial temporal lobe. However, there is circumstantial evidence that this is indeed the case. The recollection of a memory has previously been observed to prime the hippocampus to activate additional memories (Duncan et al., 2019; Duncan & Shohamy, 2016; Patil & Duncan, 2018). Consistent with the predictions of a cascading process, these additionally activated memories may include memories which are only indirectly related to an observed stimulus (Duncan et al., 2019). Moreover, this cascading retrieval process is consistent with an autocorrelated sampling process whereby earlier sampled memories influence which additional memories are sampled later in time (Bornstein & Norman, 2017). However, while these findings are consistent with the predictions of the cascading recall process, these studies have lacked the overarching theoretical framework that we have developed in the current work. Thus, efforts to combine imaging methodologies with the theory-first approach developed in the present work may offer a fruitful avenue for future research.

There was also a lack of variance between observed outcomes, as all interactions with the Fennimals during the training phase resulted in the same positive feedback signal. In contrast, decisions in everyday life are commonly based on multiple sources of noisy and potentially conflicting signals. Previous work has extended models of sequential sampling (Busemeyer & Townsend, 1993; Ratcliff, 1978; Ratcliff et al., 2016; Usher & McClelland, 2004) to capture the integration of such conflicting signals during value-based (economic) decision-making (Hare et al., 2011; Krajbich et al., 2010; Krajbich & Rangel, 2011). While future research is needed, the interplay between sequential sampling models and cascading recall may be a useful direction for further exploration. While previous work has suggested *that* such signals may be drawn from episodic memory (Mack & Preston, 2016; Shadlen & Shohamy, 2016), it has thus far remained unclear *how* these memories are sampled. The cascading recall process suggests such a source mechanism. Promisingly, the predictions of cascading recall appear to be consistent with a previously observed increased deliberation time (Shadlen & Shohamy, 2016) and hippocampal activation (Bakkour et al., 2019; Bornstein et al., 2017; Gluth et al., 2015) for decisions with where a larger number of past experiences need to be sampled before the evidence favors any one alternative.

As time passes, the memory of an event is likely subject to distortion – either by a loss of the memory entirely or a misremembering of the features that constituted the past experiences. Neither the limited duration of our experiment design, nor our conceptualization of the cascading recall process, accounts for this transience. While this degradation does not negate a person's general ability to rely on indirectly related experiences, future research is needed to test the degree to which much older (and partially forgotten) memories can be retrieved during the decision process. While there is increasing interest in the mechanisms that underlie memory transience (Richards & Frankland, 2017), future research is needed to determine the likely intricate ways in which the (partial) forgetting or misremembering of memories shape the cascading recall process.

5. Concluding remarks

How do we draw on our past experiences to guide decisions in novel situations if cued recall alone does not yield a sufficiently informative memory? In this work we synthesized previous findings on the organization of the medial temporal lobe to support a formalized first-principles approach to understanding the retrieval of episodic memories. The resulting cascading recall process bridges a wide range of previously disjointed theoretical and empirical approaches to memory-based decision-making, and offers a computationally tractable bottom-up mechanism to sample past experiences. In turn, this cascading recall process provides new avenues for future research on memory-based decision-making.

CRediT authorship contribution statement

Achiel Fenneman: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sarah T. Malamut:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Alan G. Sanfey:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the comments, feedback and support that have contributed to various iterations of the present research. In particular, we thank both the reviewers and editor for their insightful comments, which have greatly improved the quality of our manuscript. We also particularly thank the institutions which supported a large fraction of this work: the Nijmegen School of Management and the Donders Institute at Radboud University Nijmegen, as well as the Faculty of Society and Economics at the Rhein-Waal University of Applied Sciences.

Data availability

The experiment software, raw data, analyses script and simulation code are freely available via the OSF link provided in the manuscript

References

Anderson, M. C., & Neely, J. H. (1996). Interference and Inhibition in Memory Retrieval. In *Memory* (pp. 237–313). Elsevier. https://doi.org/10.1016/B978-012102570-0/50010-0.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*(2), 154–179. https://doi.org/10.1037/0033-295X.93.2.154

Backus, A. R., Bosch, S. E., Ekman, M., Grabovetsky, A. V., & Doeller, C. F. (2016). Mnemonic convergence in the human hippocampus. *Nature Communications, 7* (May), 1–9. https://doi.org/10.1038/ncomms11991

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H. R., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *ELife, 8*, 1–28. https://doi.org/10.7554/eLife.46080

Ballard, I. C., Wagner, A. D., & McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-08998-1

