Episodic retrieval for model-based evaluation in sequential decision tasks

Corey Y Zhou[1], Deborah Talmi[2], Nathaniel D Daw[3,*], and Marcelo G Mattar[1,4,*]

[1]Department of Cognitive Science, University of California San Diego, San Diego, CA, USA
[2]Department of Psychology, University of Cambridge, Cambridge, UK
[3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
[4]Department of Psychology, New York University, New York, NY, USA
[*]equal contribution
[@]corresponding author: marcelo.mattar@nyu.edu

Abstract

It has long been hypothesized that episodic memory supports adaptive decision making by enabling mental simulation of future events. Yet, memory research is often carried out in settings that are far removed from ecological contexts of decision making, and models of adaptive choice, conversely, only invoke episodic memory in highly stylized terms, if it all. To address these gaps, we propose a novel process-level model of choice that grounds model-based evaluation in empirically informed dynamics of episodic recall. In this model, the probability of retrieving each available memory sample is given by the Successor Representation (SR), a biologically plausible world model in reinforcement learning. The evolution of these probabilities based on past retrievals, in turn, is dictated by the Temporal Context Model (TCM), a prominent model of episodic retrieval. Through a series of simulations, we demonstrate that the patterns of episodic retrieval suggested by this model enables flexible computation of decision variables. On this basis, we argue that a number of previously described features of episodic memory serve an adaptive purpose in sequential decision making. For instance, we show that the classic retrieval bias known as contiguity effect, when viewed from a decision making perspective, leads to model-based rollouts for forward simulation. We also show that features of episodic memory such as emotional modulation enable generalization and efficient decisions given limited experience. By bridging theoretical models across these two domains, we make a set of theoretical predictions linking episodic memory properties to adaptive choice in sequential tasks that may guide future empirical endeavors.

Keywords: episodic memory; decision-making; successor representation; temporal context model; reinforcement learning

Introduction

What is memory for? Although laboratory studies often focus on memory performance in isolation, as if recall accuracy is the participants' only goal, an important real-world use of past experience is to guide adaptive choices. This observation has driven increasing interest in the interplay between memory and decision making, and delivered promising insights. Understanding this interplay promises both to unpack the mechanisms by which experience guides choice, and to illuminate the potential adaptive function of various seemingly arbitrary aspects of memory.

The relationship between memory and decisions is perhaps most apparent for procedural memory, where a putative neurocomputational mechanism involving dopamine, prediction errors, and stimulus-response habits has long been the shared, orthodox model in both areas (Dolan & Dayan, 2013). Building on this relationship, there has been increasing interest in how different memory systems might relate to different decision systems, including a potential correspondence between declarative memory and the cognitive maps or models thought to guide goal-directed systems for deliberative evaluation of candidate actions (Doll, Shohamy, & Daw, 2015; Eichenbaum, 2001). In particular, sequential decision tasks like spatial navigation or chess offer much evidence that the brain engages in constructive, deliberate evaluation, akin to mental simulation informed by map- or model-like information about the task (Pfeiffer & Foster, 2013; van Opheusden et al., 2021). However, we still understand relatively little about the mechanisms by which deliberative sequential decisions are achieved, or how they might draw on specific memory processes long-established in memory laboratories.

In this paper, we propose a new mechanistic theory of decision making that grounds model-based evaluation in the recall of episodic memories, or memories for individual autobiographical events (Tulving, 1972). Many decisions could benefit from recall of one-off autobiographical events. For example, to navigate through a large, unfamiliar venue, we may recall having examined the sculpture in the corridor on the right earlier that night, and use that memory to orient ourselves. Episodic memory has also been suggested to guide decisions by scaffolding the construction of hypothetical future scenarios (Schacter, Benoit, Brigard, & Szpunar, 2015). Indeed, patients with episodic memory deficits are less effective at certain decision making tasks (Gutbrod et al., 2006; Gupta et al., 2009; Bakkour et al., 2019). Relatedly, researchers in decision neuroscience have become interested in a class of *decision-by-sampling* algorithms. These algorithms bear a loose analogy to episodic memory, in that decisions are achieved by considering a small number of individual past experiences with similar actions and their outcomes (Plonsky, Teodorescu, & Erev, 2015; Bornstein, Khaw, Shohamy, & Daw, 2017; Lieder, Griffiths, & Hsu, 2018). Yet, despite the suggestive links, these previous theories are not especially informed by research on memory, and have also only been applied to a restricted class of single-step decision tasks.

Our approach instead begins with a standard model of episodic encoding and recall – the temporal context model (TCM; Howard & Kahana, 2002; Sederberg, Howard, & Kahana, 2008; Polyn, Norman, & Kahana, 2009; Talmi, Lohnas, & Daw, 2019). TCM is a descriptive (rather than normative) model originally conceived to capture patterns of episodic retrieval in tasks like word-list learning. Here, we present a series of simulations in which TCM is applied in the setting of sequential decision making. We show that, when the problem of action-outcome prediction is framed as the problem of recalling relevant past experiences (which we formalize with off-the-shelf TCM recall), the resulting algorithm provides a novel, parameterized family of decision-by-sampling estimators that are provably appropriate for sequential decision tasks. Our study builds on previous research showing that the encoding stage in TCM closely relates to model learning, enabling gradual construction of a type of world model

known as successor representation (SR) (Gershman, Moore, Todd, Norman, & Sederberg, 2012). We extend the prior work by studying the predictions of TCM with respect to memory retrieval, which we show to correspond to queries of the learned model that can be used for planning or evaluation at decision time. The result is a theoretical proposal that we call TCM-SR.

Despite its root in memory literature, TCM-SR has a quantitative mapping to reinforcement learning (RL) models in decision neuroscience that expands the connection between the two fields. We show that two special cases of our model correspond to two influential mechanisms for model-based choice: a constructive "rollout"-based simulation of future trajectories, and the use of temporal abstraction (SR) to compress such iterative serial reasoning. We then show that the full model extends and interpolates between these two extremes using intermediate parameterizations, providing a family of Monte Carlo estimators based on a generalized notion of rollouts. We also show that several other known properties of episodic memory can be viewed as rational from a decision-making standpoint. For instance, people sometimes recall events in the opposite temporal sequence to that experienced during encoding, and recall is often biased toward emotionally arousing events. Viewed in the context of our theory, these and other features of episodic memory have unanticipated advantages for choice. More broadly, the direct mapping we hypothesize between research in episodic memory and decision making sheds light on both areas, and suggests many new research directions and future experiments.

The remainder of this paper is structured as follows: We begin with a description of the normative problem of interest and the properties of the episodic memory system. We then propose a simplified TCM-inspired model of episodic memory for sample-based action evaluation; this model illustrates the key ideas behind our theory and serves as the basis for the more realistic variants that are presented subsequently. Then, in each subsequent section, we show that progressive addition of known episodic memory properties (formalized in different variants of TCM) confers unexpected decision making advantages. As we will show, our model makes a series of empirical predictions regarding the content of retrieval during decision making, how speed and accuracy are traded off during episodic-based evaluation, and how a number of known memory retrieval biases give rise to novel choice biases that are amenable to empirical testing.

Results

Decisions via model-based evaluation

We explore how episodic memory retrieval can be used to guide decisions using a stylized decision making task in which an action is followed by a sequence of states, each associated with a (potentially nonzero) reward (Fig. 1a). This task resembles a game of Plinko where a player drops the ball in one of the holes in the top row of the board. The initial ball placement (action) determines the first state of the sequence. Each subsequent state (and reward) follows from the previous through a stochastic transition, analogous to a ball falling through the Plinko board. The agent's goal is to choose the starting state that maximizes the cumulative reward. We selected this stylized task to depict graphically the process of sequential retrieval. Despite its simplicity, this problem captures a number of key aspects of more general sequential decision tasks — in particular, rewards accumulate sequentially over a series of steps, and cannot be predicted with certainty from each action. The problem of optimal decision

here can be reduced to a problem of *prediction*: estimating, for each candidate action, the resulting (sequential, stochastic) rewards. This is the function we ascribe to episodic retrieval.

We formalize this intuition using the framework of Reinforcement Learning (RL; Sutton & Barto, 2018). On a particular trial, the agent receives total discounted reward $G = \sum_{t=1}^{\infty} \gamma^t R_t$, where R_t is the reward received at timestep t and γ is a discount factor specifying the degree to which earlier rewards are favored over later rewards. The agent's goal is to select the action which, by affecting the sequence of future states, maximizes the expected G. One strategy is to estimate the expectation for each candidate action a, i.e. $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t R_t \mid A = a]$. The quantity $q(a)$ is known as the *action value*. Using this strategy, the agent can evaluate each candidate action and select the action with maximum value. Note that no further action is involved after the initial one; in RL terms, this corresponds to policy evaluation in a Markov process, a classic sub-problem for solving more elaborate choice tasks (e.g., Markov decision processes, in which actions can occur at every step).

RL offers various approaches to estimate action values, falling broadly in two categories: agents learn aggregated action values q from experience, or instead draw on a "world model" of the environmental dynamics to simulate action outcomes. The former approach is most commonly associated with the classic temporal difference (TD) algorithm (Sutton, 1988) and procedural memory, and not the focus of this paper.

Here we focus on the second class of strategies, often called *planning* or *model-based RL*. Suppose at any point of the Plinko game, the agent is capable of predicting the probability of the ball's board position at the next time step – i.e., the agent understands the step-by-step transition structure of the game, a form of world model (Fig. 1b, boards labeled as $\mathbf{T}^1$, $\mathbf{T}^2$, $\mathbf{T}^3$). By recursively predicting the position of the ball one step into the future, the agent can simulate one of many possible trajectories following a given action, along with the corresponding rewards. A complete trajectory simulated in this way is called a *rollout*, and its associated total reward provides a noisy estimate of the value of the given action. Taking the total reward across for each considered action, averaged across multiple rollouts, the agent can choose the action with maximal estimated value. Action evaluation by stochastic, iterative simulation is at the heart of numerous model-based approaches to RL, such as Monte Carlo Tree Search (Coulom, 2006). Its power — for instance, in competitive play of challenging games like Go (Silver et al., 2016) — arises from its ability to compositionally (albeit laboriously) analyze novel situations, such as never-experienced board positions (Daw & Dayan, 2014; Mattar & Lengyel, 2022).

An alternative and often more efficient RL approach is to first learn, for each action, the expected number of visits to each state in the future (formally, $\mathbf{M} = \mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \gamma^2\mathbf{T}^3 + \cdots$, where each element M_{ij} of matrix $\mathbf{M}$ represents the discounted number of visits to state j from state i). $\mathbf{M}$ is known as Successor Representation (SR; Dayan, 1993), and is thought to account for various features of human and animal behavior and neural responses (Momennejad et al., 2017; Stachenfeld, Botvinick, & Gershman, 2017; E. M. Russek, Momennejad, Botvinick, Gershman, & Daw, 2017, 2021; Piray & Daw, 2021). Like $\mathbf{T}$, the SR matrix $\mathbf{M}$ summarizes the transition structure of the world, but aggregated over multiple steps; thus it can also be understood as another form of world model (Fig. 1b, board labeled $\mathbf{M}$). If the SR is known, action values can be estimated straightforwardly by multiplying the expected number of visits to each state by the rewards present in those states (i.e., $q(a) = \mathbf{x}_a^\top \mathbf{M}\mathbf{r}$, where $\mathbf{x}_a$ is a one-hot column vector denoting the top-row state resulting from action a, and $\mathbf{r}$ is a column vector whose k^{th} element r_k indicates the reward present in state k). Thus, while still relying on the basic "world model" approach, the SR simplifies evaluation and avoids the iterative construction of trajectories by using a stored model of aggregated transition dynamics over multiple time steps. The cost of this simplification (called *temporal abstraction*) is that it limits the flexibility of the model to

work out value in novel or changed situations, because information about future events is "baked in" to $\mathbf{M}$ (E. M. Russek et al., 2017; Piray & Daw, 2021). Overall, then, these two model-based strategies for prospective evaluation have different costs and benefits.

In this paper, we show that the properties of episodic memory imply an additional approach for estimating action values, which both generalizes and interpolates between the rollout-based and SR-based approaches, balancing two different strategies for long-term prospection. Our proposal builds on the observation that episodic memory encoding has the effect of learning an SR-like model (Gershman et al., 2012). We leverage this observation to show that the *sequential retrieval* of remembered events in the same memory model implements a rollout-like (iterative) state simulation process that differs from standard (non-iterative) uses of the SR described previously. Accordingly, we next describe the processes of memory encoding and retrieval that support value estimation.

Episodic retrieval via the Temporal Context Model

Our starting point is a standard model of memory encoding and retrieval, the Temporal Context Model (TCM; Howard & Kahana, 2002), which we simplify in the first instance and progressively augment to expose the contribution of different model components. TCM aims to explain experiments where memory is the dependent variable: which stimuli tend to be recalled and in which order, as a function of factors such as their serial position during encoding (Fig. 1c). To explain these results, TCM centrally posits that such episodic retrieval is affected by a drifting *temporal context*, a continuously evolving representation composed of a recency-weighted running average of previously observed and retrieved stimuli:

$$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{x}_t. \tag{1}$$

During encoding, associations are formed between the representation of the observed stimulus $\mathbf{x}_t$ and the temporal context $\mathbf{c}_t$ present at that moment (Fig. 1d-f). During retrieval, items are sampled from memory in proportion to how well the current temporal context matches the context associated with each item during encoding (Fig. 1g-i). Retrieval is thus determined by the current context and the agent's memory (Fig. 1h,i), formalized as a set of associations developed during encoding (Fig. 1f) that represent the contexts associated with each previously-seen item. Finally, retrieval also updates the temporal context, biasing the subsequent retrieval of new items from memory (Fig. 1g-i). The updating of the temporal context when an item is recalled allows TCM to explain ubiquitous patterns of sequential retrieval in list learning tasks (see Methods for a formal description of TCM).

TCM recapitulates two recall biases often observed in list learning paradigms: the recency effect and the contiguity effect (Fig. 1c). The recency effect is the observed heightened probability of recalling the most recently-studied information; as the temporal context drifts continuously in TCM, the context at recall better matches contexts associated with the stimuli studied last. The contiguity effect refers to a tendency for subsequent recalls to contain stimuli studied in close temporal proximity; because temporal contexts tend to be similar for temporally close-by stimuli, the retrieval of one promotes retrieval of others studied close in time. Note that TCM is a descriptive model, as it aims to match rather than rationalize or justify these empirically observed patterns.

