

Episodic Retrieval for Model-Based Evaluation in Sequential Decision Tasks

Corey Y. Zhou¹, Deborah Talmi², Nathaniel D. Daw³, and Marcelo G. Mattar^{1, 4}

¹ Department of Cognitive Science, University of California, San Diego

² Department of Psychology, University of Cambridge

³ Princeton Neuroscience Institute, Princeton University

⁴ Department of Psychology, New York University

It has long been hypothesized that episodic memory supports adaptive decision making by enabling mental simulation of future events. Yet, attempts to characterize this process are surprisingly rare. On one hand, memory research is often carried out in settings that are far removed from ecological contexts of decision making. On the other hand, models of adaptive choice only invoke episodic memory in highly stylized terms, if at all. To address these gaps, we propose TCM-SR, a novel process-level model that grounds model-based evaluation in empirically informed dynamics of episodic recall. In this model, the probability of retrieving each available memory is governed by the successor representation, 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, a prominent model of episodic retrieval. Through simulations and analytical derivations, we show that the patterns of episodic retrieval suggested by this model enables flexible computation of decision variables. On this basis, a number of previously described features of episodic memory might serve an adaptive purpose in sequential decision making. For instance, we show that the contiguity effect, a well-known bias in episodic retrieval, enables mental simulation via model-based rollouts to inform decisions. We also show that backward retrieval and emotional modulation improve generalization and the efficiency of decisions given limited experience. By bridging computational models across these two domains, we make several theoretical and empirical predictions linking episodic memory to adaptive choice in sequential tasks.

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

What is memory for? Although laboratory studies often focus on memory performance in isolation, as if recall was 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, a topic now prominent in the fields of psychology, neuroscience, and machine learning. Understanding this interplay promises both to unravel the

mechanisms by which experience influences 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

This article was published Online First December 30, 2024.

Andrew Heathcote served as action editor.

Marcelo G. Mattar  <https://orcid.org/0000-0003-3303-2490>

Nathaniel D. Daw and Marcelo G. Mattar contributed equally to this article. The simulation code is available on GitHub at <https://github.com/corxyz/tcm-sr>. This study was not preregistered. An earlier version of this work was presented at the 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making, titled “Memory Mechanisms Predict Sampling Biases in Sequential Decision Tasks” by Marcelo G. Mattar, Deborah Talmi, and Nathaniel D. Daw (Mattar et al., 2019).

This work was supported in part by the Division of Information and Intelligent Systems, U.S. National Science Foundation, Grant IIS-1822571 awarded to Nathaniel D. Daw, which is a part of the Collaborative Research in Computational Neuroscience program. The authors thank Kim Stachenfeld, Tianyi Zheng, and Jason Schweinsberg for helpful discussions.

This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License (CC BY-NC-ND 4.0;

<https://creativecommons.org/licenses/by-nc-nd/4.0>). This license permits copying and redistributing the work in any medium or format for noncommercial use provided the original authors and source are credited and a link to the license is included in attribution. No derivative works are permitted under this license.

Corey Y. Zhou played a lead role in formal analysis, investigation, software, and visualization and an equal role in writing—original draft and writing—review and editing. Deborah Talmi played a supporting role in investigation, writing—original draft, and writing—review and editing. Nathaniel D. Daw played a supporting role in writing—original draft and writing—review and editing and an equal role in conceptualization, investigation, and project administration. Marcelo G. Mattar played a lead role in supervision, a supporting role in software and visualization, and an equal role in conceptualization, investigation, project administration, writing—original draft, and writing—review and editing.

Correspondence concerning this article should be addressed to Marcelo G. Mattar, Department of Psychology, New York University, New York, NY 10012, United States. Email: marcelo.mattar@nyu.edu

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 support the deliberative evaluation of candidate actions in goal-directed behavior (Doll et al., 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 modellikelike 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 article, we propose a new mechanistic theory of decision making that grounds model-based evaluation in the recall of episodic memories or memories for individual episodes from one’s past (Tulving, 1972). Many decisions could benefit from recall of one-off events. For example, to navigate through a large unfamiliar venue, we may recall having examined the sculpture in the left corridor earlier that night and use that memory to orient ourselves. Consistent with a role on decision making, episodic memory is strongly modulated by the presence of rewards and other emotionally salient information (Clewett et al., 2019; Horwath et al., 2023; Mather et al., 2015; Talmi et al., 2019). Additionally, episodic memory has been suggested to guide decisions by scaffolding the construction of hypothetical future scenarios (Schacter et al., 2015). In line with these predictions, patients with episodic memory deficits show impairments in a range of decision-making tasks (Bakkour et al., 2019; Gupta et al., 2009; Gutbrod et al., 2006).