Barron, G., & Erev, I. (2003). Small Feedback-based Decisions and Their Limited Correspondence to Description-based Decisions. *Journal of Behavioral Decision Making, 16*(3), 215–233. https://doi.org/10.1002/bdm.443

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience, 16* (10), 1492–1498. https://doi.org/10.1038/nn.3515

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology, 53*(3), 155–167. https://doi.org/10.1016/j.jmp.2008.05.006

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, Attention, and Choice. *The Quarterly Journal of Economics, 135*(3), 1399–1442. https://doi.org/10.1093/qje/qjaa007

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications, 8*(May), 1–9. https://doi.org/10.1038/ncomms15958

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience, 20*(7), 997–1003. https://doi.org/10.1038/nn.4573

Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *ELife, 9*. https://doi.org/10.7554/eLife.59360

Bowman, C. R., & Zeithamova, D. (2018). Abstract Memory Representations in the Ventromedial Prefrontal Cortex and Hippocampus Support Concept Generalization. *The Journal of Neuroscience, 38*(10), 2605–2614. https://doi.org/10.1523/JNEUROSCI.2811-17.2018

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459. https://doi.org/10.1037/0033-295X.100.3.432

Chadwick, M. J., Hassabis, D., & Maguire, E. A. (2011). Decoding overlapping memories in the medial temporal lobes using high-resolution fMRI. *Learning and Memory, 18*(12), 742–746. https://doi.org/10.1101/lm.023671.111

Cox, G. E., & Criss, A. H. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review, 127*(5), 792–828. https://doi.org/10.1037/rev0000195

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*(1), 180–209. https://doi.org/10.1037/0033-295X.106.1.180

Dragoi, G., & Buzsáki, G. (2006). Temporal Encoding of Place Sequences by Hippocampal Cell Assemblies. *Neuron, 50*(1), 145–157. https://doi.org/10.1016/j.neuron.2006.02.023

Dudai, Y. (2012). The Restless Engram: Consolidations Never End. *Annual Review of Neuroscience, 35*(1), 227–247. https://doi.org/10.1146/annurev-neuro-062111-150500

Duncan, K. D., Doll, B. B., Daw, N. D., & Shohamy, D. (2018). More Than the Sum of Its Parts: A Role for the Hippocampus in Configural Reinforcement Learning. *Neuron, 98*(3), 645–657.e6. https://doi.org/10.1016/j.neuron.2018.03.042

Duncan, K. D., Semmler, A., & Shohamy, D. (2019). Modulating the use of multiple memory systems in value-based decisions with contextual novelty. *Journal of Cognitive Neuroscience, 31*(10), 1455–1467. https://doi.org/10.1162/jocn_a_01447

Duncan, K. D., & Shohamy, D. (2016). Memory states influence value-based decisions. *Journal of Experimental Psychology: General, 145*(11), 1420–1426. https://doi.org/10.1037/xge0000231

Duncan, M., & Lewandowsky, S. (2005). The time course of response suppression: No evidence for a gradual release from inhibition. *Memory, 13*(3–4), 236–246. https://doi.org/10.1080/09658210344000233

Eichenbaum, H. (2014). Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience, 15*(11), 732–744. https://doi.org/10.1038/nrn3827

Eichenbaum, H. (2017a). On the Integration of Space, Time, and Memory. *Neuron, 95*(5), 1007–1018. https://doi.org/10.1016/j.neuron.2017.06.036

Eichenbaum, H. (2017b). Time (and space) in the hippocampus. *Current Opinion in Behavioral Sciences, 17*, 65–70. https://doi.org/10.1016/j.cobeha.2017.06.010

Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience, 3*(11), 1149–1152. https://doi.org/10.1038/80671

Erev, I., Ert, E., Plonsky, O., & Roth, Y. (2023). Contradictory deviations from maximization: Environment-specific biases, or reflections of basic properties of human learning? *Psychological Review, 130*(3), 640–676. https://doi.org/10.1037/rev0000415

Ergorul, C., & Eichenbaum, H. (2004). The hippocampus and memory for "what," "where," and "when.". *Learning and Memory, 11*(4), 397–405. https://doi.org/10.1101/lm.73304

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9*(1), 59–79. https://doi.org/10.3758/BF03196257

Garvert, M. M., Dolan, R. J., & Behrens, T. E. J. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *ELife, 6*, 1–20. https://doi.org/10.7554/eLife.17086

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology, 68*(1), 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation, 24*(1), 1–24. https://doi.org/10.1162/NECO_a_00226

Gilboa, A., Sekeres, M., Moscovitch, M., & Winocur, G. (2019). The hippocampus is critical for value-based decisions guided by dissociative inference. *Hippocampus, 29* (8), 655–668. https://doi.org/10.1002/hipo.23050

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics, 110*(3), 605–639.