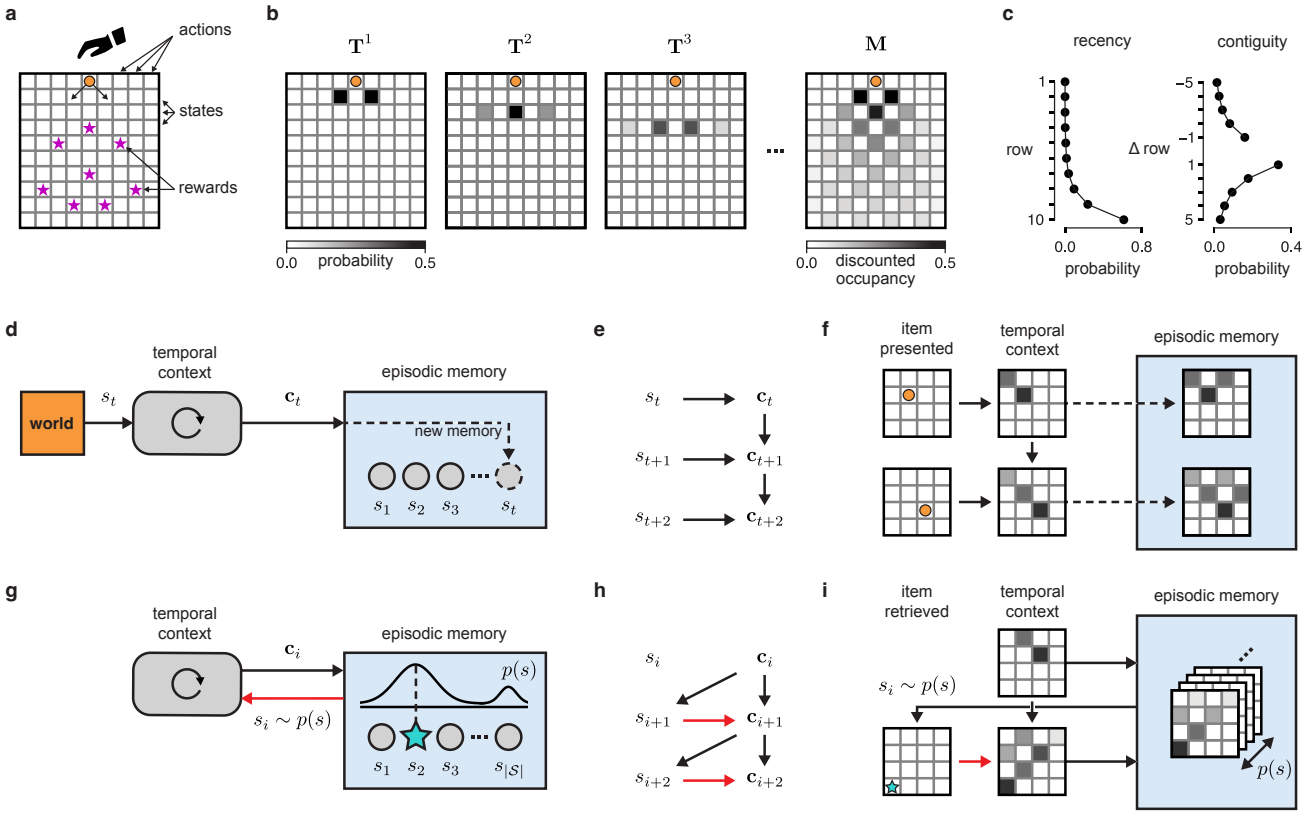

Fig. 1. Overview of the TCM-SR model. (a) Our Plinko game has 10×9 states, each represented by a small square. The agent may take any of 9 possible actions, corresponding to the 9 locations on the top row where the Plinko ball (orange circle) may be dropped. The dropped ball follows a stochastic trajectory down the board, collecting scattered rewards (purple stars) along the way (see Methods for task details). The goal of the agent is to select the action leading to a trajectory containing as many rewards as possible. **(b)** The first three Plinko boards labeled $\mathbf{T}^1$, $\mathbf{T}^2$, and $\mathbf{T}^3$ represent the probability distribution of the ball location 1, 2, and 3 time steps after the moment depicted in (a) respectively. The Plinko board labeled $\mathbf{M}$ represents the fully-learned Successor Representation (SR), given by $\mathbf{M} = \gamma^0\mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \gamma^2\mathbf{T}^3 + \cdots$. SR values correspond to the expected number of (discounted) visitations to each state on the board, starting from the action depicted in (a). **(c)** After each full trajectory is experienced and stored in memory, the recency effect (left) predicts that stimuli from the bottom rows, which have been experienced more recently, are more likely to be retrieved. The contiguity effect (right) predicts that, following each stimulus retrieved on a given row, stimuli from adjacent rows are more likely to be subsequently retrieved. **(d-f)** Encoding phase of TCM-SR. **(d)** Presentation of stimulus s_t at time t by the external world updates the temporal context $\mathbf{c}_t$. Memory encoding amounts to storing each temporal context present when a stimulus is seen. The first time each stimulus is presented, a new memory is stored (circle with dashed outline). Each subsequent time the same stimulus is presented, the associated memory is modified (not shown). **(e)** The temporal context $\mathbf{c}_i$ defines a distribution $p(s)$ over memories. It depends on the previous temporal context $\mathbf{c}_t$ and the current state s_{t+1}, corresponding to a recency-weighted representation of the stimuli (depicted in f). **(f)** Schematic of encoding two consecutive stimuli in the Plinko task. Stored memory of each stimulus (right box) includes a composite representation of temporal contexts present during each of the encoding situations. **(g-i)** Retrieval phase of TCM-SR. **(g)** The agent freely samples one or more stimuli during retrieval. The retrieved stimulus s_i is a sample from the recall distribution $p(s)$. Higher retrieval probability is assigned to stimuli whose stored context is more similar to the current context. The context associated with the sample influences the temporal context to affect subsequent retrievals. **(h)** The temporal context $\mathbf{c}_{i+1}$ depends on the previous temporal context $\mathbf{c}_i$ and the retrieved stimulus s_{i+1}, which itself depends on the previous context $\mathbf{c}_i$. The red arrow illustrates how the temporal context is affected by each retrieved stimulus. **(i)** Schematic of retrieving a stimulus in the Plinko task. The temporal context is updated by a retrieved context. See the online article for the color version of this figure.

TCM predictions for decision tasks

In the present article, we study the predictions of TCM for an agent performing a sequential decision task. While we study these predictions for a general task, we illustrate them in the context of a stylized problem, the Plinko game. Here, experienced states (s_t, ball locations in Plinko) take the place of list items (e.g., words in TCM). In the encoding phase (corresponding to learning a task) a sequence of states is experienced (a trajectory followed by the ball in Plinko, viewed as a word list in TCM) and stored in memory. In the retrieval phase states previously stored in memory can be retrieved, corresponding to locations in the Plinko board that have been previously visited by the ball. We propose that such recall of states can be used at some later decision time to evaluate actions, akin to an agent querying an episodic memory for choice-relevant information.

To understand this process, we first note that the problem of learning associations between a stimulus and its temporal context during encoding, as formalized in TCM, is equivalent to the problem of learning the SR (see Gershman et al., 2012 or the Methods section). Intuitively, learning what context precedes each stimulus is equivalent to learning that the stimulus is a likely successor of the context components (notice how the episodic memory representations in Fig. 1f share characteristics with $\mathbf{M}$ in Fig. 1b). Leveraging this observation about memory *encoding*, the rest of this paper shows that the properties of episodic *retrieval*, as envisioned in TCM, are sufficient to compute estimates of action values $q(a)$ by a rollout-like state sampling process. We further show that this hypothetical process reflects knowledge of the cognitive map or model of the task, and specifically the SR (E. M. Russek et al., 2017; Momennejad et al., 2017; Gershman, 2018; Piray & Daw, 2021)), suggesting that episodic retrieval is a candidate mechanism for *model-based* or *goal-directed* decisions in the brain.

In the following sections, we examine in increasing levels of detail how value estimation can be achieved via episodic retrieval. We begin by stripping down episodic memory of many of its defining properties by removing a number of algorithmic details from the original TCM formulation, each of which corresponds to one such property. While this approach may seem overly abstracted at first, it leads to the cleanest baseline instantiation of recursive retrieval as a sampling algorithm that estimates action values in sequential decision problems. We then gradually reintroduce properties of episodic memory, which allows us to systematically analyze how each of them confers a different advantage for action evaluation and ultimately choice. These advantages include temporal horizon extension, one- and few-shot learning, bias-variance trade-off, and sample efficiency improvement.

Independent samples from memory yield unbiased value estimates

To study how the episodic retrieval can be used for action evaluation, we start by making two simplifying assumptions which will be relaxed in subsequent sections. The first assumption is that each stimulus (state) is experienced many times during encoding: this is analogous to studying the same word in multiple lists in a free recall task, or to the Plinko ball visiting a particular state multiple times across many trajectories. We make this assumption at this moment for didactic purposes, acknowledging that episodic memory is more commonly associated with low-sample regimes in which learned stimuli are experienced once or at most a few times. The repeated exposure associates each stimulus with a single context representation, obtained by combining the contexts across every presentation of the same stimulus (see Fig. 1f for an example of composite representation in episodic memory). As

shown by Gershman et al. (2012), this composite context is equivalent to the stimulus' steady-state (i.e., fully converged) SR.

The second simplifying assumption is that the retrieval of a stimulus does not affect the temporal context. That is, during retrieval we set $\beta = 0$ and $\rho = 1$ in Eq. (1), leading to $\mathbf{c}_i = \mathbf{c}_{i-1}$ (this is equivalent to removing the red arrows in Fig. 1g-i). Note that this simplification eliminates the model's ability to explain the contiguity effect. Additionally, we do not impose the constraint often present in free-recall tasks that the same item cannot be retrieved multiple times. In this setting, retrieved stimuli can be viewed as "samples" that are independent and identically distributed (i.i.d.).

With the two assumptions above in place, the predictions of this stripped-down TCM formulation are that the set of retrieved stimuli are i.i.d. samples (second assumption) from the steady-state normalized SR (first assumption) of the queried action (Fig. 2a). This observation suggests a potential use for these samples in decision making. Specifically, an action can be evaluated by averaging the rewards associated with the episodically retrieved samples from the SR:

$$\hat{q}(a) \propto \frac{1}{N} \sum_{i=1}^{N} \mathbf{r}^{\mathsf{T}} \mathbf{x}(S_i), \tag{2}$$

where $S_1, S_2, \ldots, S_N \sim p(s)$ are samples from the normalized SR, i.e., $p(s) = \frac{\mathbf{x}_a^{\mathsf{T}} \mathbf{M}}{|\mathbf{x}_a^{\mathsf{T}} \mathbf{M}|} \mathbf{x}(s)$, $\mathbf{x}(S_i)$ is the one-hot feature vector for state S_i (for which we some times use the shorthand $\mathbf{x}_i$), and $r_i = \mathbf{r}^{\mathsf{T}} \mathbf{x}(S_i)$ is the reward present in state S_i. Thus, $\hat{q}(a)$ is obtained by averaging samples of r_i.

Intuitively, the agent first resets the temporal context to the action to be evaluated (note that this eliminates any residual effect of recent history). The agent then retrieves a sequence of successor states and their respective rewards (Fig. 2a). Eq. (2) shows that the average reward across all sampled states is a proxy for the action value, as we originally defined it. Repeating such retrieval-based evaluation for each candidate action can thus inform the agent to select the highest-valued action. Note this procedure is not derived from normative considerations (i.e., what memories an agent ought to retrieve); rather, it is a direct prediction of TCM: given the assumptions in place, TCM predicts i.i.d. sampling from the SR, retrieving states whose average reward is the normative action value. Our contribution here is to highlight and express this prediction formally and to show that these samples can be used straightforwardly to compute action values.

The action values estimated by this process depend directly on the associations learned during encoding (i.e., the SR). In particular, the temporal context drift rate determines the similarity between the contexts associated with two consecutive stimuli. During retrieval, this rate modulates the sharpness by which retrieval is biased toward states occuring soon after the starting context. In RL terms, this amounts to the temporal horizon of the SR, parameterized by the discount factor γ. This ultimately affects the overall value estimated; depending on the discount factor, the computed value ranges between (i) rewards sampled exclusively from imminent states ($\gamma = 0$, Fig. 2a,b), and (ii) rewards sampled from all future states, with a preference for earlier states ($\gamma > 0$, Fig. 2d,e). Notably, the former case ($\gamma = 0$) implements the evaluation required for bandit problems, in which action values depend only on instantaneous rewards. Indeed, a special case of the current model corresponds to a class of *decision-by-sampling* models that have been previously described and empirically tested in single-step problems like bandits (e.g. Plonsky et al., 2015; Bornstein et al., 2017; Lieder et al., 2018). The latter case ($\gamma > 0$) extends the i.i.d. decision-by-sampling approach to sequential problems. Unlike rollout-based algorithms like MCTS, which sample states serially conditional on their predecessors to produce

trajectories, this approach estimates action values by i.i.d. Monte Carlo sampling. Such sampling is possible because the SR effectively "flattens" the tree-like set of future situations in a sequential task to a set of individual future states weighted by their prevalence in the tree. Consequently, it transforms sequential decision tasks into bandit problems studied previously, extending the findings from sampling models to the sequential case.

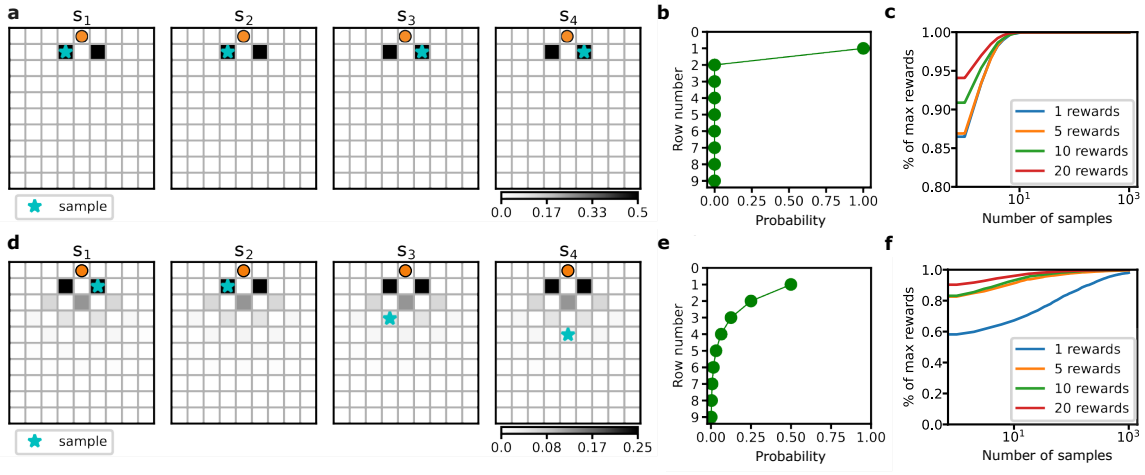

Fig. 2. Independent samples from memory yield unbiased value estimates. (a-c) Sampling from a distribution with a short temporal horizon. Parameters: $\rho = 1$, $\beta = 0$, $\gamma = 0$. **(a)** An example of querying an action (orange circle) through memory recall (cyan stars). s_i shows the i^{th} stimulus sampled, where the same state can be sampled multiple times. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board. **(c)** We simulate an agent who evaluates two actions (at top-center state and the state immediately adjacent to the right) using the procedure from Eq. (2), and then selects the action with the larger estimated value. The image shows the fraction of maximum rewards (y-axis) expected as more samples are drawn (x-axis, shown in log-scale) as a function of different numbers of rewards placed on the Plinko board. **(d-f)** As in **a-c**, but using parameters: $\rho = 1$, $\beta = 0$, $\gamma = 0.5$. See the online article for the color version of this figure.

As more sampled rewards are averaged, the action value estimate approaches the truth, enabling better decisions. However, more samples typically require more time and resources. This leads to the question: how many samples should one draw for a decision? The answer depends on one's goal. Accurate action value estimation in our task entails dozens or hundreds of samples, as each sample provides reward information about only one of various successor states. However, many fewer samples are usually needed for action selection, as illustrated in the following two scenarios. First, if the value of one action dominates the others (i.e. one action leads to much larger rewards than the others), it can be identified with many fewer samples than needed to estimate all action values accurately. Second, if no action value dominates the others, identifying the optimal action requires a large number of samples, but the extra computation will not lead to a substantially larger payoff. Either way, a large fraction of the available payoff can be achieved with relatively few samples (Fig. 2c,f): in Plinko, over 80% of maximum available reward can be obtained with fewer than 10 samples, unless the available rewards are extremely sparse (e.g., a single reward placed in the Plinko board). This prediction aligns with previous work demonstrating that surprisingly few samples are needed for effective decisions in bandit problems (Vul, Goodman, Griffiths, & Tenenbaum, 2014); here, by extending the decision-by-sampling approach we show that a similar observation applies to sequential problems as well.

In sum, if retrieval does not update the temporal context, action values can be estimated straightforwardly by sampling stimuli i.i.d. from episodic memory and averaging the corresponding rewards. That is, in this parameter regime, TCM-SR embodies the SR's strategy for forecasting future events by temporal abstraction: it records long-run sequential contingencies experienced at encoding time, so as to easily recapituate them by retrieval at choice time. However, unlike previous invocations of SR in decision neuroscience and RL, this retrieval is accomplished by sampling individual future states rather than by exhaustive summation. This brings temporally abstract prospection into contact with episodic retrieval and decision-by-sampling models. The next section shows that episodic retrieval can also lead to rollout-based prospective simulation.

The contiguity effect enables value estimation via rollouts

The previous section considered a simplified setting in which the retrieval of a stimulus does not affect subsequent retrievals, giving rise to i.i.d. samples that the agent could average to obtain action values estimates. However, a prominent feature of episodic memory is that consecutive retrievals are *not* independent. Indeed, the simplifying assumptions from the previous section eliminate the model's ability to explain the contiguity effect, ubiquitous in list learning experiments. Thus, we now consider a different parameter regime of TCM, in which stimulus retrieval *does* affect subsequent retrievals. We focus initially on the extreme case where retrieval depends only on the immediately preceding retrieved stimulus (i.e., we set $\beta = 1$ and $\rho = 0$ in Eq. (1) to yield $\mathbf{c}_i = \mathbf{x}_i$), while assuming that this update is driven by a static, task-independent representation of each stimulus — another simplifying assumption that we also relax in the last section. TCM operationalizes this setting by fully updating the temporal context with the last retrieval, retaining no information retrieved before that. Thus, in contrast to the i.i.d. setting, this setting produces correlated samples (forming a Markov chain), which can also be used to estimate action values.

As previously, the temporal context drift rate has a direct impact on sharpness of the distribution over retrieved states. In particular, a quickly evolving temporal context during encoding leads to the learning of an SR with a low discount factor γ. In the extreme of $\gamma = 0$, the first retrieved memory is an immediate successor of the considered action (because $\mathbf{M} = \mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \cdots = \mathbf{T}^1$ when $\gamma = 0$, Fig. 1b). Upon retrieving the first memory and updating the temporal context, the second retrieved memory is an immediate successor of the first sample (Fig. 3a). Repeating this sampling process recursively leads to a rollout (in Plinko, this process amounts to a simulation of a trajectory through which the ball might plausibly fall; Fig. 3a,b).

How can these samples used to estimate action values? As described in the RL literature (Tesauro & Galperin, 1996; Coulom, 2006), the sampled rewards in a traditional rollout can be added to produce an estimate of the action value:

$$\hat{q}_{\tilde{\gamma}=1}(a) \propto \sum_{i=1}^{N} \mathbf{r}^\mathsf{T}\mathbf{x}(S_i), \tag{3}$$

where $S_1, S_2, \ldots, S_N$ are samples from the normalized SR with $\gamma = 0$, each represented by a one-hot feature vector $\mathbf{x}_i = \mathbf{x}(S_i)$, with $p(S_1 = s) = \frac{\mathbf{x}_a^\mathsf{T}\mathbf{M}}{|\mathbf{x}_a^\mathsf{T}\mathbf{M}|}\mathbf{x}(s)$ representing the SR of the queried action, $p(S_2 = s) = \frac{\mathbf{x}_1^\mathsf{T}\mathbf{M}}{|\mathbf{x}_1^\mathsf{T}\mathbf{M}|}\mathbf{x}(s)$ representing the SR of the first sample, and so on. Note that each stimulus of the trajectory $S_1, S_2, \ldots, S_N$ is drawn from a different distribution.