Thus far, most of the work characterizing the role of episodic memory on model-based evaluation starts from a decision-making perspective. For example, there has been growing interest 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 past events with similar actions and their outcomes (Bornstein et al., 2017; Bornstein & Norman, 2017; Lieder et al., 2018; Plonsky et al., 2015). Yet, despite the suggestive links, these previous connections still lack in their explanatory power: They are often limited to single-step bandit tasks (Bornstein et al., 2017; Bornstein & Norman, 2017; Duncan & Shohamy, 2016; Rouhani et al., 2018) and conceive episodic memory’s role in decisions as a veridical record of one’s past events (Braun et al., 2018; Duncan & Shohamy, 2016; Nicholas et al., 2022). The same is true of algorithms that augment reinforcement learning (RL) agents with the ability to store and retrieve past experiences (Blundell et al., 2016; Gershman & Daw, 2017; Lengyel & Dayan, 2007; Pritzel et al., 2017; Ritter et al., 2018). While these algorithms show clear improvements in performance in sequential decision settings, they also abstract away the most intriguing aspects of episodic memory. Here, we aim to address these limitations.

In contrast to the predominant approach in decision making, our approach instead begins with a standard model of episodic encoding and recall—the temporal context model (TCM; Howard & Kahana, 2002a). TCM is a descriptive (rather than normative) model that reproduces various patterns of episodic retrieval in tasks like list–list learning. In the past 2 decades, TCM has been augmented with

additional assumptions and mechanisms to reproduce an increasingly larger number of episodic memory phenomena (Cohen & Kahana, 2022; Healey & Kahana, 2016; Polyn et al., 2009a; Sederberg et al., 2008; Talmi et al., 2019). Common to all these models is the existence of a slowly changing context representation to which list items are linked, enabling subsequent retrieval. Building off of this rigorous work, we examine the implications of these assumptions to sequential decision making. Through a series of simulations and analytical derivations, 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 associations formed during *encoding* in TCM correspond to the successor representation (SR), a type of world model that supports efficient and flexible decision making (Gershman et al., 2012). We extend this 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. In other words, we show that TCM offers a mechanism for decision making based on the SR. The result is a theoretical proposal that we call temporal context model–successor representation (TCM-SR).

The retrieval mechanism in TCM-SR, like the original TCM, reproduces fundamental properties of episodic memory, such as the tendency to retrieve items in the same temporal order in which they were experienced. And despite its root in memory research, TCM-SR offers a quantitative mapping to RL models in decision neuroscience, thereby expanding the connection between the two fields. We show that two special settings of the retrieval mechanism in 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, providing a family of Monte Carlo estimators based on a generalized notion of rollouts.

Equipped with a model that formalizes the link between retrieval and model-based evaluation, we proceed to 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. Crucially, our model also makes several empirical predictions about decision making, including 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. More broadly, we hope that the mapping we offer between research in episodic memory and decision making sheds light on both areas and suggests many new research directions and future experiments.

Results

Decisions via Model-Based Evaluation

To illustrate the role of episodic retrieval on action evaluation, we consider a stylized decision-making task inspired by the game of