Gluth, S., Sommer, T., Rieskamp, J., & Büchel, C. (2015). Effective Connectivity between Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from Memory. *Neuron, 86*(4), 1078–1090. https://doi.org/10.1016/j.neuron.2015.04.023

Goldstone, R. L., & Son, J. Y. (2012). Similarity. *In The Oxford Handbook of Thinking and Reasoning (Issue October)*. https://doi.org/10.1093/oxfordhb/9780199734689.013.0010

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591–635. https://doi.org/10.1016/S0364-0213(03)00031-4

Hare, T. A., Schultz, W., Camerer, C., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences of the United States of America, 108*(44), 18120–18125. https://doi.org/10.1073/pnas.1109322108

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences of the United States of America, 104*(5), 1726–1731. https://doi.org/10.1073/pnas.0610561104

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences, 11*(7), 299–306. https://doi.org/10.1016/j.tics.2007.05.001

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1263–1271. https://doi.org/10.1098/rstb.2008.0296

Henson, R. N. A. (1998a). Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5), 1162–1181. https://doi.org/10.1037/0278-7393.24.5.1162

Henson, R. N. A. (1998b). Short-Term Memory for Serial Order: The Start-End Model. *Cognitive Psychology, 36*(2), 73–137. https://doi.org/10.1006/cogp.1998.0685

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431–440. https://doi.org/10.1037/a0027373

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science, 19*(8), 802–808. https://doi.org/10.1111/j.1467-9280.2008.02160.x

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, &amp Computers, 16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L., & Ludlam, G. (1980). Differential forgetting of prototypes and old instances: Simulation by an exemplar-based classification model. *Memory & Cognition, 8*(4), 378–382. https://doi.org/10.3758/BF03198278

Homa, D., et al. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101*(1), 116–122. https://doi.org/10.1037/h0035772

Honey, R. C., Iordanova, M. D., & Good, M. (2014). Associative structures in animal learning: Dissociating elemental and configural processes. *Neurobiology of Learning and Memory, 108*, 96–103. https://doi.org/10.1016/j.nlm.2013.06.002

Horner, A. J., Bisby, J. A., Bush, D., Lin, W. J., & Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications, 6*(May). https://doi.org/10.1038/ncomms8462

Horner, A. J., & Burgess, N. (2013). The associative structure of memory for multi-element events. *Journal of Experimental Psychology: General, 142*(4), 1370–1383. https://doi.org/10.1037/a0033626

Horner, A. J., & Burgess, N. (2014). Pattern completion in multielement event engrams. *Current Biology, 24*(9), 988–992. https://doi.org/10.1016/j.cub.2014.03.012

Horner, A. J., & Doeller, C. F. (2017). Plasticity of hippocampal memories in humans. *Current Opinion in Neurobiology, 43*, 102–109. https://doi.org/10.1016/j.conb.2017.02.004

Howard, M. W., & Eichenbaum, H. (2013). The hippocampus, time, and memory across scales. *Journal of Experimental Psychology: General, 142*(4), 1211–1230. https://doi.org/10.1037/a0033621

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: Transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 391–407. https://doi.org/10.1037/a0015002

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*(2), 208–233. https://doi.org/10.1037/0033-295X.96.2.208

Iordanova, M. D., Good, M. A., & Honey, R. C. (2008). Configural learning without reinforcement: Integrated memories for correlates of what, where, and when. *Quarterly Journal of Experimental Psychology, 61*(12), 1785–1792. https://doi.org/10.1080/17470210802194324

Iordanova, M. D., Good, M., & Honey, R. C. (2011). Retrieval-mediated learning involving episodes requires synaptic plasticity in the hippocampus. *Journal of Neuroscience, 31*(19), 7156–7162. https://doi.org/10.1523/JNEUROSCI.0295-11.2011

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. *Nature Reviews Neuroscience, 16*(9), 521–534. https://doi.org/10.1038/nrn4000

Kahana, M. J. (2020). Computational Models of Memory Search. *Annual Review of Psychology, 71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Komorowski, R. W., Manns, J. R., & Eichenbaum, H. (2009). Robust conjunctive item - Place coding by hippocampal neurons parallels learning what happens where. *Journal of Neuroscience, 29*(31), 9918–9929. https://doi.org/10.1523/JNEUROSCI.1378-09.2009