Intuitively, for each action being evaluated, the agent retrieves a plausible sequence of states and the rewards associated with them. The total reward across all sampled states is an estimator for the action value. This is equivalent to an agent recalling a previous study list, and evaluating its worth based on the number of rewarded items it recalled. Again, this is a descriptive observation about TCM rather than a normative prescription about memory: a specific parameter regime of TCM implies that stimuli will be retrieved in sequences that correspond to a rollout in RL. Our contribution is to make this observation explicit and note that such rollouts can be used to estimate action values.

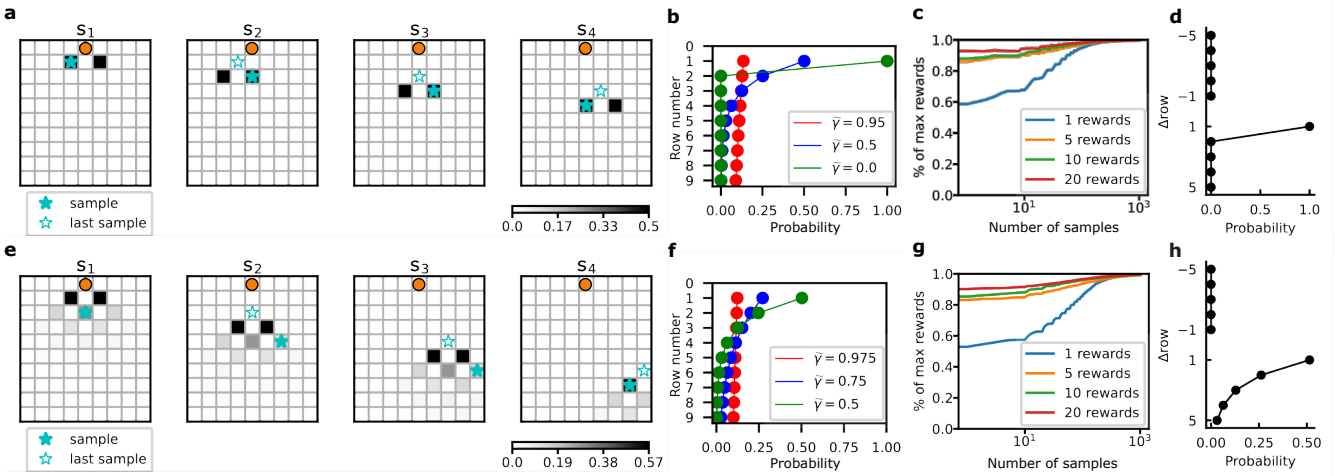

Fig. 3. Recall-dependent context updates lead to rollouts. **(a-d)** Sampling from a distribution with a short temporal horizon. Parameters: $\rho = 0$, $\beta = 1$, $\gamma = 0$. **(a)** An example sequence of memory retrieved when initiating the temporal context as the top-center state (orange circle) through memory recall (cyan stars). $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board. We illustrate these distributions for three values of p_{stop} (0.05, 0.5, and 1), each leading to an effective temporal discount factor $\tilde{\gamma} = 1 - p_{\text{stop}}$. **(c)** We simulate an agent that evaluates two actions (at top-center state and the state immediately adjacent to the right) using the procedure from Eq. (4), and then selects the action with the larger estimated value. The image shows the fraction of maximum rewards (y-axis) expected as more samples are drawn (x-axis, shown in log-scale), setting $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d)** Probability that a sample is drawn from each row of the Plinko board, as a function of the distance to the previously sampled row. **(e-h)** As in **a-d**, but using parameters: $\rho = 0$, $\beta = 1$, $\gamma = 0.5$. See the online article for the color version of this figure.

Note that each rollout incorporates all future rewards with equal weight. This leads to an action value estimate for a unit discount factor (which we denote $\hat{q}_{\tilde{\gamma}=1}(a)$), even though the sampling distributions specified by the normalized SR are encoded with $\gamma = 0$. This happens because, by retrieving n consecutive memories and summing the rewards according to Eq. (3), one is concatenating n one-step predictions (i.e., $\gamma = 0$), which is equivalent to performing one n-step prediction (i.e., $\tilde{\gamma} = 1$). However, weighing all future rewards equally is not always desirable or even possible, as it would require each rollout to continue forever. We circumvent this issue by positing a fixed probability of interrupting the retrieval process at any moment, denoted p_{stop}. The larger the interruption probability, the less likely is the rollout to continue far into the future. This probability, in turn, allows the agent to control

the effective discount factor of the constructed values during retrieval:

$$\hat{q}_{\tilde{\gamma}}(a) \propto \sum_{i=1}^{N} \mathbf{r}^{\mathsf{T}} \mathbf{x}(S_i), \tag{4}$$

where $S_1, S_2, \ldots, S_N$ are samples from the normalized SR with $\gamma = 0$. The effective discount factor is given by $\tilde{\gamma} = 1 - p_{\text{stop}}$, where p_{stop} is the interruption probability (see Methods for details). Using this sampling scheme, we can measure the empirical distribution that each sample is drawn from each row for different values of p_{stop} (Fig. 3b). This confirms the relationship between p_{stop} and the effective discount factors during retrieval, $\tilde{\gamma}$.

The reliability of a value estimate is again proportional to the number of samples and rollouts performed. As before, over 80% of maximum available reward can be obtained with fewer than 10 samples (i.e., one full rollout), unless the available rewards are extremely sparse (e.g., a single reward placed in the Plinko board; Fig. 3c). Note that since each retrieved item promotes the retrieval of successor states, this regime explains part of the contiguity effect: it predicts the recall of items encoded *after*, but not *before*, the just-recalled item (Fig. 3d).

All this raises a potentially confusing notational and conceptual point. The current model now involves two discount factors, because it uses serial retrieval to extend the temporal range of the encoded associations. The parameter γ refers to the timescale of associations formed when building an SR *at encoding time*. Sampling directly from this encoded SR i.i.d. (as in the previous section) estimates action values q reflecting that discount factor (i.e., in which future rewards lose value exponentially with rate γ; this happens in TCM-SR because the corresponding states are less likely to be retrieved). However, by performing iterative sequential retrieval from the same model, it is possible to extend this timescale at retrieval time to give more weight to later rewards, i.e. to estimate values reflecting a larger discount factor than the encoding γ. We denote the effective discount factor achieved at retrieval by $\tilde{\gamma}$. Using rollouts from a one-step model ($\gamma = 0$) to compute long-run action values is a familiar case of this construction; we develop further examples next.

Going beyond the extreme case of $\gamma = 0$ studied above, we now study the case of a general encoding timescale $\gamma > 0$. Here, the first retrieved item is a sample from the normalized SR at the candidate action, and each subsequent recall is a sample from the SR of the previous sample (Fig. 3e-h). Sequential retrieval again resembles a rollout, but due to the longer timescale of the SR, two consecutive samples can be separated by multiple rows. We call such a jumpy, state-skipping rollout a *generalized* rollout. To estimate action values using generalized rollouts, the sampled rewards can again be added to produce a sample of the cumulative return, exactly as in Eq. (4). Moreover, by specifying an interruption probability, the effective discount factor produced during retrieval can be controlled and corresponds to $\tilde{\gamma} = \gamma p_{\text{stop}} + (1 - p_{\text{stop}})$ (see Methods for details).

Why is this useful? Just as rollouts construct long-run predictions from a one-step model, generalized rollouts construct longer-run predictions from an SR. The timescale of the encoded world model may not be under the control of the agent. For example, it may be constrained by biological factors such as those governing neural plasticity (e.g., the temporal decay of intracellular concentrations that maintain eligibility traces) and/or by the statistics of experience, such as the timescales of the trajectories that they encounter. By contrast, we posit that p_{stop} is likely under the control of the agent. A chess player, for example, can decide how much time to spend simulating a particular sequence of moves (E. Russek, Acosta-Kane, van Opheusden, Mattar, & Griffiths, 2022). This highlights a remarkable feature of episodic memory: even if the learned associations at encoding have a short timescale (in the

extreme, a myopic SR with $\gamma = 0$, equivalent to a one-step transition model of the world), the retrieval phase can *extend* this timescale to implement any desired discount factor simply by continuously sampling successor memories. The effective discount factor thus increases as the simulated trajectories lengthen. This allows the agent to decouple the discount factor from timescale of the world model.

In sum, we have shown that when each retrieval completely resets the temporal context, action values can be estimated by accumulating sampled rewards drawn sequentially from episodic memory. This procedure implements a generalized rollout algorithm whose "skippiness" γ is specified by the drift rate at encoding, and whose effective discount factor $\tilde{\gamma}$ can be controlled by the probability of interrupting the retrieval process. Overall, the case of rollouts studied here, as well as the i.i.d. case studied previously, represent two distinct modes of operation of episodic memory, which TCM formalized as extreme settings of the parameter space. Next, we consider intermediate, more general — and likely more realistic — settings.

Data from free recall experiments suggest an intermediate regime

The previous sections examined two different strategies for predicting future events, corresponding to extreme settings in parameter space of TCM. The first section established that when retrieval does not modulate the temporal context, action values can be estimated via i.i.d. sampling from a model whose learned associations span future states over some temporal horizon. The second section showed that if retrieval completely resets the temporal context, sequential retrieval chains together predictions to extend this horizon, and action values can be estimated via generalized rollouts. Yet behavioral data from memory tasks suggest that human memory operates in neither of these two extreme modes, but rather displays signatures of both (Howard & Kahana, 2002). Indeed, the best fitting parameters describing context update in free recall experiments usually fall between the two extremes (i.e., $0 < \beta < 1$ in Eq. (1)), suggesting that each retrieval updates the temporal context but only *partially*. We now consider this intermediate regime and show that here, too, episodic memory can help compute action values.

The partially-updated temporal context at retrieval can be understood as a mixture of the current test context and the encoding context. For instance, immediately after the first retrieval, the context mixture enables sampling from either the SR of the queried action (the original context), or from the SR of the first sample (the retrieved context). Thus, the second sample either starts a new rollout with probability $1 - \beta$, or continues an existing rollout with probability β. Hence, β interpolates between the two distinct settings discussed above. Each action can be evaluated according to:

$$\hat{q}_{\tilde{\gamma}}(a) \propto \beta \sum_{i=1}^{N} \mathbf{r}^{\mathsf{T}} \mathbf{x}(S_i), \tag{5}$$

where $\beta > 0$ and $S_1, S_2, \ldots, S_N \sim p(s)$ are samples from the normalized SR $p(s)$ corresponding to some effective discount factor $\tilde{\gamma}$. Note that this estimator is only unbiased given an infinite number of samples and otherwise an underestimate (see Methods for details); however, a relatively large number of samples is sufficient for an estimate that's close to the truth (Fig. 4c,f).

The same insights gained in the previous sections apply here, including extension of the effective discount factor with a larger β (Fig. 4b,e) and the sample efficiency during decision making (Fig. 4c,f). Notably, due to the partial updating, implemented by setting $\rho = \beta = 0.5$, the effective discount fac-

tors as computed in the generalized rollout case (i.e., fully updating the temporal context with the last retrieval with $\rho = 0, \beta = 1$; lines in Fig. 4b,e) no longer capture the empirical sampling distributions under the same p_{stop} unless $p_{\text{stop}} = 1$ (dots in Fig. 4b,e). Recall that the larger the β, the further into the future later samples reach: i.e., β controls the degree to which the timescale at retrieval is extended (Fig. 4a,d). Thus both increasing β and decreasing the interruption probability extend the agent's effective temporal horizon for action evaluation, with the exception that the resultant sampling distribution may not correspond to any specific $\tilde{\gamma}$ as it is not necessarily an exponential distribution (e.g. red dots in Fig. 4b).

Hence in the more realistic setting of partial context updates, action values can still be estimated from retrieved episodic samples. This suggests that by modulating β (i.e. how drastically context is shifted to reflect each new sample), the agent can modulate its reliance on temporal abstraction vs constructive, rollout-based simulation, allowing it to balance the costs and benefits of these evaluation regimes depending on circumstances. This is similar to other examples in which, it has been argued, the brain adjusts its decision computations due to similar cost-benefit tradeoffs (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011; Nicholas, Daw, & Shohamy, 2022).

All simulations so far only consider the case of unlimited experience (i.e., multiple rounds of encoding; sampling from a converged SR). The next section extends our predictions to settings when only limited experience is available.

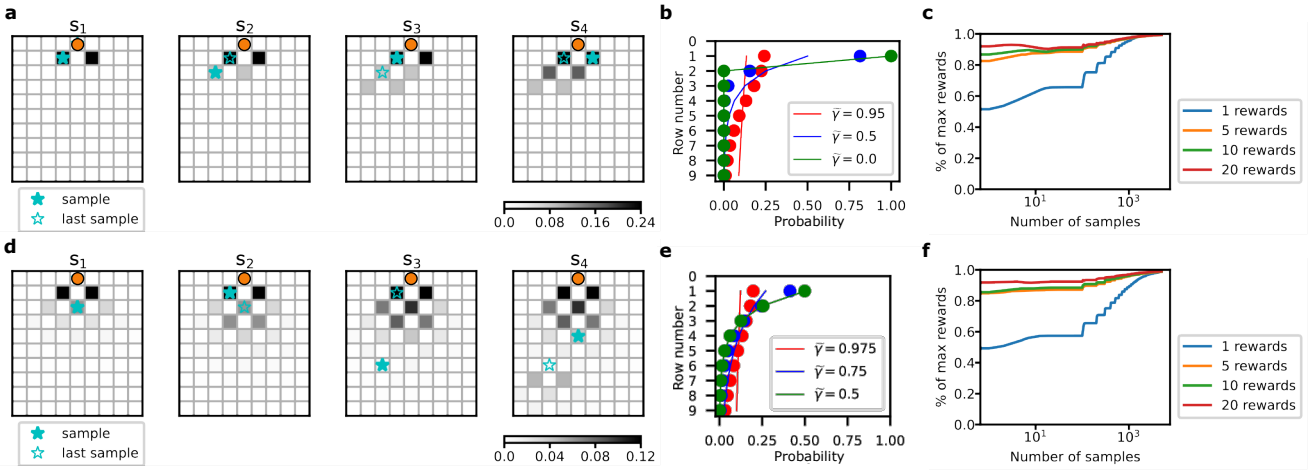

Fig. 4. An intermediate regime between i.i.d. sampling and rollouts. (a-c) Parameters: $\rho = 0.5$, $\beta = 0.5$, $\gamma = 0$. **(a)** An example sequence of memory retrieved when initiating the temporal context as the top-center state (orange circle) of a Plinko board. $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board, in this intermediate sampling regime (dots) versus generalized rollout (lines, same as Fig. 3b) given the same discount factors. We illustrate these distributions for three values of p_{stop} (0.05, 0.5, and 1), each leading to an effective temporal discount factor $\tilde{\gamma} = 1 - p_{\text{stop}}$. **(c)** We simulate an agent that evaluates two actions (at top-center state and the state immediately adjacent to the right), and then selects the action with the larger estimated value. The image shows the fraction of maximum rewards (y-axis) expected as more samples are drawn (x-axis, shown in log-scale), setting $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d-f)** As in **a-c**, but using parameters: $\rho = 0.5$, $\beta = 0.5$, $\gamma = 0.5$. See the online article for the color version of this figure.

With limited experience, retrieval is based on trajectories

Our simulations thus far assumed that the retrieval simulations we describe are preceded by an extensive encoding phase in which each state (location on the Plinko board) is encoded a large number of times. With repeated exposure, the associations formed between stimuli and contexts converge to the true steady-state SR (Gershman et al., 2012). Yet episodic memory is generally believed to be most useful, and perhaps most frequently used, when our experience with stimuli is limited. Indeed, this belief underlies most previous models of decision making informed by episodic memory (Lengyel & Dayan, 2007; Gershman & Daw, 2017; Ritter et al., 2018). We investigate this low-sample setting below, showing how unbiased value estimates are possible from states sampled along few experienced trajectories. In this case, the encoded model approximates the true task dynamics using this sparse set of encoded trajectories. Apart from that, the flexible prospection properties of the model remain the same.

Consider first that the agent has encountered only a single trajectory. TCM's account of encoding this trajectory into episodic memory is equivalent to the RL account for learning an SR from this same experience (e.g., via temporal difference learning; Gershman et al., 2012). This forms associations corresponding to the sequential contingencies experienced by the agent. If this encoding is followed by TCM retrieval, only states along the experienced trajectory will be retrieved (Fig. 5a, "Trial 1"), with states early in the trajectory having higher retrieval probability due to the temporal discount factor γ. Each subsequent stimulus is drawn from a distribution that depends on the degree of context update β. As before, this leads to a sampling scheme resembling i.i.d. sampling, rollouts, or both: but over a sparsely populated transition model consisting of only the encoded trajectory.

The extension to multiple experienced trajectories is straightforward. For instance, if an action has been executed twice, both trajectories should be encoded in the learned SR. Here, states belonging to either trajectory can be retrieved, with dynamics again depending on the degree of context updating (Fig. 5a, "Trial 2"). The learned SR comes to represent a composite of possible trajectories as experiences expand, eventually converging to the steady-state SR (Fig. 5a, right). Thus, TCM-SR predicts that retrieval is based on experienced trajectories when experience is limited; as the agent acquires more experience, our model predicts the limit cases studied in previous sections.