Plinko. In Plinko, a ball is dropped from the top of a board and bounces off pegs, gaining points as it descends. The player’s goal is to drop the ball at the location that will earn them as many points as possible. The spot where the ball is initially placed represents the player’s action, and each subsequent location visited by the ball represents a state. Each bounce and the resulting direction and points mimic the unpredictable outcomes following the player’s initial decision. Like the ball’s trajectory, the process involves random transitions from one state to the next, accumulating rewards along the way. While Plinko is obviously a real-life game, in this article, we use it as a metaphor for a generic decision-making task where rewards are gathered over time and the outcome of each action is uncertain. In this class of tasks, optimal decision 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 the problem of prediction using the framework of RL (Sutton & Barto, 2018). On each trial, the agent chooses an action $A = a$ and receives the return $G = \sum_{t=1}^{\infty} \gamma^{t-1} R_t$, where R_t is the reward received at time step 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 return G . This requires estimating, for each candidate action a , the expectation $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t | A = a]$, known as the *action value*. If action values are known, the agent can select the action with maximum value. Note that, in the class of tasks we study here, the agent takes no further action after the first one. This setting represents either a one-step (i.e., bandit) task with temporally extended outcomes or a sequential decision problem where future decisions are not optimized. In RL, the latter case corresponds to policy evaluation in a Markov process, a classic subproblem for solving more elaborate choice tasks (e.g., Markov decision processes [MDP] 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 it is not the focus of this article.

Here, we focus on the second class of strategies, often called *planning* or *model-based* RL. Suppose that 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—that is, the agent understands the step-by-step transition structure of the game, a form of world model (Figure 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 any of the 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. Averaging the total reward across multiple rollouts yields an estimate of $q(a)$, and by repeating this process for each candidate action—a temporally extended process—the agent can choose the action with maximal estimated value. Note that this type of action evaluation by stochastic, iterative simulation is at the heart of numerous model-based approaches to RL, such as the 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 entirely novel situations, such as a never-experienced board position (Daw & Dayan, 2014; Mattar & Lengyel, 2022). However, such flexibility may incur a high cost in terms of compute and time.

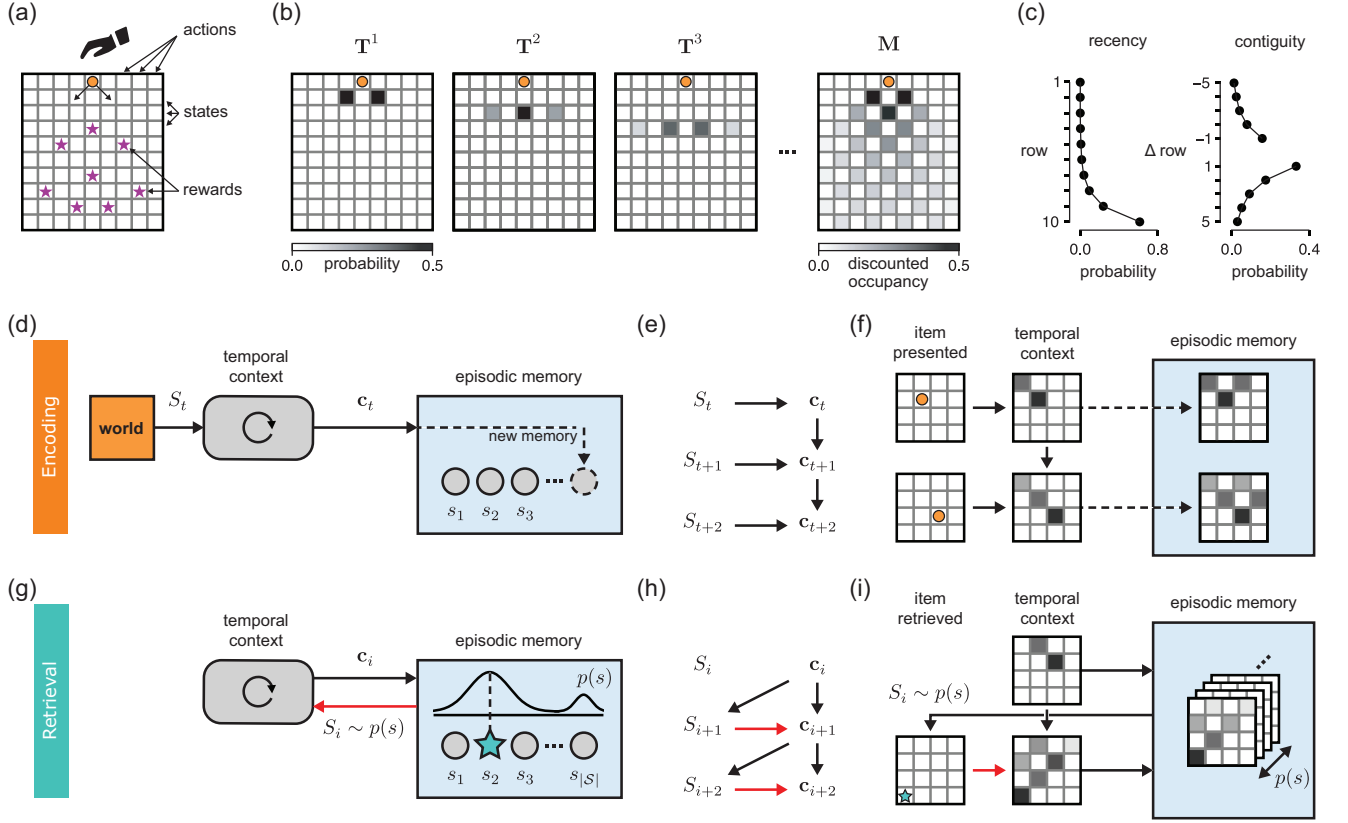
An alternative and often more efficient RL approach for estimating action values is to first learn, for each action, how many visits to each future state can be expected—formally, $\mathbf{M} = \mathbf{T}^1 + \gamma^1 \mathbf{T}^2 + \gamma^2 \mathbf{T}^3 + \dots$, where each element M_{ij} of matrix \mathbf{M} represents the discounted number of visits to state j from state i . \mathbf{M} is known as SR (Dayan, 1993), a predictive representation that humans and animals appear to learn and use, based on behavioral and neural evidence (Momennejad et al., 2017; Piray & Daw, 2021; E. M. Russek et al., 2017, 2021; Stachenfeld et al., 2017). Like \mathbf{T} , the SR matrix \mathbf{M} summarizes the transition structure of the world but aggregated over multiple steps; thus, like \mathbf{T} , it can also be understood as a form of world model (Figure 1b, board labeled \mathbf{M}). With the SR, action values can be estimated straightforwardly by multiplying the expected number of visits to each state by the rewards present in those states—that is, $q(a) = \mathbf{x}_a^T \mathbf{M} \mathbf{r}$, 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, 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} (Piray & Daw, 2021; E. M. Russek et al., 2017). In sum, rollouts and the SR are two model-based strategies for action evaluation with their unique advantages and disadvantages.