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13*(10), 1292–1298. https://doi.org/10.1038/nn.2635

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences of the United States of America, 108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review, 119*(3), 573–616. https://doi.org/10.1037/a0028681

Leutgeb, J. K., Leutgeb, S., Moser, M. B., & Moser, E. I. (2007). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science, 315*(5814), 961–966. https://doi.org/10.1126/science.1135801

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review, 122*(2), 337–363. https://doi.org/10.1037/a0039036

Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *NeuroImage, 127*, 144–157. https://doi.org/10.1016/j.neuroimage.2015.12.015

Mack, M. L., Preston, A. R., & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology, 23*(20), 2023–2027. https://doi.org/10.1016/j.cub.2013.08.035

Mack, M. L., Preston, A. R., & Love, B. C. (2020). Ventromedial prefrontal cortex compression during concept learning. *Nature Communications, 11*(1), 1–11. https://doi.org/10.1038/s41467-019-13930-8

Manning, J. R., Polyn, S. M., Baltuch, G. H., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences of the United States of America, 108*(31), 12893–12897. https://doi.org/10.1073/pnas.1015174108

Manns, J. R., Howard, M. W., & Eichenbaum, H. (2007). Gradual Changes in Hippocampal Activity Support Remembering the Order of Events. *Neuron, 56*(3), 530–540. https://doi.org/10.1016/j.neuron.2007.08.017

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society B: Biological Sciences, 262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

Matute, H., Lipp, O. V., Vadillo, M. A., & Humphreys, M. S. (2011). Temporal contexts: Filling the gap between episodic memory and associative learning. *Journal of Experimental Psychology: General, 140*(4), 660–673. https://doi.org/10.1037/a0023862

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Mcclelland, J. L., Mcnaughton, B. L., & Reilly, R. C. O. (1995). *Why There Are Complementary Learning Systems in the Hippocampus and Neocortex : Insights From the Successes and Failures of Connectionist Models of Learning and Memory., 102*(3), 419–457.

McKenzie, S., Frank, A. J., Kinsky, N. R., Porter, B., Rivière, P. D., & Eichenbaum, H. (2014). Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron, 83*(1), 202–215. https://doi.org/10.1016/j.neuron.2014.05.019

McKenzie, S., Robinson, N. T. M., Herrera, L., Churchill, J. C., & Eichenbaum, H. (2013). Learning causes reorganization of neuronal firing patterns to represent related experiences within a hippocampal schema. *Journal of Neuroscience, 33*(25), 10243–10256. https://doi.org/10.1523/JNEUROSCI.0879-13.2013

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., Hefft, S., Merkow, M., Polyn, S. M., Jacobs, J., Kahana, M. J., & Schulze-Bonhage, A. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science, 342*(6162), 1111–1114. https://doi.org/10.1126/science.1244056

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*(6), 609–626. https://doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*(2), 183–203. https://doi.org/10.1037/0033-295X.100.2.183

Murphy, G. L., & Ross, B. H. (1994). Predictions From Uncertain Categorizations. *Cognitive Psychology, 27*(2), 148–193. https://doi.org/10.1006/cogp.1994.1015

Norman, K. A., & O'Reilly, R. C. (2003). Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary-Learning-Systems Approach. *Psychological Review, 110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 700–708. https://doi.org/10.1037/0278-7393.14.4.700

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus, 4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

O'Reilly, R. C., Norman, K. A., & McClelland, J. L. (1998). A hippocampal model of recognition memory. *Advances in Neural Information Processing Systems*, 73–79.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311. https://doi.org/10.1037/a0038692

Patil, A., & Duncan, K. D. (2018). Lingering Cognitive States Shape Fundamental Mnemonic Abilities. *Psychological Science, 29*(1), 45–55. https://doi.org/10.1177/0956797617728592

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review, 122*(4), 621–647. https://doi.org/10.1037/a0039413

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A Context Maintenance and Retrieval Model of Organizational Processes in Free Recall. *Psychological Review, 116*(1), 129–156. https://doi.org/10.1037/a0014420

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3, Pt.1), 353–363. https://doi.org/10.1037/h0025953

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences, 20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Richards, B. A., & Frankland, P. W. (2017). The Persistence and Transience of Memory. *Neuron, 94*(6), 1071–1084. https://doi.org/10.1016/j.neuron.2017.04.037

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences, 17*, 114–123. https://doi.org/10.1016/j.cobeha.2017.07.016

Rolls, E. T. (2013). The mechanisms for pattern completion and pattern separation in the hippocampus. *Frontiers in Systems Neuroscience, 7*(OCT), 1–21. https://doi.org/10.3389/fnsys.2013.00074

Rolls, E. T. (2016). Pattern separation, completion, and categorisation in the hippocampus and neocortex. *Neurobiology of Learning and Memory, 129*, 4–28. https://doi.org/10.1016/j.nlm.2015.07.008

Rudy, J. W., & Sutherland, R. J. (1995). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *Hippocampus, 5*(5), 375–389.