Note that the TCM predictions above share commonalities with previous proposals for how episodic memory might be used for decision making (Lengyel & Dayan, 2007; Gershman & Daw, 2017). In particular, Gershman and Daw (2017) proposed that agents store individual trajectories in memory, such that when a familiar state is encountered, action values can be computed by summing the rewards along a trajectory and averaging across trajectories: the very prediction given by $\beta = 1$ and $\gamma = 0$ in TCM-SR. However, our model also predicts sampling along novel trajectories. e.g. given trajectories ABDE and ACDF, our model predicts that rollouts along ABDF or ACDE are possible. For more general parameter settings, our model predicts state-skipping (if $\gamma > 0$) or backward jumping (if $\beta < 1$). Furthermore, states in the beginning of an experienced trajectory (predictions of the near-future vs. distant-future) are prioritized for retrieval due to discount factor. These differences result from the critical assumption of our model that agents retrieve individual states, rather than trajectories.

In sum, when limited experience is available, action values can be estimated by sampling states along (a composite of) previously experienced trajectories, facilitating few-shot estimation of action values as formalized in previous models. The next section considers additionally how preferentially retrieving emotionally salient stimuli, as observed empirically, can lead to faster evaluation.

Emotional modulation of memory yields bias-variance trade-off

The sections thus far formalize how temporal contingencies at encoding affect retrieval at a later time, and why retrieval dynamics in the TCM-SR are suited well for action evaluation. Yet, so far we have ignored another prominent feature of episodic memory that ought to affect retrieval-based evaluation during decision making: the psychological impact of states that are rewarded, compared to those that are not.

Episodic retrieval is strongly affected by signs that some stimuli are more important than others. For example, in the phenomenon of value-directed remembering, memory for high-reward stimuli is better than memory for low-reward stimuli (Stefanidi, Ellis, & Brewer, 2018). Even when reward is not signalled overtly, signals that some stimuli should be prioritized promotes their retrieval (Mather, Clewett, Sakaki, & Harley, 2015). In fact, stimuli that attract processing resources are remembered better even when retaining them in memory is not obviously goal-congruent. One well-known example is that emotionally salient stimuli are retrieved preferentially even when participants have no external incentive (Yonelinas & Ritchey, 2015). Formal models of emotionally enhanced memory have attributed the effect either to a differential learning rate (Talmi et al., 2019; Cohen & Kahana, 2019) or differential information decay (Zhou, Guo, & Yu, 2020) during encoding. Given that emotional salience modulates episodic memory, it follows that it should also modulate action evaluation in TCM-SR. We examine this issue below. For present purposes, we gloss over the many differences between emotional stimuli, prioritized stimuli, and rewards and punishments with varied magnitude, referring to all of them as 'emotionally salient' or 'important' states, and speak generally about 'emotional modulation' to refer to their effect on memory (Talmi, Kavaliauskaite, & Daw, 2018).

To study the effect of emotional modulation in the Plinko game, we first note that when there is a single state with nonzero reward, the optimal actions are the ones capable of reaching that state. But if samples are prioritized based purely on temporal contingencies, that key state will be sampled very rarely among the many background states, and the agent might need a large number of samples to discover which actions are most likely to obtain it. Indeed, this sort of "needle in the haystack" effect accounts for the relatively poor performance for TCM-SR with few samples in our simulations thus far (Figs 2c,f; 3c,g; 4c,f). While performance can be improved by drawing more samples, this longer deliberation can be costly in terms of time and effort.

A potentially more effective way to find the best action might be to bias sampling toward the most relevant states (here, the goal), even if biasing the sampling procedure might lead to biases of the estimated payoff q (Lieder et al., 2018). Here we suggest that such favorable biasing can be accomplished by (and, conversely, helps to justify) emotionally modulated retrieval, where we operationalize emotionally salient states as those with unusually large (or small) rewards.

Computationally, an emotionally modulated retrieval results in a *bias-variance trade-off*: preferential retrieval of emotionally-salient stimuli disproportionally influences the final evaluation, resulting in an *estimation bias*, that is, either an over- or an under-estimation of true action values. When most samples come from the smaller set of "important" states, samples are less varied, resulting in lower *estimation variance*. Consequently, fewer samples are required to be reasonably precise and fewer retrievals are needed to arbitrate between competing actions. Nevertheless, the eventual decision can be suboptimal, in the sense that the action selected may not be the one associated with most reward. The larger the retrieval preference towards emotionally-salient stimuli, the larger the estimation bias and smaller the variance – thus, a bias-variance trade-off. A similar observation has been previously made in bandit

settings (Lieder et al., 2018). Here, we extend this class of Monte Carlo models to sequential tasks, and show that the same observation applies. The main contribution of this section is that TCM-SR allows us to expose how action evaluation in sequential tasks relates to episodic memory, helping to rationalize emotional memory effects.

To illustrate this effect in our Plinko environment, we follow previous modeling work and employ a higher learning rate to encode emotionally salient stimuli into memory (Talmi et al., 2019). This means that the learned SR will be skewed towards the rewarded states (Fig. 5b). Consequently, in the Plinko game, states associated with rewards are sampled more frequently during retrieval (Fig. 5d, right). Without emotional modulation, rewarded states would have been sampled only rarely (Fig. 5d, left). The consequences of operationalizing emotional modulation in TCM-SR such that rewarded states are encoded with a larger learning rate are threefold. First, the action value estimates no longer converges to the correct action values. Second, convergence will be faster, resulting in a bias-variance trade-off (Fig. 5f, compare with Fig. 5e). Third, if the agent selects actions according to this regime, a higher fraction of rewards can be obtained for a given number of samples (Fig. 5c), suggesting that biased retrieval can be more favorable, in terms of ultimately guiding choice, than unbiased retrieval.

Retrieving a learned context allows backward sampling

Starting from a simplified model of episodic memory, the previous sections examined the effect of various known properties of episodic memory on action evaluation and choice. A key insight of the model is that forward contiguity gives rise to predictive state rollouts. However, in list learning data, contiguity also runs in reverse: stimuli are also more likely to be recalled if they were experienced *before* as well as after the just-recalled stimulus (Fig. 1c). From the perspective of mental simulation, this property seems counterintuitive: in our example, it corresponds to rollouts in which the Plinko ball, impossibly, runs uphill. Here we suggest that this type of reversible simulation is actually adaptive for many tasks other than Plinko.

The reason our simulations thus far reproduced only the forward contiguity (Fig. 3d,h) is because of one final simplification that have not yet been re-examined. We have assumed that when a memory is retrieved, it updates the temporal context with a static, task-independent representation of the retrieved stimulus ($\mathbf{x}_t$ in Eq. (1); Fig. 1h). In contrast, the original TCM model explains the two-sided contiguity effect by positing that context update caused by retrieving a stimulus is not static and task-independent; rather, memory retrieval updates the temporal context with a dynamic, task-dependent representation, a representation that changes each time that stimulus is experienced. In particular, TCM assumes that the temporal context is updated by a retrieved *context* associated with a given stimulus, instead of being updated by the stimulus representation $\mathbf{x}_t$ itself. Formally, the temporal context is updated during retrieval according to $\mathbf{c}_i = \rho\mathbf{c}_{i-1} + \beta\mathbf{c}_i^{\text{IN}}$, where $\mathbf{c}_i^{\text{IN}} = \mathbf{M}\mathbf{x}_i$, i.e., $\mathbf{c}_i^{\text{IN}}$ is the column of the SR indexed by the stimulus. Importantly, this modification only concerns retrieval; it does not affect encoding, where the context still evolves according to Eq. (1) with $\mathbf{c}_i^{\text{IN}} = \mathbf{x}_t$. Thus, whenever stimulus S is retrieved, the temporal context is updated by the context associated with S, which is similar to the contexts associated with both subsequent stimuli and preceding stimuli. This results in the classic, bidirectional contiguity effect, often reported in list learning experiments (Fig. 6a,b).

What might be the adaptive purpose of a bidirectional pattern of retrieval? This pattern might appear counterintuitive since an action value is determined by the expectation of *future* rewards. Indeed, in our previous simulations, action values were estimated via strictly forward-looking rollouts, i.e., in

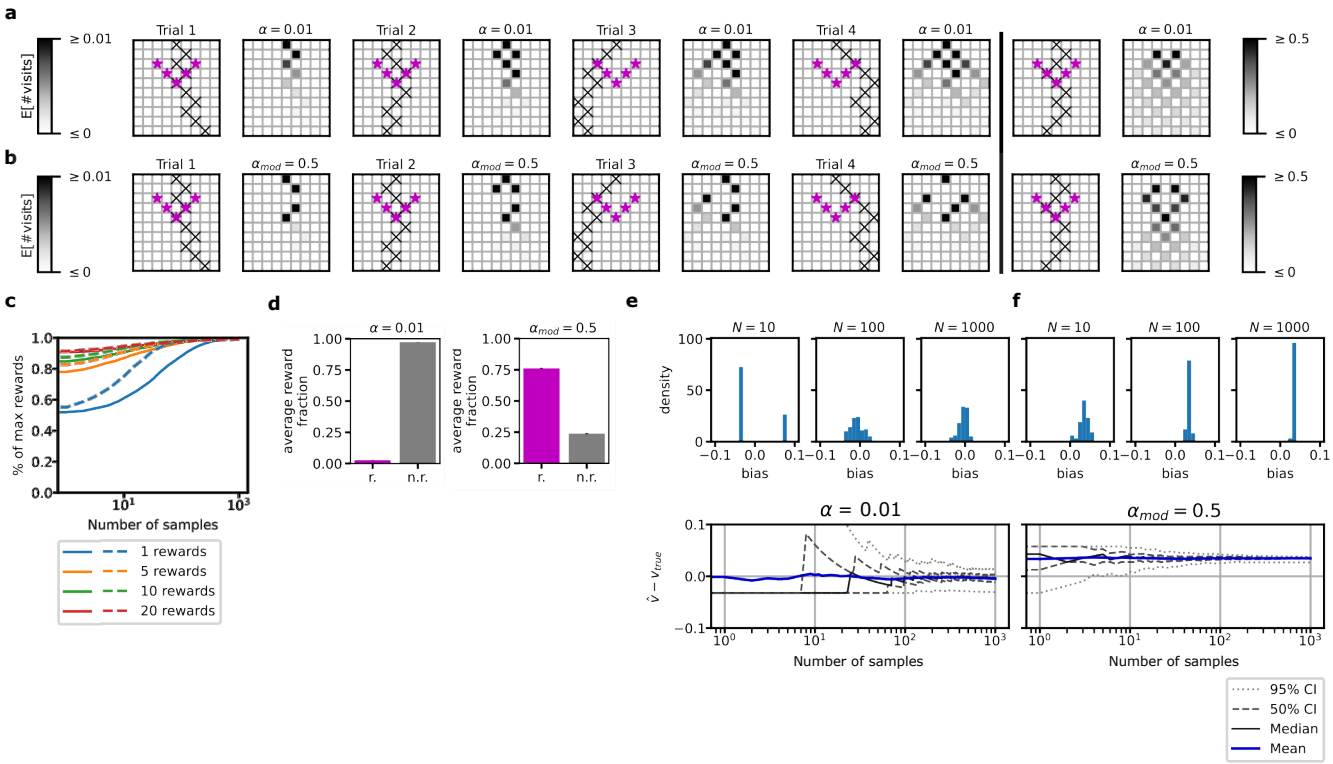

Fig. 5. Retrieval with limited experience and with emotional modulation. (a) Each pair of panels represent a 'trial' where the agent observes the trajectory that follows a single action (left, each visited state denoted in x's, and each rewarded state in *) and the ensuing learned SR after convergence ($\gamma = 0.9$) (right). The impact of accumulated experience is shown by comparing Trials 1, 2, 3, 4, and Trial $\to \infty$, presented in the five pairs of panels going from left to right, all without emotional modulation ($\alpha = 0.01$). **(b)** The same as (a) but now with emotional modulation ($\alpha = 0.01$ for unrewarded states and $\alpha = 0.5$ for rewarded states). **(c)** We simulate an agent that evaluates two actions (at top-center state and the state immediately adjacent to the right) using the procedure from Eq. (2), with (dashed lines) and without emotional modulation (solid lines). The agent selects the action whose estimated value is larger. The image shows the fraction of maximum rewards (y-axis) expected as more samples are drawn (x-axis, shown in log-scale), setting $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d)** Average fraction of sampled states with (r.) and without a reward (n.r.). Error bars indicate s.e.m. across experiments. Left: no emotional modulation. Right: with emotional modulation. **(e-f)** Bias and variance convergence based on a single observation for $\gamma = 0.9$ without emotional modulation (e) and with modulation (f). Top: mean bias of estimates based on 10, 100, 1000 samples. Bottom: mean discrepancy between the true value and the estimated value as a function of number of samples on a log scale. See the online article for the color version of this figure.

terms of future rewards alone. With a bidirectional pattern of retrieval, sampling no longer respects the temporal order of events experienced during encoding. We argue that, in most realistic tasks, the experienced temporal ordering of events is only one of all possible orderings; most state transitions experienced in one order can also be traversed in the reverse order. Although this is never the case in Plinko (since gravity strictly pulls the ball downward), it is often the case in tasks like spatial navigation. In other tasks (like chess), many actions are reversible while some others (e.g. capturing a piece) are not. An agent operating in the low-data regime can leverage this reversibility to infer, after experiencing state A followed by B (A $\rightarrow$ B), that transitioning from B to A (B $\rightarrow$ A) is likely also possible. Similarly, given only a few experiences in an environment, the agent can infer an exponentially larger number of unexperienced but likely possible trajectories (e.g., extrapolating A $\rightarrow$ B $\rightarrow$ C to not only C $\rightarrow$ B $\rightarrow$ A, but also A $\rightarrow$ B $\rightarrow$ A, C $\rightarrow$ B $\rightarrow$ C, etc), which in turn generalizes action evaluation. Ideally, the relative strength of forward vs. reverse continguity (biased forward in classic list learning data) would reflect the chance that a newly encountered action is reversible; this might, in turn depend on context.

As an example, consider an experience where an action is followed by A $\rightarrow$ B $\rightarrow$ C, and that the agent retrieves stimulus B. The generalized rollout studied previously permits a subsequent sample of C but not A due to its strictly forward-looking nature. By assuming that the retrieved stimulus updates the temporal context with a retrieved context, the next retrieval can be either C or A, consistent with the assumption of reversibility. This can improve sample efficiency, as multiple (plausible) sequences of events can be simulated despite having encoded only a single experience.

To simulate this scenario, we modified our Plinko task to eliminate gravity so that the agent can move diagonally in any direction, and it may start from any board position. The agent's goal is to select an adjacent state to move into, after which each subsequent states is selected at random from between the neighbors of the previous state. In this "reversible Plinko", the value of each state is affected by all rewards on the board, with nearby rewards contributing a higher weight to the value. If an agent only experiences top-to-bottom trajectories in the reversible Plinko task, and uses a strictly forward-looking rollout to evaluate actions, the resulting values will correspond to values under the gravity-bound Plinko rules. While they are in line with the agent's experiences, they do not match the true values under the reversible Plinko rules (Fig. 6d). A retrieved context aids the agent to go beyond unidirectional experience and correctly estimate the values for the reversible Plinko (Fig. 6c). Hence we suggest that the ubiquitous human tendency to recall stimuli in the opposite order than experienced may allow a more efficient use of one's limited experience.

Discussion

Summary of Findings

We proposed TCM-SR, a process-level model of how episodic memory informs decision making. What is extraordinary about this model is that it applies — essentially unmodified — a standard theory of episodic memory function to an entirely different setting: that of sequential decision tasks. The resulting hybrid implements and extends a prominent class of theories of how the brain makes sequential decisions. The proposed grounding of decision variables and choices in specific episodic retrieval dynamics brings to bear much of our knowledge of episodic memory, including a richly developed behavioral and neural framework. It also suggests many testable predictions for choice manipulation

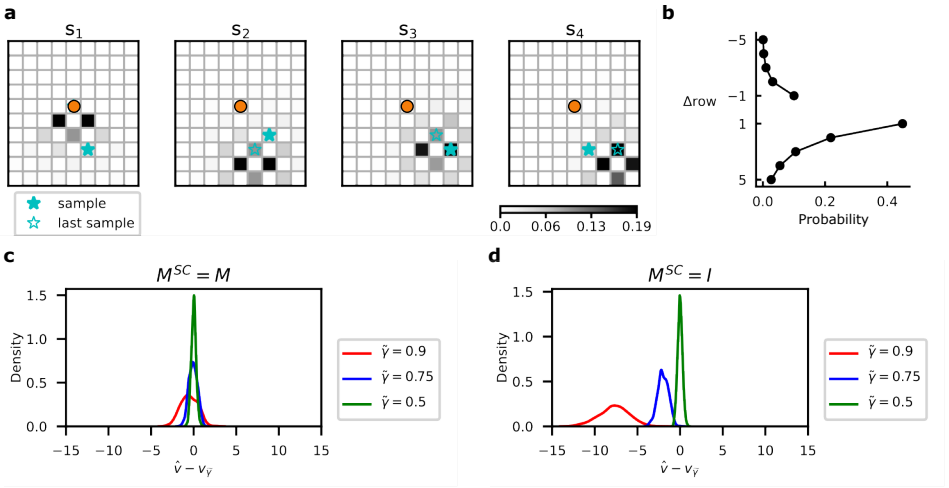

Fig. 6. Retrieving a learned context allows backward sampling. **(a)** An example sequence of memory retrieved when initiating the temporal context with the state shown as an orange circle, and using $\gamma = 0.5$. $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Contiguity curve implied by the sampled states with respect to their corresponding row number given $\gamma = 0.5$ (zero omitted). Note that both forward and backward sampling are predicted. **(c)** Distribution of estimation error using the SR as the feature-to-context association matrix. Errors are computed as the difference between the sampling-based value estimation and the ground-truth value in a reversible MDP (i.e., a grid world rather than a Plinko game). **(d)** As in (c), but using the identity matrix as the feature-to-context association matrix (as in the previous simulations). See the online article for the color version of this figure.

via manipulations known to affect memory encoding or retrieval. Conversely, the theory rationalizes seemingly arbitrary features of episodic memory — such as emotional memory effects and the bidirectionality of temporal contiguity — which appear counterintuitive from the traditional RL perspective, but turn out to be adaptive for choice.