In this article, we show that the properties of episodic memory imply an additional approach for estimating action values. This approach generalizes and interpolates between the rollout-based and SR-based approaches, balancing two different strategies for long-term prospection and evaluation. 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 rolloutlike (iterative) state simulation process that differs from standard (noniterative) uses of the SR described previously. Accordingly, we next describe the processes of memory encoding and retrieval that we will later link to value estimation.

Episodic Retrieval via the TCM

Our starting point is a standard model of memory encoding and retrieval, the TCM (Howard & Kahana, 2002a), 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 (Figure 1c). To explain these results, TCM centrally posits that such episodic retrieval is affected by a drifting *temporal context* \mathbf{c} , a continuously evolving representation given by:

Figure 1
Overview of the TCM-SR Model



Note. (a) Our Plinko game has 10×9 states, each represented by a small square. The agent may take any of nine possible actions, corresponding to the nine 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 (pink stars) along the way. 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 T^1 , T^2 , and T^3 represent the probability distribution of the ball location one, two, and three time steps after the moment depicted in (a), respectively. The Plinko board labeled M represents the successor representation (SR), given by $M = \gamma^0 T^1 + \gamma^1 T^2 + \gamma^2 T^3 + \dots$. SR values corresponding 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 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 c_{t+1} depends on incoming sensory information S_{t+1} and on the previous temporal context c_t . (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. Dashed arrows indicate accumulative change to the stored episodic memory M^{SC} . (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 c_{i+1} depends on the previous temporal context c_i and the retrieved stimulus S_{i+1} , which itself also depends on the previous context 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, whose associated stimulus is sampled using the stored episodic memory M^{CS} . TCM = temporal context model; SR = successor representation; SC = stimulus-to-context matrix; CS = context-to-stimulus matrix.

$$c_t = \rho c_{t-1} + \beta c_t^N. \quad (1)$$

At each moment t , the temporal context c_t is updated by input c_t^N , typically due to information arriving either externally through the senses or internally through memory retrieval. Yet, this update is only partial, with $0 < \beta < 1$ representing how much new information is assimilated and $0 < \rho < 1$ representing how much of the previous context is retained (we will constrain $\rho + \beta = 1$ for simplicity). As

such, the temporal context c_t is a recency-weighted average of past inputs, with older information decaying quickly for high values of β (and low values of ρ) and older information decaying slowly for low values of β (and high values of ρ).