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences, 20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26*(1), 3–8. https://doi.org/10.1002/hipo.22523

Schlichting, M. L., & Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences, 1*, 1–8. https://doi.org/10.1016/j.cobeha.2014.07.005

Shadlen, M. N. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron, 90*(5), 927–939. https://doi.org/10.1016/j.neuron.2016.04.036

Shohamy, D., & Daw, N. D. (2015). Integrating memories to guide decisions. *Current Opinion in Behavioral Sciences, 5*, 85–90. https://doi.org/10.1016/j.cobeha.2015.08.010

Shohamy, D., & Wagner, A. D. (2008). Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events. *Neuron, 60*(2), 378–389. https://doi.org/10.1016/j.neuron.2008.09.023

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8*, 203–220. https://doi.org/10.3758/BF03196157

Squire, L. R., Stark, C. E. L., & Clark, R. E. (2004). the Medial Temporal Lobe. *Annual Review of Neuroscience, 27*(1), 279–306. https://doi.org/10.1146/annurev.neuro.27.070203.144130

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62*(6), 1041–1062. https://doi.org/10.1080/17470210902747112

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *ELife, 12*, 1–21. https://doi.org/10.7554/eLife.77185

Sutherland, R. J., & Rudy, J. W. (1989). Configural Association theory: The contribution of the hippocampus to learning, memory, and amnesia. *Psychobiology, 17*(2), 129–144.

Teyler, T. J., & Rudy, J. W. (2007). The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus, 17*(12), 1158–1169. https://doi.org/10.1002/hipo.20350

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic Hypothesis Generation and Human Judgment. *Psychological Review, 115*(1), 155–185. https://doi.org/10.1037/0033-295X.115.1.155

Todd, P. M., & Hills, T. T. (2020). Foraging in Mind. *Current Directions in Psychological Science, 29*(3), 309–315. https://doi.org/10.1177/0963721420915861

Tonegawa, S., Liu, X., Ramirez, S., & Redondo, R. (2015). Memory Engram Cells Have Come of Age. *Neuron, 87*(5), 918–931. https://doi.org/10.1016/j.neuron.2015.08.002

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology, 53*(1), 1–25.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 381–391. https://doi.org/10.1016/S0022-5371(66)80048-8

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111*(3), 757–769. https://doi.org/10.1037/0033-295X.111.3.757

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38*(4), 599–637. https://doi.org/10.1111/cogs.12101

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77*(1), 1–15. https://doi.org/10.1037/h0028569

Wimmer, G. E., & Büchel, C. (2016). Reactivation of reward-related patterns from single past episodes supports memory-based decision making. *Journal of Neuroscience, 36*(10), 2868–2880. https://doi.org/10.1523/JNEUROSCI.3433-15.2016

Wimmer, G. E., & Shohamy, D. (2012). *Preference by Association : Science, 338*(October), 270–273. https://doi.org/10.1126/science.1223252

Yassa, M. A., & Reagh, Z. M. (2013). Competitive Trace Theory: A Role for the Hippocampus in Contextual Interference during Retrieval. *Frontiers in Behavioral Neuroscience, 7*. https://doi.org/10.3389/fnbeh.2013.00107

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences, 34*(10), 515–525. https://doi.org/10.1016/j.tins.2011.06.006

Zeithamova, D., & Bowman, C. R. (2020). Generalization and the hippocampus: More than one story? *Neurobiology of Learning and Memory, 175*, Article 107317. https://doi.org/10.1016/j.nlm.2020.107317

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168–179. https://doi.org/10.1016/j.neuron.2012.05.010

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience, 6*(2012–03), 1–14. https://doi.org/10.3389/fnhum.2012.00070

Zhang, Z., Wang, S., Good, M., & Hristova, S. (2021). Retrieval-constrained valuation : Toward prediction of open-ended decisions. DOI: 10.1073/pnas.2022685118/-/DCSupplemental.Published.

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments. *Psychological Review, 127*(5), 719–748. https://doi.org/10.1037/rev0000190