Our model inherits from TCM a drifting temporal context that integrates the agent's recent experience during memory encoding and guides retrieval. The agent evaluates actions by retrieving memories that correspond to task's states and the rewards associated with them, according to the prediction of TCM. As we have shown, such recursive retrieval implements a parameterized family of sampling algorithms that, when applied to sequential decision problems, gives rise to action value estimates. Our model thus provides a novel mechanistic account of model-based evaluation, incorporating aspects of both SR theories and iterative rollout-based planning, the hallmarks of both of which have been previously seen in neural and behavioral data (Momennejad et al., 2017; Stachenfeld et al., 2017; E. M. Russek et al., 2017; Momennejad, Otto, Daw, & Norman, 2018; Mattar & Daw, 2018; E. M. Russek et al., 2021; Liu, Mattar, Behrens, Daw, & Dolan, 2021). Crucially, many previous ideas (both theoretically justified or empirically observed) about the role of episodic memory on decision making arise naturally as subcases of our model.

The theoretical derivations and simulation results we presented established that TCM can compute decision variables in stylized tasks, but we have not focused on applying these insights to specific experiments. This is because the two classes of theories our account merges — TCM and SR — are already supported well in each domain, with large bodies of experiments and simulations, which we do not repeat. In particular, our model inherits TCM's account of a panoply of list learning phenomena (e.g.,

primacy, recency, and contiguity effects; Howard & Kahana, 2002; Sederberg et al., 2008; Polyn et al., 2009; Talmi et al., 2019). Meanwhile, since its strategy encompasses model-based and SR-based choice, it can explain the full range of behavioral phenomena that suggest that the brain recruits cognitive maps or world models in decisions (e.g., nimble replanning, revaluation and transfer, and credit assignment in multi-step MDPs; Daw et al., 2005; Keramati et al., 2011; E. M. Russek et al., 2017). It also explains occasional slips of action consistent with the use of an SR (Momennejad et al., 2017; Piray & Daw, 2021). Furthermore, the decision-time sampling process is broadly consistent with neural results showing that these types of model-based choices are at least sometimes accompanied by replay or reinstatement reminiscent of rollouts (Pfeiffer & Foster, 2013; Momennejad et al., 2018; Mattar & Daw, 2018).

TCM-SR produces a number of new and untested predictions in both the decision and memory domains. We have argued that recall biases like contiguity and emotional memory enhancement have corresponding effects on choices. If deliberative evaluation is indeed grounded in free recall, these decision effects should be quantitatively comparable to their counterparts measured in list learning, that is, model fits should reveal they reflect the same within- and between-individual best-fitting parameters. Additionally, other manipulations that affect memory, like proactive and retroactive interference, should also have concomitant effects on decisions via enhancement or suppression of particular states and/or outcomes. Conversely, the rationalization of these parameterized memory effects as enabling more efficient choice in various settings suggests that the parameters governing them are potentially malleable, adapting to the statistics of the study material to optimize choice (Nicholas et al., 2022). For instance, when states or study items reflect non-reversible environmental dynamics, a rational RL agent would be expected to dial back the reversibility assumption when learning an SR. In turn, this may also attenuate the backward contiguity effect as measured via memory recall. In another example, the usefulness of emotional memory enhancement (Fig 5) at improving choices strongly depends on the statistics of the emotionally salient rewards, such as their sparsity. If the degree of emotional enhancement is normatively adjusted to reflect its circumstantial suitability, this may also impact memory. This line of reasoning may suggest an explanation for findings in the memory domain showing that these effects are modulated by how emotional and neutral items are clustered during study (Talmi et al., 2019).

Successor Representation and Generalization to Sequential Decision Problems

Our account is consistent with a class of models that use a handful of selective samples to construct decision variables (Plonsky et al., 2015; Bornstein et al., 2017; Lieder et al., 2018). While these models address a number of empirical phenomena in choice, and suggest a broad analogy with episodic recall, they consider only the special case of one-step decision problems and incorporate few insights about known episodic memory mechanisms. We argue that incorporating the effects of contiguity, established in episodic memory research, is key to extending sampling models beyond bandit problems into the sequential realm — a broader, more realistic, and more challenging classes of decision making problems. Bandit-like evaluation then arises as a special case in our model, allowing it to both incorporate the results from previous models while extending many of these ideas (like bias-variance tradeoffs in the small-sample domain) to the sequential domain.

Another crucial ingredient for generalizing from one-step bandits to sequential decision problems is the successor representation (SR). Prior work suggests the SR is biologically plausible to learn, given its

ability to explain patterns in human behavior (Momennejad et al., 2017; E. M. Russek et al., 2017) and in the activity of hippocampal neurons (Brea, Gaál, Urbanczik, & Senn, 2016; Stachenfeld et al., 2017; Garvert, Dolan, & Behrens, 2017). Building on the previously established equivalence between TCM encoding and SR learning (Gershman et al., 2012), our model extends the insight to formalize, for the first time, how TCM retrieval amounts to sample-based action evaluation in sequential settings. This temporally extended sampling process marks a departure from the canonical view in SR models from both neuroscience and AI, where state values are instead computed instantaneously via a dot product $\mathbf{v} = \mathbf{M}_\gamma \mathbf{r}$ (Dayan, 1993; Momennejad et al., 2017; E. M. Russek et al., 2017). Apart from drawing a connection with episodic retrieval, our sampling variant places it in the context of rollout-based models, allows the model iteratively to construct forecasts beyond its innate temporal scope.

Relation to Existing Models

Episodic Control

Previous episodic control models in RL have often been stylized in design, treating "episodic" memory chiefly as a store of individual instances. Our model improves on them by incorporating known mechanistic details of episodic memory. For example, a recent model of this class assumes that action values are computed by considering all relevant trajectories the agent has experienced (Gershman & Daw, 2017). In contrast, TCM-SR assumes that trajectories are only encoded indirectly via state-context associations, while maintaining the ability to simulate rollouts during retrieval. The two achieve similar action values, except in cases where our model retrieves rollout samples by merging different trajectories. Importantly, our model explicates the *process* of retrieval, predicting that (1) individual states rather than trajectories are retrieved, and (2) retrieved samples may skip over intermediate states. Future work should investigate whether these predictions better describe how humans evaluate actions.

Episodic vs. Model-Based Evaluation

Previous work has often distinguished between at least two types of decisions, model-based (goal-directed, deliberative) and model-free (habitual, automatic) (Daw et al., 2005). It remains unclear, though, both what is the exact neural and computational basis for the planning-like behaviors associated with model-based control, and whether any contributions of episodic memory to choice are distinct from this. The recruitment of constructive rollouts in our model suggests an intriguing possibility that what has been attributed to model-based evaluation might be wholly or partially explained by episodic retrieval. Several lines of empirical results support this hypothesis: patients with hippocampal damage tend to exhibit a lower degree of model-based control (Gutbrod et al., 2006; Vikbladh et al., 2019); the hippocampus is often active in tasks requiring model-based control (Bornstein & Daw, 2013); and finally, inactivating the hippocampus in rats causes their behavior to shift from model-based to model-free (Miller, Botvinick, & Brody, 2017).

All this casts doubt on the influential hypothesis that episodic control represents a distinct "third way" that departs from the model-based vs. model-free dichotomy (Lengyel & Dayan, 2007). Instead, TCM-SR predicts that episodic retrieval can give rise to evaluations resembling either episodic or semantic model-based control, depending on the amount of experience the agent has been able to accumulate, which determines the sparsity of its memory representation. Given ample experience, like a world

model, SR only retains statistical commonalities across experience, and thereby facilitates model-based rollouts for action evaluation. This is consistent with the complementary learning systems account whereby semantic representations are obtained by extracting regularities across individual experiences via a process of consolidation (O'Reilly, Bhattacharyya, Howard, & Ketz, 2014; Kumaran, Hassabis, & McClelland, 2016). When experience is limited, SR represents individual trajectories, and recall largely follows them as experienced. Therefore, despite different formalizations, TCM-SR in fact agrees in spirit with a prediction of the earlier model (Lengyel & Dayan, 2007) that agents might rely more on evaluations grounded in distinct episodic records when experience is limited, giving way to control based on a more statistical model as more experience is gathered. Together, the empirical and modeling evidence suggest a close link between episodic and model-based evaluation as a function of experience. Importantly, these considerations point to the importance of future investigation in a memory regime that has not seen much study in list learning: how episodic recall is affected by repeated exposure to lists with overlapping items (Gershman et al., 2012), analogous to the hypothetical transition from individual trajectories to an SR in our model.

Additional computational considerations

TCM has been extended in a series of successor models, such as TCM-A (Sederberg et al., 2008) and CMR (Polyn et al., 2009). These extensions are all compatible with our approach. Indeed, the additional features of memory addressed there — particularly clustering of recall not just by temporal context but also semantic and source-memory similarity — may have unappreciated consequences in the decision domain that TCM-SR does not yet address.

A major computational simplification of our model is that we treat candidate decisions as only a single choice made at the first step, after which the model predicts further states as though the task played out passively, like a falling Plinko ball. For most sequential decision tasks, such as mazes, actions must additionally be chosen at each subsequent step, and these choices impact the value of the action at the first step (Sutton & Barto, 2018). Like other SR-based models, our present model does not fully solve this broader class of tasks. That said, the RL literature typically considers the action evaluation problem our model does solve (termed "policy evaluation" in the RL literature) to be the key subproblem for addressing the more general policy optimization problem. Critically, even in a task where every step involves a decision, an SR or other model can learn the world's dynamics *under a particular assumption about which actions are chosen*, called a policy. This turns a problem with decisions at every step to one, like Plinko, in which the state evolves passively (because subsequent decisions are assumed known). In this way, the agent can evaluate the consequences of any candidate action choice at any particular step, temporarily assuming the others are fixed. This local "policy improvement" process can iterate, with the SR continually relearned, recomputed, or adjusted, to reflect improved policies as learning proceeds. Future work could address a number of alternative approaches to this problem, including nonlinear SR variants that approximate maximization at intermediate steps (Piray & Daw, 2021), or rollout/retrieval dynamics that include some degree of maximization biasing the choice at each rollout step (E. M. Russek et al., 2017), similar to traditional value iteration algorithms.

To conclude, the contribution of TCM-SR is in bridging two distinct families of theory and cognition, each explaining, but hitherto separately, a large body of empirical data. We have suggested how to address decision problems using the tools of memory research, and how to apply normative insights from decision problems back to emory mechanisms. By doing so, and pointing to a number of possible

new avenues rife for exploration, we hope to bring research literatures that have evolved separately closer to each other.

Acknowledgments

This work was supported in part by the U.S. National Science Foundation, grant IIS-1822571 (ND), part of the Collaborative Research in Computational Neuroscience program.

Methods

Task Details

We wish to formalize how action values in sequential decision problems can be estimated via episodic memory samples, taking into account several known properties about retrieval dynamics in free-recall. We illustrate this process with a temporally extended game called Plinko (Fig. 1a). This game is an analogy to a generic sequential decision task where each action leads to a stochastic sequence of states, and where each state can be reached by potentially multiple actions. We selected the game of Plinko because it allows the visual depiction of the sequential retrieval process in a didactic manner (as rows represent both time and space). The game should therefore not be interpreted literally as choices in a real game of Plinko are unlikely to be guided by episodic memory.

In Plinko, the agent chooses a place on the top row of the board to drop a ball. At each step, the ball falls diagonally either to the left or to the right by one row, with equal probability. If the ball is at the left edge of the board, it falls diagonally to the right with probability 1. Similarly, if the ball is at the right edge, it falls diagonally to the left with probability 1. A trial starts when the ball is dropped on the top row and ends when the ball reaches the bottom of the board. Rewards, which are scattered across the board, can be collected whenever they are hit by the falling ball. An experiment is composed of multiple trials having a single reward placement.

The agent must decide where to drop the ball in order to collect as much reward as possible. To decide, we assume that the agent estimates the goodness of each candidate location along the top row so as to support effective decision making. The goodness of each action is the total expected reward resulting from that action. We further assume that the agent has had prior experience with this task stored in episodic memory. Whenever the agent needs to select an action, it evaluates each candidate action by retrieving episodic memories. No other source of information is available to the agent.

Formal setting

We formalize this problem as Markov Reward Process (MRP) – a discrete-time stochastic process which extends a Markov Chain by adding a reward to each state. Unlike in a Markov Decision Process (MDP), the state dynamics in an MRP are not under control of the agent (note that an MRP is obtained by fixing the agent's policy in an MDP). Thus, in an MRP we are typically concerned with the problem of reward

prediction (e.g., how much reward will follow from each state on the top row of Plinko) and not *control* (e.g., which actions to select at each Plinko state). Nonetheless, we use the notation of MDPs in this paper to remain consistent with the decision making literature.

The task is formalized by a 5-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ denotes the set of states, and $\mathcal{A}$ denotes the set of actions corresponding to each state in the top row, i.e., $\mathcal{A} = \{s_{a=1}, s_{a=2}, \ldots, s_{a=|\mathcal{A}|}\}$. $\mathcal{P} : \mathcal{S} \mapsto \mathcal{S}$ is the Markov transition function that defines the probability distribution $\mathcal{P}(s'|s)$ of transitioning from state s to state s'. $\mathcal{R} : \mathcal{S} \mapsto \mathbb{R}$ is the reward function $\mathcal{R}(s)$ specifying the reward magnitude received upon visiting state s, and $\gamma \in [0, 1)$ is the discount factor that controls the temporal horizon of computations by reducing the importance of rewards in distant future.

The goal of the agent is to choose the action that maximizes the cumulative discounted return $G = \sum_{t=1}^{\infty} \gamma^t \mathcal{R}(S_t)$, where S_t is a random variable denoting the state at time t. As a shorthand, R_t is a random variable denoting the reward obtained at time t. Upon selecting an action, the agent experiences a sequence of states, each drawn with probability $\mathrm{P}\big(S_{t+1} = s' \mid S_t = s\big) = \mathcal{P}(s'|s)$. This gives rise to a "trajectory" given by $S_1, R_1, S_2, R_2, S_3, R_3, \ldots, S_H, R_H$, where $H = \frac{|\mathcal{S}|}{|\mathcal{A}|}$ is the number of rows in the Plinko board. After reaching the bottom of the Plinko board, we assume that the ball is transferred to an unrewarded, absorbing state outside the board.

The value of state s, denoted $v(s)$, is defined as the expected return when starting in s: $v(s) = \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k R_{t+k} \mid S_t = s\right]$. The value of taking action a, denoted $q(a)$, is defined as the expected return when taking action a in the beginning of a trial: $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t R_t \mid A = a]$, and can also be defined strictly in terms of states: $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t \mathcal{R}(S_t) \mid S_1 = s_a]$. We refer to $q(a)$ as "action value".

In order to select an action, the agent estimates $q(a)$ for each candidate action. The field of RL describes various methods for estimating $q(a)$, broadly divided into model-free and model-based methods. Model-free methods are those where the agent learns to estimate $q(a)$ directly from experience. The classic temporal difference (TD) algorithm, for example, iteratively updates the agent's estimate $Q(a)$ as $Q(a) \leftarrow Q(a) + \alpha\left(\gamma R_1 + \gamma^2 v(S_2) - Q(a)\right)$ whenever action a is performed. In model-based methods, in contrast, the agent uses a model of the world (i.e., an estimate of $\mathcal{P}$ and $\mathcal{R}$) to estimate $q(a)$. If both $\mathcal{P}$ and $\mathcal{R}$ are perfectly known, the agent can generate a plausible trajectory $S_1, R_1, S_2, R_2, S_3, R_3, \ldots, S_T, R_T$ resulting from a, where $S_1 = s_a$, $S_{i+1} \sim \mathcal{P}(.|S_i)$, and $R_i = \mathcal{R}(S_i)$. Each such trajectory is called "rollout", alluding to the fact that states (and rewards) are sampled recursively (Tesauro & Galperin, 1996). The total discounted reward along a rollout trajectory is a Monte Carlo estimate of the action value, i.e., $Q(a) = \sum_{i=1}^{H} \gamma^i R_i$.