During encoding, each observed stimulus S_t becomes associated with the temporal context c_t present at that moment (Figure 1d–f). Formally, $M^{CS} \leftarrow M^{CS} + x_t c_t^T$, where x_t (shortform for $x(S_t)$) is the representation of stimulus S_t in terms of features, and M^{CS} stores the

associations between item and context. During retrieval, the probability of retrieving an item is proportional to how well the context associated with that item matches the current temporal context, or $p(s_k) \propto \mathbf{M}^{\text{CS}} \mathbf{c}_i \cdot \mathbf{x}_k$ (Figure 1g–i). Retrieval is thus determined by the current context and the agent’s memory (Figure 1h and i), that is, the set of associations between each previously seen item and the corresponding stored context. Once an item is retrieved, the temporal context is updated by Equation 1, which in turn affects the retrieval of subsequent items (Figure 1g–i). The assumption of a temporal context that changes with each retrieval is essential to explain the patterns of sequential retrieval observed in free-recall experiments.

TCM recapitulates two recall biases often observed in free recall: the recency effect and the contiguity effect (Figure 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, as a descriptive model, the goal of TCM is to reproduce rather than rationalize or justify these empirically observed patterns.

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. In this task, states S_i are ball locations in Plinko (corresponding to words in a free-recall study). In the learning phase (the encoding phase in TCM), the agent learns the trajectories that can follow from each action by experiencing episodes of the ball dropping through the Plinko board. Each episode is a sequence of states S_1, S_2, \dots, S_H representing a trajectory followed by the ball in Plinko (viewed as a word list stored in memory). In the decision phase (the retrieval phase in TCM), the agent must decide which action to select by using the information previously learned about the trajectories. We propose that an action can be evaluated by retrieving memories of locations that may follow that action and the rewards in those locations, akin to an agent querying an episodic memory for choice-relevant information.

To understand this process, we first note that the associations between stimulus and context, learned during the encoding phase of TCM, amount to learning which stimuli *precedes* a given stimulus. Equivalently, the encoding phase of TCM amounts to learning that a given stimulus is a *successor* all of the preceding stimuli. Indeed, after extensive experience, the associations between item and context encode the SR, that is, $\mathbf{M}^{\text{CS}} \rightarrow \mathbf{M}^{\text{T}}$ (see Gershman et al., 2012, or the Method section for a formal demonstration; also, notice how the episodic memory representations in Figure 1f share characteristics with \mathbf{M} in Figure 1b). This crucial equivalence is the basis for the remainder of this article, which leverages this observation about the encoding phase of TCM to examine predictions regarding memory *retrieval*. In particular, we will show that the retrieval process assumed by TCM can be used to compute action values $q(a)$, a potential mechanism for *model-based* or *goal-directed* decisions in the brain.

In the following sections, we examine the role of each known episodic memory property, formalized in TCM, in supporting value estimation. We begin with a stripped-down version of TCM that illustrates the key ideas behind our theory and serves as the basis for the more realistic variants that are presented subsequently. This initial model makes assumptions that simplify both the encoding and the retrieval process assumed by TCM. The choice to start our investigation this way is purely didactic and by no means suggests that human behavior happens in this manner—we only use it as a simplified baseline to easily analyze and understand the function of individual episodic memory features (akin to ablation studies). We then relax the assumptions, one at a time, starting with more realistic retrieval dynamics based on human free-recall data, moving to emotional modulation by reward, and finally examining the full TCM model. We show that gradually adding each known property of episodic memory (formalized in different variants of TCM) leads not only to more realistic models of evaluation but also to unexpected advantages for decision making. These advantages include the online control of temporal horizon, a speed–accuracy tradeoff, and improvements in sample efficiency.

Independent Samples From Memory Yield Unbiased Value Estimates

To study how episodic retrieval supports evaluation, we start by making two simplifying assumptions about TCM that subsequent sections later relax. The first assumption is that each stimulus (state) is presented many times during encoding. We make this assumption to simplify our predictions of retrieval, acknowledging that episodic memory thrives in the exact opposite scenarios of one- or few-shot learning. As discussed previously, this simplifying assumption results in associations between item and context that correspond to the SR, that is, $\mathbf{M}^{\text{CS}} = \mathbf{M}^{\text{T}}$.

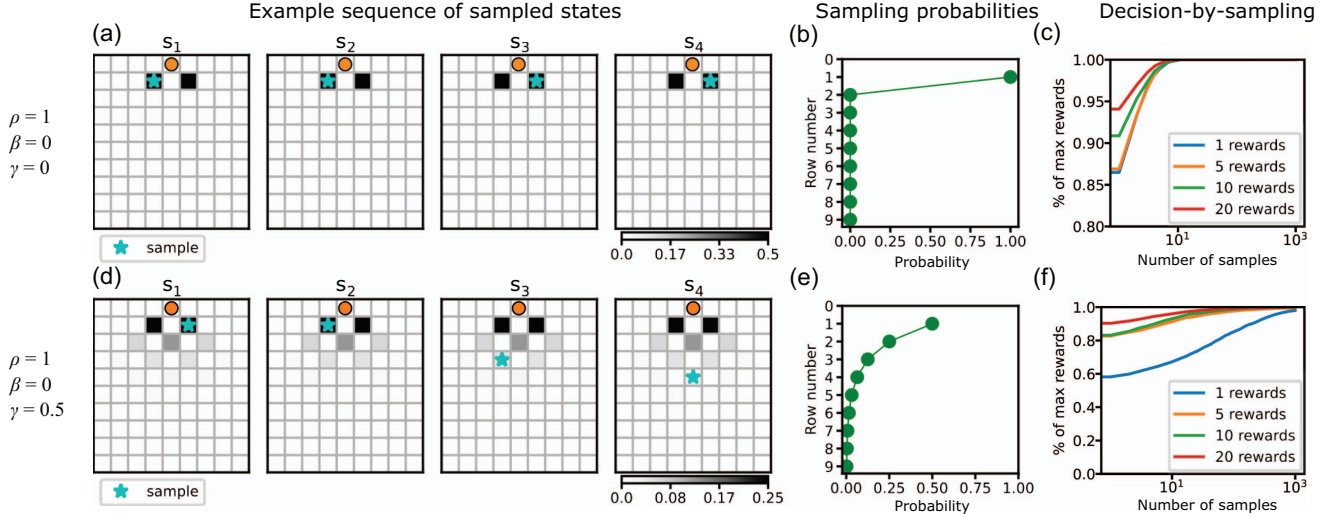
The second simplifying assumption is that the retrieval of a stimulus does not affect the temporal context. That is, we set $\beta = 0$ and $\rho = 1$ in Equation 1, leading to $\mathbf{c}_i = \mathbf{c}_{i-1}$ (this is equivalent to removing the red arrows in Figure 1g–i). Note that, because the context is not updated during retrieval, 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 simplified setting, retrieved stimuli can be viewed as independent “samples” drawn from the same underlying distribution, that is, they are independent and identically distributed (i.i.d.) samples.

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 (Figure 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:

$$\hat{q}_{\beta=0}(a) \propto \frac{1}{N} \sum_{i=1}^N \mathbf{r}^{\text{T}} \mathbf{x}(S_i), \quad (2)$$

where $S_1, S_2, \dots, S_N \sim p(s)$ are samples from the normalized SR, that is, $p(s) = \frac{\mathbf{x}^{\text{T}} \mathbf{M}}{\|\mathbf{x}^{\text{T}} \mathbf{M}\|} \mathbf{x}(s)$. In this equation, $r(S_i) = \mathbf{r}^{\text{T}} \mathbf{x}(S_i)$ is the reward

Figure 2
Independent Samples From Memory Yield Unbiased Value Estimates



Note. (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. Grayscale 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 Equation 2 and then selects the action with the larger estimated value. At least one reward is placed on the second row from the top, with no reward on the topmost row or the bottom three rows. Rewards are reachable from either action. 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: $\rho = 1$, $\beta = 0$, $\gamma = 0.5$. Rewards are uniformly placed between the second and the seventh rows (inclusive) and are reachable from either action. S = sequence of states.

present in state S_i . Thus, $\hat{q}(a)$ is obtained by averaging reward samples (see Lemma 1 in the Method section).