Successor Representation

Consider the one-step state-transition matrix $\mathbf{T} \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ whose (i, j)-th entry T_{ij} corresponds to the probability of transitioning from state i to state j: $T_{ij} = \mathrm{P}\big(S_{t+1} = s_j \mid S_t = s_i\big)$. Consider also the one-step reward vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ whose k-th entry r_k corresponds to the reward present in state k. The

value function can be expressed in vector form as:

$$\begin{aligned}\mathbf{v} &= \mathbf{T}^1\mathbf{r} + \gamma^1\mathbf{T}^2\mathbf{r} + \gamma^2\mathbf{T}^3\mathbf{r} + \cdots \\ &= \left(\sum_{k=0}^{\infty} \gamma^k \mathbf{T}^k\right)\mathbf{T}\mathbf{r} \\ &= \left(\mathbf{I} - \gamma\mathbf{T}\right)^{-1}\mathbf{T}\mathbf{r}.\end{aligned} \tag{6}$$

The matrix $\left(\mathbf{I} - \gamma\mathbf{T}\right)^{-1}\mathbf{T}$ is the successor representation (SR), denoted by $\mathbf{M}_\gamma$ (Dayan, 1993). The (i, j)-th entry M_{ij} corresponds to the expected sum of future visits to state j from state i, discounted according to γ. The SR can be learned directly from experience using TD learning. If the "true" SR is available to the agent, all state values can be estimated simultaneously by $\mathbf{v} = \mathbf{M}_\gamma\mathbf{r}$.

We note that our definition differs from the more traditional $\left(\mathbf{I} - \gamma\mathbf{T}\right)^{-1}$. The inclusion of an additional $\mathbf{T}$ in the definition simply indicates that the value of some state does not depend on rewards present in that same state, but only on rewards present in future states. This is a matter of definition, and is equivalent to stating that, in Plinko, rewards are collected upon *entering*, but not *exiting* a state.

Temporal Context Model

Overview

TCM is a computational model of episodic memory designed to explain human behavior in free recall experiments. In a free recall experiment, subjects first study a list of items (often words or word-pairs) presented one at a time for a brief period. Then, subjects are asked to recall the items in any order they wish. TCM models memory encoding through associative learning between recently studied stimuli and the temporal context at presentation time. The learned associations then guide retrieval - specifically, the stimulus most similar to the current context is retrieved (Howard & Kahana, 2002).

TCM posits that each stimulus is represented by a feature vector, while the abstract temporal context is formalized as the combination of recently experienced stimuli. The temporal context in TCM is updated during both the encoding and retrieval of memories, while specifying the recall probability of each individual stimulus.

TCM predicts a *recency effect*, as observed in human free recall: since the temporal context at the beginning of recall is most similar to the one maintained near the end of the list during encoding, the last few stimuli are more likely to be recalled by association (Fig. 1c, left). Indeed, when a distractor task is introduced to delay recall, the recency effect is significantly attenuated (Greene, 1986), likely because the context when retrieval started has evolved away from the end-of-list context.

TCM also predicts a *temporal contiguity effect* observed in human free recall. Kahana (1996) used the lag conditional response probability (lag-CRP) to quantify such effect. The lag-CRP is computed as the conditional probability that, given the most recently recalled stimulus and its serial position i during encoding, the subsequently recalled stimulus comes from serial position $i + j$, where j is a signed integer representing the lag. Crucially, TCM captures the temporal contiguity effect using its evolving temporal context. At an arbitrary point in time, the context is composed of two components – one

that encodes the associations formed during the experiment thus far, encompassing both encoding and retrieval, and one that's primarily associated with the most recently experienced stimulus. The former, called experimental context, is part of the temporal context both before and after the recent stimulus presentation; but the latter, called pre-experimental context, is only introduced at the moment of stimulus presentation (or recall). Thus while the former shares similarity to other stimuli as a symmetrical function around the stimulus, the latter is likely dissimilar to all preceding stimuli and is only incorporated in ensuing contexts. As a result, TCM predicts lag-CRP to be asymmetrical, with higher probability to recall subsequent stimuli than preceding ones (Fig. 1c, right)

Notably, temporal contiguity effect is approximately scale-invariant - it has been observed both within individual lists and across lists spanning extended amount of time (e.g. Howard & Kahana, 1999, Howard, Youker, & Venkatadass, 2008), suggesting maintenance of temporal contexts over multiple timescales and entities.

In summary, the temporal context and its evolution dynamics in TCM provides an algorithmic hypothesis of how human episodic memory, especially aspects that are captured in free recall paradigms, manifests specific retrieval dynamics contingent on the relative temporal order.

Formal Model Description

Let $\mathbf{c}_t$ denote the experimental context at time t and $\mathbf{c}_t^{\text{IN}}$ denote the pre-experimental context at t, both as column vectors. Additionally, let $\mathbf{x}(S_t)$ be the feature representation of the stimulus encoded at time t, e.g., a one-hot $|S|$-dimensional vector. As a shorthand, we write $\mathbf{x}_t$ in place of $\mathbf{x}(S_t)$. Likewise, we denote the stimulus retrieved at time i as $\mathbf{x}_i$. i.e., we use $t \in \{1, 2, \ldots, T\}$ to index encoding time and $i \in \{1, 2, \ldots, N\}$ to index retrieval steps. Additionally, TCM achieves associative learning via a context-to-stimulus matrix $\mathbf{M}^{\text{CS}}$ and a stimulus-to-context matrix $\mathbf{M}^{\text{SC}}$. The learning and update rules are summarized in Table 1.

Table 1: Summary of the Temporal Context Model

Name	**Expression**	
Context-to-Feature Matrix	$\mathbf{M}^{\text{CS}} = \sum_t \mathbf{x}_t \mathbf{c}_t^{\mathsf{T}}$	(7)
Input Context	$\mathbf{c}_t^{\text{IN}} = \mathbf{M}^{\text{SC}} \mathbf{x}_t$	(8)
Context Update	$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{c}_t^{\text{IN}}$	(9)
Feature Retrieval	$\mathbf{x}_i = \mathbf{M}^{\text{CS}} \mathbf{c}_i$	(10)

TCM posits that when a stimulus $\mathbf{s}_t$ is experienced either in encoding or retrieval, the following sequence of events take place in order: first, presenting $\mathbf{x}_t$ evokes its associated context $\mathbf{c}_t^{\text{IN}}$ via the stimulus-to-context matrix according to Eq. (8). If the stimulus is unique, $\mathbf{c}_t^{\text{IN}}$ is equivalent to the stimulus' pre-experimental context; if the stimulus is repeated, $\mathbf{c}_t^{\text{IN}}$ also contains the (weighted) experimental context where it was previously experienced. Next, the retrieved context updates the current context $\mathbf{c}_t$ according to Eq. (9). Note that ρ and β are chosen so that $\mathbf{c}_t$ remains a unit vector. Finally, $\mathbf{M}^{\text{CS}}$ and $\mathbf{M}^{\text{SC}}$ are updated as needed and the above sequence ensues. If Hebbian learning is assumed, for instance, $\mathbf{M}^{\text{CS}}$ at time t during encoding is updated by the outer product of the recently encoded stimulus $\mathbf{x}_t$ and its temporal context $\mathbf{c}_t$ as shown in Eq. (7).

At the beginning of each new experiment, $\mathbf{M}^{\text{CS}}$ may be reset for simplicity. Howard and Kahana (2002) derived a learning rule for the stimulus-to-context matrix $\mathbf{M}^{\text{SC}}$ such that it behaves in a desirable manner when a stimulus is repeated after a long delay. Since we are interested in sequential decision making scenarios with distinct stimuli, we will not discuss the details in this paper.

The Successor Representation in the Temporal Context Model

Consider the special case where $\mathbf{M}^{\text{SC}}$ is the identity matrix. It follows that $\mathbf{c}_t^{\text{IN}} = \mathbf{x}_t$. i.e. the associated context of a stimulus is exactly its corresponding features. Thus Eq. (9) is reduced to Eq. (1).

Assuming one-hot encoding, we can use the delta function to map each retrieved $\mathbf{x}_t$ to an abstract state vector indexed by time. Thus the j-th entry of $\mathbf{c}_t$ satisfies

$$\mathbf{c}_{t+1}(s_j) = \begin{cases} \rho \mathbf{c}_t & \text{if } S_t \neq s_j \\ \rho \mathbf{c}_t + 1 & \text{if } S_t = s_j, \end{cases}$$

which is analogous to the eligibility trace over $\mathbf{X}$, the set of all possible feature vectors. In particular, Gershman et al. (2012) showed that if stimuli are unique and $\rho = \lambda\gamma\beta$, learning of the context-to-stimulus associations according to Eq. (7) and the transpose of the SR matrix are equivalent under the TD(λ) learning algorithm. Specifically, each TD update can be then written as

$$\mathbf{M}_{t+1}^{\text{CS}} \leftarrow \mathbf{M}_t^{\text{CS}} + \alpha \mathbf{c}_{t+1}(\mathbf{x}'_{t+1} + \gamma \mathbf{x}'_{t+1}\mathbf{M}_t^{\text{CS}} - \mathbf{x}'_t \mathbf{M}_t^{\text{CS}}). \tag{11}$$

If each experience with a certain stimulus is treated as a visitation to a unique state in some abstract memory state space, the context-to-stimulus matrix $\mathbf{M}^{\text{CS}}$ is exactly equivalent to the transpose of the SR. i.e. $\mathbf{M}^{\text{CS}} = \mathbf{M}'$. Additionally, because of the uniqueness, the prediction error is always zero, so Eq. (11) is reduced to $\mathbf{M}_{t+1}^{\text{CS}} \leftarrow \mathbf{M}_t^{\text{CS}} + \alpha \mathbf{c}_{t+1}\mathbf{x}'_{t+1}$; this is exactly Hebbian learning rule for $\mathbf{M}^{\text{CS}}$ in Eq. (7).

On the other hand, if the visited state are not unique, Eq. (7) predicts that context-to-stimulus association will grow without bound, whereas Eq. (11) avoids this issue while maintaining the same functional form in the case of unique stimuli (Gershman et al., 2012).

Value Computation in TCM-SR

Setting $\mathbf{M}^{\text{CS}}$ to the transpose of SR gives rise to a family of sample-based action value computation techniques, which we call TCM-SR. As a special case, consider the problem of estimating the state value of some S_0. Let $\mathbf{m}^{\pi}_{S_0,*;\gamma}$ denote the row in $\mathbf{M}^{\pi}_{\gamma}$ corresponding to S_0 and $m^{\pi}_{S_0,S';\gamma}$ the entry corresponding to a future state S' of the current state S_0 (i.e., expected number of future visits to S' from S_0). Further define $r(S)$ as the one-step expected reward by visiting state S. By expressing values in terms of the

SR and one-step rewards, the state value of S_0 can consequently be rewritten as

$$v_\pi(S_0) = \mathbf{m}^\pi_{S_0,*;\gamma}\mathbf{r} = \sum_{S'} m^\pi_{S_0,S';\gamma} r(S'). \tag{12}$$

Note that each row of $\mathbf{M}^\pi_\gamma$ sums to $1/(1-\gamma)$. Thus we may treat the normalized vector $\frac{1}{1/(1-\gamma)}\mathbf{m}^\pi_{S_0,*;\gamma}$ as a probability distribution over successor states of S_0, which in turn supports standard Monte Carlo sampling techniques to obtain an estimate of $v_\pi(S_0)$ corresponding to a specific discount factor. As a straightforward example, we can draw N i.i.d. successor states (samples) $S_1, S_2, \ldots, S_N$ according to the normalized row $\mathbf{m}^\pi_{S_0,*;\gamma}$. i.e., $S_i \sim (\mathbf{m}^\pi_{S_0,*;\gamma}/\|\mathbf{m}^\pi_{S_0,*;\gamma}\|)$. The Monte Carlo estimator of $\widetilde{v}_\pi(S_0)$ is

$$\widetilde{v}_\pi(S_0) = \frac{1}{N(1-\gamma)} \sum_{i=1}^{N} r(S_i). \tag{13}$$

Setting $\rho = 1, \beta = 0, \mathbf{M}^{\text{CS}} = \mathbf{M}', \mathbf{M}^{\text{SC}} = \mathbf{I}_{|\mathcal{S}|}$ in TCM gives rise to this exact sampling scheme, as the temporal context is never updated with contexts of the sampled states and stays at $\mathbf{m}_{S_0,*;\gamma}$.

However, in general, TCM draws are not i.i.d., because a non-zero β would cause the temporal context to drift towards the most recently experienced stimulus. Subsequent recalls are therefore dependent on preceding memory samples, as manifested by the contiguity effect where subsequent recalls are biased towards successors of the previous sample. In particular, $\mathbf{x}_i$ may be obtained via

$$\mathbf{x}_i \sim \frac{1}{Z}\mathbf{M}^{\text{CS}}\mathbf{c}_i,$$

where Z is the normalization constant and $\mathbf{c}_i = \rho\mathbf{c}_{i-1} + \beta\mathbf{M}^{\text{SC}}\mathbf{x}_{i-1}$.

Importantly, by leveraging the temporal correlation of samples in TCM, value computation can be performed in a flexible manner despite various learning constraints. For example, the discount factor restricts the timescale over which future rewards are considered in the successor representation. The decay of eligibility traces also limits the extent to which reward information is propagated during encoding. Nonetheless, samples drawn, during retrieval, using the drifting temporal context could effectively extend the horizon such that an TCM-SR agent with a small discount factor appears farsighted. When $\gamma = 0$ and $\beta = 1$, TCM-SR produces a standard rollout such that successive samples form a full trajectory, even though the SR at each time step is completely myopic. With a larger γ, the agent could skip multiple steps at a time and compute expected return by searching over an extended temporal scope. With a smaller but non-zero β, the agent interpolates between i.i.d. sampling from the normalized SR (the flattened distribution over successors) and rollouts iteratively over successors' successors.

TCM-SR generates samples analogous to stochastically and recursively constructing a tree over states. At each time step, a state is retrieved from the current temporal context and added to the tree. Because contexts are linear combinations of individual state contexts, suppose S' is drawn from the context of some state S with probability p. An edge between S and a realization $S' = s$ is then added with probability equal to $p(1-\gamma)m^\pi_{S,S';\gamma}$. i.i.d. sampling ($\beta = 0$) results in a random tree with one root node equal to the starting state and all children as leaf nodes (i.e., a star tree). In contrast, the generalized rollout scheme ($\beta = 1$) produces a linear graph - a single chain of state following the starting

state. In expectation, an intermediate β gives rise to an interpolated tree structure of these two types. Simulations 1-3 demonstrate the behavior of each of these cases, and we prove the exact state value computation in the next section.

Furthermore, emotion is known to influence memory. Emotional salience tends to modulate memory retrieval. This effect may be explained by differential rates of stimulus encoding (Talmi et al., 2019) or faster decay of less salient outcomes (Zhou et al., 2020). From the reinforcement learning perspective, both accounts effectively lead to over-representation of particularly rewarding (or detrimental) states, or a utility-weighted memory encoding (Lieder et al., 2018). While enhanced availability of certain samples may bias decisions, when data are sparse and deliberation time is limited, such bias provides a practical advantage to consider rare but critical future possibilities. Noting this link between emotional salience and memory encoding, TCM-SR predicts over-representation of certain events in memory translates to those events having an enhanced impact on decision variables. Similar to Lieder et al. (2018), we simulate emotional modulation with importance sampling, implying a bias-variance trade-off; namely, although over-representation creates a bias in estimation, fewer samples are required for a confident estimate. We give a formal derivation in the next section.

Finally, because SR is dependent on the behavioral policy under which it is learned, a large change in the transition structure or reward function may render the previously obtained SR fruitless. For instance, if a behavioral policy poorly represents certain state transitions around the reward location, an agent using its corresponding SR will be inflexible and perform suboptimally in transfer learning (e.g. Momennejad et al., 2017; Lehnert, Tellex, & Littman, 2017). On the other hand, humans can solve a wide range of transfer learning problems, and perform tasks such as counterfactual reasoning that require simulations of strictly never-seen scenarios. As our main objective is to understand how memory can facilitate effective decision making with limited experience, it is important for the TCM-SR agent to learn values in a flexible manner beyond what the SR prescribes.

Up until now, for simplicity's sake, we have assumed $\mathbf{M}^{\text{SC}}$ to be the identity matrix - that is, the context associated with a state is exactly its feature vector. Alternatively, $\mathbf{M}^{\text{SC}}$ could encode some backward transitions such as the transpose of $\mathbf{M}^{\text{CS}}$, so memory search proceeds in never-experienced directions. Crucially, retrieval of memory samples and subsequent value computation would depend less on the behavioral policy during study. This amounts to regularizing a directional policy to include the possibility of backtracking. We argue that restoring this key feature of the encoding model produces a representation that diverges from the SR, but in so doing corrects one of its key deficiencies.

Theory Details

We now formally prove the relevant properties of the TCM-SR model instantiated as in the Results section. In each of the following cases, the main goal is to prove that the model can be used to compute an unbiased estimate of some queried action a (i.e., $\hat{q}(a)$) in the limit of sample size. For simplicity, we assume that a leads to a deterministic transition to some state S_0. e.g. in the Plinko game, the agent chooses to place the ball in one of the states on the top row of the board. Thus the problem is equivalent to solving $v(S_0)$, or the value of the state corresponding to action a.

In addition, derivations and proofs in this section assume all feature vectors are one-hot coded, and that the starting context is the same as the feature vector associated with the starting state. i.e. $\mathbf{c}_0 = \mathbf{x}_0$. We use $\mathbf{x}(s_n)$ to indicate the location of one at s_n in feature vector $\mathbf{x}$. For clarity, the policy π and discount

factor γ during the encoding phase are implicit in the following proofs. e.g. using $\mathbf{M}$ as a shorthand for $\mathbf{M}_\gamma^\pi$.

Independent samples from memory yield unbiased value estimates

We first consider the case where $\rho = 1, \beta = 0, \mathbf{M}^{\text{CS}} = \mathbf{M}', \mathbf{M}^{\text{SC}} = \mathbf{I}_{|\mathcal{S}|}$, which is the i.i.d. sampling regime.