To see how the retrieval phase of TCM can give rise to this sampling scheme, recall that the probability of retrieving an item is proportional to how well the context associated with that item matches the current temporal context, or $p(s) \propto [\mathbf{M}^{\text{CS}} \mathbf{c}] \cdot \mathbf{x}(s)$. If the temporal context is set to the action to be evaluated to eliminate any residual effect of recent history (i.e., $\mathbf{c} = \mathbf{x}_a$) and leveraging the first assumption above (i.e., $\mathbf{M}^{\text{CS}} = \mathbf{M}^{\text{T}}$), then TCM predicts that the probability of retrieving an item s is given by $p(s) \propto [\mathbf{M}^{\text{T}} \mathbf{x}_a] \cdot \mathbf{x}(s) = [\mathbf{x}_a^{\text{T}} \mathbf{M}] \cdot \mathbf{x}(s)$. For a one-hot representation of the current context \mathbf{x}_a , the first term $\mathbf{x}_a^{\text{T}} \mathbf{M}$ is the row of the SR matrix M corresponding to action a . Thus, with the simplifying assumptions above, TCM predicts that the probability of retrieving each item is proportional to the SR of the queried action.

Intuitively, the agent retrieves a sequence of successor states and their respective rewards (Figure 2a). Equation 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 that 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 (see Theorem 2 in the Method section). 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 during encoding β_{enc} 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 occurring soon after the starting context (note that we distinguish the drift rate at encoding β_{enc} from the drift rate at retrieval, which we assumed to be $\beta = 0$). In RL terms, the drift rate at encoding modulates the temporal horizon of the SR, parameterized by the discount factor $\gamma = 1 - \beta_{\text{enc}}$.

By affecting the temporal discount factor, the drift rate at encoding ultimately affects the overall value estimated during retrieval. Depending on the discount factor, the computed value ranges between (a) rewards sampled exclusively from imminent states ($\gamma = 0$, Figure 2a and b) and (b) rewards sampled from all future states, with a preference for earlier states ($\gamma > 0$, Figure 2d and 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., Bornstein et al., 2017; Lieder et al., 2018; Plonsky et al., 2015). The latter case ($\gamma > 0$) extends the i.i.d. decision-by-sampling approach to sequential problems. Unlike rollout-based algorithms like Monte Carlo Tree Search, 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 treelike 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 temporally extended decisions into bandit problems studied previously, extending the findings from sampling models to the sequential case.

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 efficient 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), then it can be identified with many fewer samples than needed to estimate all action values accurately. Second, if no action value dominates the others, then 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.

In our Plinko simulations with $\gamma = 0$, over 85% of the maximum available rewards can be obtained with a single sample, in line with results obtained from bandit problems (which the setting of $\gamma = 0$ corresponds to; Figure 2c). This prediction aligns with previous work demonstrating that surprisingly few samples are needed for effective decisions in bandit problems (Vul et al., 2014). For $\gamma = 0.5$, corresponding to an average drift rate at encoding, the SR extends further into the future, leading to a much larger number of states that can be sampled (Figure 2d). While more samples are needed in this case to yield the same fraction of rewards, we found that over 80% of maximum available reward can be obtained with fewer than 10 samples (Figure 2f), unless the available rewards are extremely sparse (e.g., a single reward placed in the Plinko board). These results suggest that, by transforming a temporally extended task into a bandit problem, previous arguments about the efficiency of a decision-by-sampling approach also applies to temporally extended problems.

In sum, if retrieval does not update the temporal context (i.e., $\beta = 0$), action values can be estimated straightforwardly by sampling stimuli i.i.d. from episodic memory and averaging the corresponding rewards. That is, 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 recapitulate them by retrieval at a choice time. However, unlike previous invocations of SR in decision neuroscience and RL, this retrieval is accomplished by iteratively sampling of individual future states rather than by an instantaneous 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.