Lemma 1. *Recall the feature vector associated with the i-th sampled state S_i is $\mathbf{x}_i$. Given $\rho = 1, \beta = 0$, the sampling distribution of S_i is*

$$\mathbb{P}(S_i) = (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_i.$$

Proof. (proof by induction) Base case: $i = 1$. Since each row of $\mathbf{M}$ sums to $1/(1-\gamma)$,

$$\begin{aligned}
\mathbb{P}(S_1) &= \frac{1}{1/(1-\gamma)} \left(\mathbf{M}^{\text{CS}} \left(\rho \mathbf{c}_0 + \beta \mathbf{M}^{\text{SC}} \mathbf{x}_0\right)\right)' \mathbf{x}_1 && \text{Eq. (10)} \\
&= (1-\gamma) \left(\mathbf{M}^{\text{CS}} \mathbf{x}_0\right)' \mathbf{x}_1 \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_1
\end{aligned}$$

Now consider arbitrary time step $i > 1$. By Eq. (9), $\mathbf{c}_i = \mathbf{c}_{i-1} = \cdots = \mathbf{c}_0 = \mathbf{x}_0$. Thus $\mathbb{P}(S_i) = (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_i$. □

Theorem 2. *Given $\rho = 1, \beta = 0$, and N samples $S_1, S_2, \ldots, S_N$, the value of state S_0, $v(S_0)$, satisfies*

$$v(S_0) = \frac{1}{N(1-\gamma)} \mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right].$$

Proof. Denote the feature representation of state $s_k \in \mathcal{S}$ as $\mathbf{x}(s_k)$. Consider the expected reward of the i-th sample:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{r}(S_i)\right] &= \sum_{k=1}^{|\mathcal{S}|} \mathbb{P}(S_i = s_k)\mathbf{r}(s_k) \\
&= \sum_{k=1}^{|\mathcal{S}|} (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}(s_k)\mathbf{r}(s_k) && \text{Lemma 1} \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{M} \sum_{k=1}^{|\mathcal{S}|} \mathbf{x}(s_k)\mathbf{r}(s_k) \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{r} \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{v}.
\end{aligned}$$

By linearity of expectation,

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right] = \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{r}(S_i)\right] = N(1-\gamma)\mathbf{x}_0'\mathbf{v} = N(1-\gamma)v(S_0).$$

Rearranging the terms, we have

$$v(S_0) = \frac{1}{N(1-\gamma)}\mathbb{E}\left[\sum_{i=1}^{N}\mathbf{r}(S_i)\right].$$

□

In summary, in an i.i.d. sampling regime, an action can be evaluated in an unbiased manner by taking the mean across rewards retrieved from episodically sampling the encoded SR.

The contiguity effect suggests value estimation via rollouts

We now consider the case where $\rho = 0, \beta = 1, \mathbf{M}^{\text{CS}} = \mathbf{M}', \mathbf{M}^{\text{SC}} = \mathbf{I}_{|\mathcal{S}|}$, corresponding to the generalized rollout sampling regime.

Lemma 3. *Given $\rho = 0, \beta = 1$, the sampling distribution of the i-th sampled state S_i is*

$$\mathbb{P}(S_i) = (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{x}_i.$$

Proof. (proof by induction) Base case: $i = 1$. This is equivalent to the i.i.d. sampling case. By Lemma 1, the base case holds. Induction hypothesis: for arbitrary $i > 0$, $\mathbb{P}(S_i) = (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{x}_i$.

$$\begin{aligned}
\mathbb{P}(S_{i+1}|S_i) &= \frac{1}{Z}\left(\mathbf{M}^{\text{CS}}\left(\rho\mathbf{c}_i + \beta\mathbf{M}^{\text{SC}}\mathbf{x}_i\right)\right)'\mathbf{x}_{i+1} \\
&= \frac{1}{Z}\left(\mathbf{M}^{\text{CS}}\left(\mathbf{M}^{\text{SC}}\mathbf{x}_i\right)\right)'\mathbf{x}_{i+1} \\
&= \frac{1}{Z}\mathbf{x}_i'\mathbf{M}\mathbf{x}_{i+1},
\end{aligned}$$

where $Z = \mathbf{x}_i'\mathbf{M}\mathbf{1} = 1/(1-\gamma)$ is the normalizing factor. Therefore,

$$\begin{aligned}
\mathbb{P}(S_{i+1}) &= \sum_{s_k}\mathbb{P}(S_i = s_k)\mathbb{P}(S_{i+1}|S_i = s_k) \\
&= \sum_{s_k}(1-\gamma)^i \mathbf{x}_0'\mathbf{M}^i\mathbf{x}(s_k)\cdot(1-\gamma)\mathbf{x}(s_k)'\mathbf{M}\mathbf{x}_{i+1} \\
&= (1-\gamma)^{i+1}\mathbf{x}_0'\mathbf{M}^i\sum_{s_k}\left(\mathbf{x}(s_k)\mathbf{x}(s_k)'\right)\mathbf{M}\mathbf{x}_{i+1} \\
&= (1-\gamma)^{i+1}\mathbf{x}_0'\mathbf{M}^{i+1}\mathbf{x}_{i+1}.
\end{aligned}$$

□

Theorem 4. *Given $\rho = 0, \beta = 1$, and arbitrary encoding γ, the value of S_0 for $\tilde{\gamma} = 1$, $v_{\tilde{\gamma}=1}(S_0)$, satisfies*

$$v_{\tilde{\gamma}=1}(S_0) = \frac{1}{(1-\gamma)}\mathbb{E}\left[\sum_{i=1}^{N}\mathbf{r}(S_i)\right].$$

Proof. Consider the expected reward of the i-th sample:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{r}(S_i)\right] &= \sum_{k=1}^{|\mathcal{S}|} P(S_i = s_k)\mathbf{r}(s_k) \\
&= \sum_{k=1}^{|\mathcal{S}|} (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{x}(s_k)\mathbf{r}(s_k) && \text{Lemma 3} \\
&= (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \sum_{k=1}^{|\mathcal{S}|} \mathbf{x}(s_k)\mathbf{r}(s_k) \\
&= (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{r}.
\end{aligned}$$

By linearity of expectation,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right] &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{r}(S_i)\right] \\
&= \sum_{i=1}^{N} (1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{r} \\
&= (1-\gamma)\mathbf{x}_0' \left(\mathbf{T} + \gamma\mathbf{T}^2 + \gamma^2\mathbf{T}^3 + \dots\right)\mathbf{r} + (1-\gamma)^2 \mathbf{x}_0' \left(\mathbf{T}^2 + 2\gamma\mathbf{T}^3 + 3\gamma^2\mathbf{T}^4 + \dots\right)\mathbf{r} + \dots \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{T}\mathbf{r} + (\gamma(1-\gamma) + (1-\gamma)^2)\mathbf{x}_0'\mathbf{T}^2\mathbf{r} + (\gamma^2(1-\gamma) + 2\gamma(1-\gamma)^2 + (1-\gamma)^3)\mathbf{x}_0'\mathbf{T}^3\mathbf{r} + . \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{T}\mathbf{r} + (1-\gamma)\mathbf{x}_0'\mathbf{T}^2\mathbf{r} + (1-\gamma)\mathbf{x}_0'\mathbf{T}^3\mathbf{r} + \dots \\
&= (1-\gamma)\mathbf{x}_0' \left(\mathbf{T}\mathbf{r} + \mathbf{T}^2\mathbf{r} + \mathbf{T}^3\mathbf{r} + \dots\right) \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{v}_{\gamma=1}.
\end{aligned}$$

Rearranging the terms, we have

$$v_{\tilde{\gamma}=1}(S_0) = \frac{1}{(1-\gamma)}\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right].$$

□

Now consider a fixed probability p_{stop} that interrupts the sampling process of the generalized rollout regime at any moment. i.e., there is a p_{stop} probability that the trial terminates immediately after the current retrieval, regardless whether the trial has reached the end or not (e.g., reaching the bottom row of the Plinko game). The temporal context that guides retrieval is reset following termination. Hence if $p_{\text{stop}} = 1$, the agent always resets the context after sampling one stimulus - equivalent to the i.i.d. sampling regime. If $p_{\text{stop}} = 0$, the agent carries on with the generalized rollout until some pre-defined end state(s) is reached so each trial results in a full trajectory with possible skips over time steps. The latter corresponds to the case proved in Theorem 4.

Proposition 5. *Given $\rho = 0, \beta = 1, p_{stop} \in [0, 1]$, and arbitrary encoding γ, the effective discount factor $\tilde{\gamma}$ of the estimated value satisfies $\tilde{\gamma} = \gamma p_{stop} - p_{stop} + 1$.*

Proof. Consider retrieval at some time i. Let A_i denote the event that the sampling process is yet terminated at time i. Thus by the above definition of p_{stop}, $\mathbb{P}(A_i) = (1 - p_{\text{stop}})^{i-1}$ for all $i \geq 1$. Further assume that upon termination, all remaining samples have reward zero (even though technically no more samples are drawn). By Theorem 4, we know

$$\mathbb{E}\left[\mathbf{r}(S_i)\right] = \mathbb{P}(A_i)(1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{r} + \mathbb{P}(A_i^{\mathsf{c}}) \cdot 0 = (1 - p_{\text{stop}})^{i-1}(1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{r}.$$

By linearity of expectation,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right] &= \sum_{i=1}^{N} \mathbb{E}\left[\mathbf{r}(S_i)\right] \\
&= \sum_{i=1}^{N} (1 - p_{\text{stop}})^{i-1}(1-\gamma)^i \mathbf{x}_0' \mathbf{M}^i \mathbf{r} \\
&= (1-\gamma)\mathbf{x}_0' \left(\mathbf{T} + \gamma \mathbf{T}^2 + \gamma^2 \mathbf{T}^3 + \dots\right) \mathbf{r} + (1 - p_{\text{stop}})(1-\gamma)^2 \mathbf{x}_0' \left(\mathbf{T}^2 + 2\gamma \mathbf{T}^3 + 3\gamma^2 \mathbf{T}^4 + \dots\right) \mathbf{r} + \\
&= (1-\gamma)\mathbf{x}_0' \mathbf{T}\mathbf{r} + (\gamma(1-\gamma) + (1 - p_{\text{stop}})(1-\gamma)^2) \mathbf{x}_0' \mathbf{T}^2 \mathbf{r} \\
&\quad + (\gamma^2(1-\gamma) + 2(1 - p_{\text{stop}})\gamma(1-\gamma)^2 + (1 - p_{\text{stop}})^2(1-\gamma)^3) \mathbf{x}_0' \mathbf{T}^3 \mathbf{r} + \dots \\
&= (1-\gamma)\mathbf{x}_0' \mathbf{T}\mathbf{r} + (1-\gamma)(\gamma p_{\text{stop}} - p_{\text{stop}} + 1) \mathbf{x}_0' \mathbf{T}^2 \mathbf{r} + (1-\gamma)(\gamma p_{\text{stop}} - p_{\text{stop}} + 1)^2 \mathbf{x}_0' \mathbf{T}^3 \mathbf{r} + \dots \\
&= (1-\gamma)\mathbf{x}_0' \left(\mathbf{T}\mathbf{r} + (\gamma p_{\text{stop}} - p_{\text{stop}} + 1) \mathbf{T}^2 \mathbf{r} + (\gamma p_{\text{stop}} - p_{\text{stop}} + 1)^2 \mathbf{T}^3 \mathbf{r} + \dots\right).
\end{aligned}$$

Interpreting $\gamma p_{\text{stop}} - p_{\text{stop}} + 1$ as the discount factor, we get

$$\mathbb{E}\left[\sum_{i=1}^{N} \mathbf{r}(S_i)\right] = (1-\gamma)\mathbf{x}_0' \mathbf{v}_{\tilde{\gamma} = \gamma p_{\text{stop}} - p_{\text{stop}} + 1}.$$

Therefore, in effect, the additional interruption probability permits modification of the temporal horizon during retrieval (and consequently, evaluation) beyond the intrinsic encoding discount factor γ. In particular, assuming the agent has control over this interruption probability, by varying p_{stop} between 0 and 1, it can interpolate $\tilde{\gamma}$ between the encoding γ and 1. Note $\tilde{\gamma} = 1$ corresponds to the rollout sampling regime proven by Theorem 4. □

In summary, in a generalized rollout sampling regime, an action can be evaluated in an unbiased manner by adding up rewards retrieved from episodically sampling the encoded SR. Specifically, the estimated action value corresponds to a discount factor of 1, or an undiscounted estimate. This implication may be problematic for tasks with an infinite horizon, as termination is undefined and the sum of rewards may be infinite. Thus we introduce an additional interruption probability p_{stop} at any given moment during retrieval/evaluation, which the agent is assumed to have control over. The result is an effective discount factor $\tilde{\gamma}$ that can be flexibly interpolated between the encoding discount factor γ and 1. For clarity, in the main text, we refer to the effective discount factor $\tilde{\gamma}$ whenever applicable, making p_{stop} implicit in our arguments.

Data from free recall experiments suggests an intermediate regime

Observe that the sequentially obtained samples can be conceptualized as a random tree with root at S_0. At each retrieval step i where $i > 0$, a node S_i is inserted into the existing tree T_{i-1} such that an edge is drawn between the current node S_i and some existing node S_j ($i > j \geq 0$) if S_i is drawn from the SR-defined distribution at S_j. Because each context is a linear combination of successor distributions of experienced stimuli, in theory, we can identify a sample as the successor of some previously retrieved state given the context it is drawn from. Let $pa(i) = j$ denote the event that S_j is the parent of S_i. For instance, $\mathbb{P}(pa(1) = 0) = 1$ since S_1 is always drawn from the distribution $(1-\gamma)\mathbf{x}_0'\mathbf{M}$ regardless of the value of ρ and β. $\mathbb{P}(pa(2) = 0) = \rho$ and $\mathbb{P}(pa(2) = 1) = \beta$ according to Eq. (9). In general, for any $i > j \geq 0$ we have

$$\mathbb{P}(pa(i) = j) = \begin{cases} \rho^{i-1} & \text{if } j = 0 \\ \rho^{i-j-1}\beta & \text{if } j > 0, \end{cases}$$

Note that the construction necessarily results in a tree because of the sequential nature of the sampling process, namely a newly inserted node has an index strictly larger than that of any existing node. The resultant tree with all N nodes plus the root node is T_N. Observe that if $\rho + \beta = 1$, then $\forall j. \sum_{i=0}^{j-1} \mathbb{P}(pa(i) = j) = 1$, so the distribution is a proper probability distribution.

Lemma 6. *Assume $\rho + \beta = 1$, $\rho, \beta > 0$. As $N \to \infty$, T_N is expected to be a tree with $1/(1-\rho)$ degrees at the root and linear graphs thereafter.*

Proof. Consider $d_N(i)$, the number of children nodes S_i has in tree T_N. It suffices to show that

$$\lim_{N\to\infty} \mathbb{E}[d_N(i)] = \begin{cases} 1/(1-\rho) & \text{if } i = 0 \\ 1 & \text{if } i > 0. \end{cases}$$

For arbitrary $N \in \mathbb{N}$, $\mathbb{E}[d_N(0)] = \sum_{i=1}^{N} \mathbb{P}(pa(i) = 0) = \sum_{i=1}^{N} \rho^{i-1} = \frac{1-\rho^N}{1-\rho}$, and $\forall j > 0$. $\mathbb{E}[d_N(j)] = \sum_{i=j+1}^{N} \mathbb{P}(pa(i) = j) = \sum_{i=j+1}^{N} \rho^{i-j-1}\beta = \frac{\beta(1-\rho^{N-j})}{1-\rho} = 1-\rho^{N-j}$. Thus, $\lim_{N\to\infty} \mathbb{E}[d_N(0)] = 1/(1-\rho)$, $\lim_{N\to\infty} \mathbb{E}[d_N(j)] = 1$ for all positive j. □

Corollary 7. *Given $\rho + \beta = 1$, $\rho, \beta > 0$, if N is large but finite, T_N is expected to have $(1-\rho^N)/(1-\rho)$ children, while the number of children of early samples are subcritical.*

Proof. The proof follows directly from Lemma 6 with finite N, noting that when j is small, $N - j$ is close to N so $\mathbb{E}[d_N(0)] \approx 1 - \rho^N < 1$. □

Theorem 8. *Given $\rho + \beta = 1$, $\rho, \beta > 0$,*

$$v_{\gamma=1}(S_0) = \frac{\beta}{(1-\gamma)} \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbf{r}(S_i)\right].$$

Proof. Here we provide a sketch of the formal proof: note that the extreme cases where one of ρ, β is 1 can be realized as a random recursive tree describe above. Specifically, as $N \to \infty$, $\rho = 1$ corresponds to a tree with height 1 and infinitely many branches at the root; $\rho = 0$ corresponds to a path graph with

infinite height. Importantly, given such a tree, we know $v_{\gamma=1}(S_0)$ may be computed as the expected total return of an arbitrary path from the root node to a leaf node in a random tree T_∞. i.e., sum along paths and average across paths from Theorem 2 and 4 respectively. The result then directly follows from Lemma 6 noting $1-\rho=\beta$. □

Lemma 9. *Given $\rho+\beta=1$, $\rho,\beta>0$, there is a non-zero probability that the shortest path from the root node to a leaf has length 2.*

Proof. Without loss of generality, consider the event l_2 that no vertex is attached to node 2 (equivalently, no future sample is drawn from the successor distribution of S_2). Then

$$\begin{aligned}
&\mathbb{P}(l_2)=\prod_{i=3}^{N}(1-\mathbb{P}(pa(i)=2))\\
\implies &\log\mathbb{P}(l_2)=\sum_{i=3}^{N}\log(1-\rho^{i-3}\beta)=\sum_{i=0}^{N-3}\log(1-\rho^{i}\beta)\approx-\sum_{i=0}^{N-3}\rho^{i}\beta<\infty\\
\implies &\mathbb{P}(l_2)>0
\end{aligned}$$

□

Proposition 10. *Given $\rho+\beta=1$, $\rho,\beta>0$, $N<\infty$, the value estimator in Theorem 8 is biased.*

Proof. Lemma 9 implies any random tree resulted from the constructive process likely has a short path (a "stub"), thus applying Theorem 8 tends to underestimate $v(S_0)$. □

Therefore, in the intermediate sampling regime that interpolates between the i.i.d. and generalized rollout regimes, an action can be evaluated in an unbiased manner by first adding up the rewards retrieved from episodically sampling the encoded SR and then scaling the sum by β, which acts like the branching factor in the limit of sample size. We have explicitly shown that such estimator may be biased downwards in the case of relatively small number of samples, but like previous cases, given a sufficiently large number of samples, the estimate approaches the true value.

Emotional Modulation of memory yields bias-variance trade-off

We implement emotional modulated learning similar to Talmi et al. (2019) by employing a fixed learning rate that is higher for emotionally salient than non-salient stimuli to learn $\mathbf{M}^{\text{CS}}$. For clarity, a state s either contains nothing (i.e. $\mathcal{R}(s)=0$) or a small reward ($\mathcal{R}(s)=1$). All else being equal, the resultant, emotionally modulated TC-SR agent is thus more likely to obtain a rewarding sample than an unmodulated agent. Denote the unbiased context-to-stimulus associative matrix $\mathbf{M}'$ and the biased $\overline{\mathbf{M}}'\neq\mathbf{M}'$. To compute an estimation of expectation, it needs *importance sampling* to translate distributions of $\mathbf{M}$ to $\overline{\mathbf{M}}$.

For simplicity, consider $\rho=1,\beta=0$ (i.i.d. sampling). By Lemma 1, the unbiased sampling distribution of the i-th sample S_i is $P(S_i)=(1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_i$, while the biased sampling distribution of S_i is $Q(S_i)=$

$(1-\gamma)\mathbf{x}_0'\overline{\mathbf{M}}\mathbf{x}_i$. To correct for the difference between P and Q, each sample S_i is reweighed by

$$w_{S_i} = \frac{P(S_i)}{Q(S_i)} = \frac{m_{S_0,S_i}}{\overline{m}_{S_0,S_i}}.$$

Using $\overline{\mathbf{M}}'$, the expected total reward for the i-th sample may be estimated as

$$\begin{aligned}
\mathbb{E}\left[\mathbf{r}(S_i)\right] &= \sum_{k=1}^{|\mathcal{S}|} P(S_i = s_k)\mathbf{r}(s_k) \\
&= \sum_{k=1}^{|\mathcal{S}|} Q(S_i = s_k)\frac{P(S_i = s_k)}{Q(S_i = s_k)}\mathbf{r}(s_k) \\
&= \sum_{k=1}^{|\mathcal{S}|} w_{S_i} Q(S_i = s_k)\mathbf{r}(s_k).
\end{aligned}$$

Theorem 2 can be then applied to estimate a specific state value. In general, $\widetilde{\mathbf{v}}$ is biased if N is finite. Specifically, $\widetilde{\mathbf{v}}$ demonstrates a bias-variance trade-off, such that extreme events are over-represented in the samples due to the biased associative matrix, but value estimates also tend to be less varied.

Similarly, if $\rho = 0, \beta = 1$ (generalized rollout), by Lemma 3, the unbiased distribution of the i-th sampled state S_i is $P(S_i) = (1-\gamma)^i\mathbf{x}_0'\mathbf{M}^i\mathbf{x}_i$, while the biased sampling distribbution of S_i is $Q(S_i) = (1-\gamma)^i\mathbf{x}_0'\overline{\mathbf{M}}^i\mathbf{x}_i$. Denote the (S_0, S_i)-th entry of $\mathbf{M}^i$ as $\left(\mathbf{M}^i\right)_{S_0,S_i}$ and that of $\overline{\mathbf{M}}^i$ as $\left(\overline{\mathbf{M}}^i\right)_{S_0,S_i}$. To correct for the difference between P and Q, each sample S_i should be reweighed by

$$w_{S_i} = \frac{P(S_i)}{Q(S_i)} = \frac{\left(\mathbf{M}^i\right)_{S_0,S_i}}{\left(\overline{\mathbf{M}}^i\right)_{S_0,S_i}}.$$

The expected total reward proceeds similarly as stated in Theorem 4 with reweighing. For demonstration purposes, we use the i.i.d. regime to illustrate the effect of emotional modulation in simulations.

Simulation Details

All simulations used a Plinko game of size 10x9 (i.e. $H = 10$, $|\mathcal{S}| = 90$, excluding the absorbing state which is outside the main board). Binary rewards were randomly placed in locations between row 1 and row 6 (inclusive; top row is row 0) such that all of them were reachable from the starting state. Each experiment was characterized by its reward placement. Details of each simulation are specified below.

Details of Simulation 1: Independent samples from memory yield unbiased value estimates

We set $\rho = 1, \beta = 0$ to simulate the effect of a stationary context, which gave rise to independent draws of memory samples in TCM-SR. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 2a-c) and $\gamma = 0.5$ (Fig. 2d-f) during encoding, with the latter corresponding to a slower rate of temporal drift (i.e., longer timescale). The stimulus-to-context associative matrix $\mathbf{M}^{\text{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments (games) were conducted for each different discount factor, with 50 trials per experiment and 1000 (independent) samples per trial (i.e., $N = 1000$). At least one reward was placed within the agent's temporal horizon. e.g., given $\gamma = 0$, row 2 contained at least one reward. The sampling distributions over rows (Fig. 2b,e) reflect trial averages if starting from the top-center state (marked with a orange circle in Fig. 2a,d).

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was computed as the average across samples and trials. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 2c,f). The number of rewards were chosen to reflect a spectrum of reward abundance ranging from a single reward to about 50%. The percentage of maximum rewards obtained of a particular game pmr was computed as

$$pmr = \frac{v(S_{\text{chosen}})}{v(S^*)},$$

where S_{chosen} is the state selected by the deterministic policy, S^* is the state with higher expected total return, and $v(\cdot) : \mathcal{S} \mapsto \mathbb{R}$ is the state value function. Note an optimal choice implies $pmr = 1$. Fig. 2c,f show the average pmr across 100 experiments.

Details of Simulation 2: Recall-dependent context updates lead to rollouts

We set $\rho = 0, \beta = 1$ to simulate the effect of a context fully determined by the most recent retrieval, which gave rise to generalized rollouts in TCM-SR. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 3a-d) and $\gamma = 0.5$ (Fig. 3e-h) during encoding. For each γ, simulation were repeated using three different probabilities of interruption $p = 0.05, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s for each underlying true γ at retrieval (Fig. 3b,f). Thus as long as the ball had not reached the bottom of the Plinko board, at each time step, there was a p probability that the trial will terminate, regardless of the ball's location. Consequently, each trial started from the top-center state (marked with a orange circle in Fig. 3a,e) and ended if either the ball hit the bottom of the board or the sampling process terminated due to the non-zero interruption probability. The stimulus-to-context associative matrix $\mathbf{M}^{\text{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments were conducted for each combination of discount factor and interruption probability. The sampling distributions over rows (Fig. 3b,f) reflect averages across 1000 trials per experiment if starting from the top-center state. The implied contiguity curves (Fig. 3d,h) were computed similarly using the same starting state.

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was obtained by summing samples within each of 5000 trials and averaging across trials. 100 games were simulated and each trial consists of a variable number of correlated samples (at most nine, or $H-1$). The interruption probability is fixed at 0.05. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 3c,g). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 3c,g show the average pmr across 100 experiments.

Details of Simulation 3: An intermediate regime between i.i.d. sampling and rollouts

We set $\rho = \beta = 0.5$ to simulate the effect of an intermediate context updating regime in TCM-SR that better explains human behavioral data on free recall tasks. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 4a-c) and $\gamma = 0.5$ (Fig. 4d-f) during encoding. For each γ, simulations were repeated using three different probabilities of interruption $p = 0.05, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s for each underlying true γ at retrieval (Fig. 4b,e). Thus as long as the ball had not reached the bottom of the Plinko board, at each time step, there was a p probability that the trial will terminate, regardless of the ball's location. Consequently, each trial started from the top-center state (marked with a orange circle in Fig. 4a,d) and ended when the ball hit the bottom of the board or the sampling process terminated due to the non-zero interruption probability. The stimulus-to-context associative matrix $\mathbf{M}^{\text{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments were conducted for each combination of discount factor and interruption probability. The sampling distributions over rows (Fig. 4b,e) reflect averages across 100 trials per experiment if starting from the top-center state.

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was obtained by summing samples within each of 5000 trials and averaging across trials. 100 games were simulated and each trial consists of a variable number of correlated samples. The interruption probability is fixed at 0.05. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 4c,f). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 4c,f show the average pmr across 100 experiments.

Details of Simulation 4: Retrieval with limited experience and with emotional modulation

We chose the i.i.d. sampling regime (i.e., $\rho = 1, \beta = 0$) to illustrate the effect of limited experiences and emotional modulation. The stimulus-to-context associative matrix $\mathbf{M}^{\text{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

The intermediate and converged SR matrices of the top-center state (4 panels to the left in Fig. 5a,b) were learned via TD(λ), where $\lambda = 0.7, \gamma = 0.9$. A ball was dropped four times from the top-center position of a board with predetermined reward locations and reached the bottom following a sequence of transitions, resulting in 4 complete trajectories. An intermediate SR was computed after observation of each complete trajectory. The unmodulated and modulated learning rates were initialized to 0.01 and 0.5 respectively. i.e., $\alpha_0 = 0.01, \alpha_{mod,0} = 0.5$. Both the unmodulated agent (Fig. 5a) and the modulated agent (Fig. 5b) were trained using the same exponential decay schedule such that the learning rates upon observing trajectory t was defined as

$$\alpha_t = \alpha_0 * e^{-kt}$$
$$\alpha_{mod,t} = \alpha_{mod,0} * e^{-kt},$$

where decay rate $k = 0.001$. In both cases, the SR converged after observing 10000 trajectories.

We used 100 random experiments (games) and drew 1000 samples from the TD-learned SR after one observation (trajectory) in each experiment to compute the average fraction of samples that contained a reward (Fig. 5d). The same set of samples (i.e., after observing a single trajectory) were used to compute the bias and variance in the value estimate of the top-center state, with a random number of binary rewards between 20 (inclusive) and 40 (exclusive) placed on the board (Fig. 5e,f).

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was computed as the average across 1000 i.i.d. samples and 50 trials. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 5c). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 5c shows the average pmr across 100 experiments.

Details of Simulation 5: Retrieving a learned context allows backward sampling

We chose the generalized rollout regime (i.e., $\rho = 0, \beta = 1$) to illustrate the effect of retrieving a learned context associated with a stimulus as opposed to a task-independent feature representation. The stimulus-to-context associative matrix $\mathbf{M}^{\mathrm{SC}}$ was equal to the SR matrix $\mathbf{M}$. Simulations used $\gamma = 0.5$ during encoding and three different interruption probabilities $p = 0.2, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s at retrieval (Fig. 6c,d). Each simulation consisted of 500 experiments and 1000 trials (rollouts) per experiment from the top-center state.

The true state value of the top-center state was computed by assuming full reversibility (i.e., symmetry of conditional transition probabilities), while the estimates are computed similar to Simulation 2 (i.e., as generalized rollouts; Fig. 6c,d).

The simulation code will be made available at `https://github.com/corxyz/tcm-sr` upon publication.

References

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., … Shohamy, D. (2019, jul). The hippocampus supports deliberation during value-based decisions. *eLife*, *8*, e46080. doi: 10.7554/eLife.46080

Bornstein, A. M., & Daw, N. D. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLoS Computational Biology*, *9*.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. (2017, 06). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*, 15958. doi: 10.1038/ncomms15958

Brea, J., Gaál, A. T., Urbanczik, R., & Senn, W. (2016). Prospective coding by spiking neurons. *PLoS Computational Biology*, *12*.

Cohen, R. T., & Kahana, M. J. (2019). Retrieved-context theory of memory in emotional disorders. *bioRxiv*, 817486.

Coulom, R. (2006). Efficient selectivity and backup operators in monte-carlo tree search. In *Computers and games*.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704-1711.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613-624. doi: 10.1162/neco.1993.5.4.613

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312 - 325.

Doll, B. B., Shohamy, D., & Daw, N. D. (2015). Multiple memory systems as substrates for multiple decision systems. *Neurobiology of Learning and Memory*, *117*, 4-13.

Eichenbaum, H. (2001). The hippocampus and declarative memory: cognitive mechanisms and neural codes. *Behavioural Brain Research*, *127*, 199-207.

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*.

Gershman, S. J. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience*, *38*, 7193 - 7200.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The successor representation and temporal context. *Neural Computation*, *24*, 1553-1568.

Greene, R. L. (1986). A common basis for recency effects in immediate and delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 413-418. doi: 10.1037/0278-7393.12.3.413

Gupta, R., Duff, M. C., Denburg, N. L., Cohen, N. J., Bechara, A., & Tranel, D. (2009). Declarative memory is critical for sustained advantageous complex decision-making. *Neuropsychologia*, *47*, 1686-1693.

Gutbrod, K., Krouel, C., Hofer, H., Müri, R. M., Perrig, W. J., & Ptak, R. (2006). Decision-making in amnesia: Do advantageous decisions require conscious knowledge of previous behavioural choices? *Neuropsychologia*, *44*, 1315-1324.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of experimental psychology. Learning, memory, and cognition*, *25 4*, 923-41.

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

Howard, M. W., Youker, T. E., & Venkatadass, V. S. (2008). The persistence of memory: Contiguity effects across hundreds of seconds. *Psychonomic Bulletin & Review*, *15*, 58-63.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103-109.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5), e1002055.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, *20*(7), 512–534.

Lehnert, L., Tellex, S., & Littman, M. L. (2017). Advantages and limitations of using successor features for transfer in reinforcement learning. *ArXiv*, *abs/1708.00102*.

Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: The third way. In *Nips*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*, 1–32.

Liu, Y., Mattar, M. G., Behrens, T. E., Daw, N. D., & Dolan, R. J. (2021). Experience replay is associated with efficient nonlocal learning. *Science*, *372*(6544), eabf1357.

Mather, M., Clewett, D. V., Sakaki, M., & Harley, C. W. (2015). Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behavioral and Brain Sciences*, *39*.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, *21*(11), 1609–1617.

Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*.

Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature neuroscience*, *20*, 1269 - 1276.

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*.

Momennejad, I., Russek, E., Cheong, J., Botvinick, M., Daw, N., & Gershman, S. (2017). The successor representation in human reinforcement learning. *Nature Human Behavior*, *1*(9), 680-692. doi: 10.1038/s41562-017-0180-8

Nicholas, J., Daw, N. D., & Shohamy, D. (2022). Uncertainty alters the balance between incremental learning and episodic memory. *Elife*, *11*, e81679.

O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive science*, *38*(6), 1229–1248.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place cell sequences depict future paths to remembered goals. *Nature*, *497*, 74 - 79.

Piray, P., & Daw, N. D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications*, *12*(1), 4942.

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological review*, *122 4*, 621-47.

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.

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., & Botvinick, M. (2018). Been there, done that: Meta-learning with episodic recall. In *International conference on machine learning* (pp. 4354–4363).

Russek, E., Acosta-Kane, D., van Opheusden, B., Mattar, M. G., & Griffiths, T. (2022). Time spent thinking in online chess reflects the value of computation.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2021). Neural evidence for the successor representation in choice evaluation. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2021/08/31/2021.08.29.458114 doi: 10.1101/2021.08.29.458114

Schacter, D. L., Benoit, R. G., Brigard, F. D., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, *117*, 14-21.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological review*, *115*(4), 893.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643–1653.

Stefanidi, A., Ellis, D. M., & Brewer, G. A. (2018). Free recall dynamics in value-directed remembering. *Journal of Memory and Language*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9-44.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Talmi, D., Kavaliauskaite, D., & Daw, N. D. (2018). In for a penny, in for a pound: examining motivated memory through the lens of retrieved context models. *Learning & Memory*, *28*, 445 - 456.

Talmi, D., Lohnas, L. J., & Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological review*, *126 4*, 455-485.

Tesauro, G., & Galperin, G. (1996). On-line policy improvement using monte-carlo search. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems* (Vol. 9). MIT Press.

Tulving, E. (1972). Organization of memory. In E. Tulving & W. Donaldson (Eds.), (p. 381-402). Academic Press.

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Ma, W. J., et al. (2021). Revealing the impact of expertise on human planning with a two-player board game.

Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., ... Daw, N. D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron*, *102*(3), 683–693.

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.

Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: an emotional binding account. *Trends in Cognitive Sciences*, *19*, 259-267.

Zhou, C. Y., Guo, D., & Yu, A. J. (2020). Devaluation of unchosen options: A bayesian account of the provenance and maintenance of overly optimistic expectations. In *Cogsci 2020* (Vol. 42, p. 1682-1688).