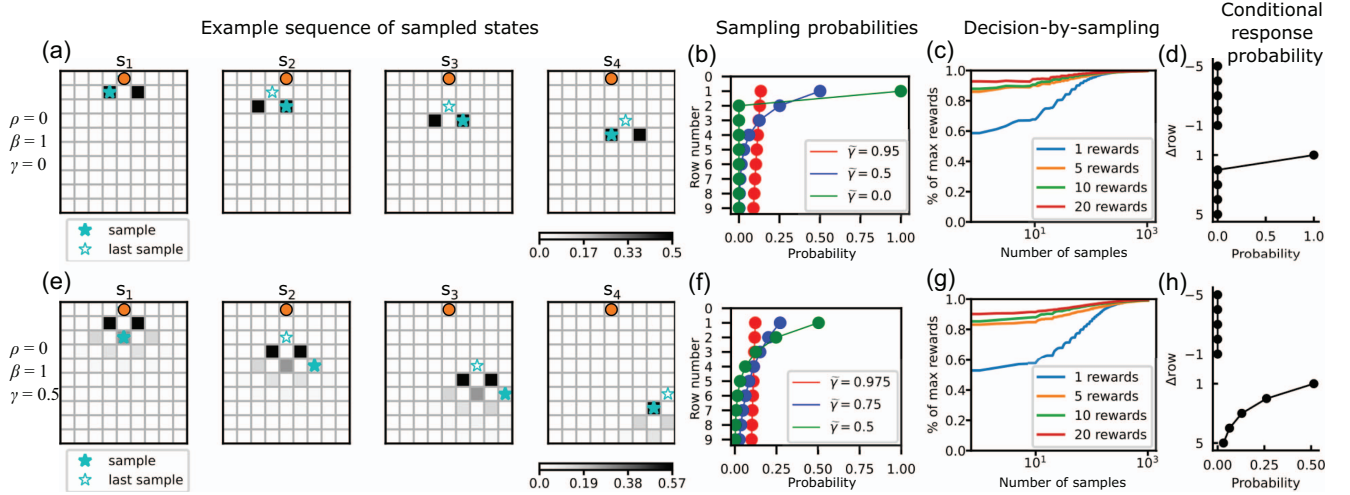
We focus initially on the extreme case where retrieval depends only on the immediately preceding retrieved stimulus—that is, we set $\beta = 1$ and $\rho = 0$ in Equation 1 to yield $\mathbf{c}_i = \mathbf{c}_i^{\text{IN}}$. We also make another simplifying assumption that this update is driven by a static, task-independent representation of each stimulus, or $\mathbf{c}_i^{\text{IN}} = \mathbf{x}_i$, an assumption that we also relax in the last section. In this setting, the temporal context is completely updated by each retrieval, that is, $\mathbf{c}_i = \mathbf{x}_i$, retaining no information retrieved before that. Since retrieval depends on the temporal context, which in turn depends solely on the memory most recently retrieved, this setting leads to a Markov chain where each sample depends on the last sample (and, given the last sample, it is independent of previous samples). We will show that this regime of correlated samples can also be used to estimate action values.

As seen previously, the probability of retrieving an item is given by $p(s) \propto [\mathbf{M}^{\text{CS}} \mathbf{c}_i] \cdot \mathbf{x}(s)$. With the assumptions that $\mathbf{c}_i = \mathbf{x}_i$ and that $\mathbf{M}^{\text{CS}} = \mathbf{M}\mathbf{I}$, this distribution is simplified to $p(s) \propto [\mathbf{M}\mathbf{I} \mathbf{x}_i] \cdot \mathbf{x}(s) = [\mathbf{x}_i^{\text{T}} \mathbf{M}] \cdot \mathbf{x}(s)$. Thus, with the simplifying assumptions above, TCM predicts that the probability of retrieving each item is proportional to the SR of the previously retrieved item \mathbf{x}_i .

As previously, the temporal context drift rate at encoding 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 case of $\gamma = 0$, the first retrieved memory is an immediate successor of the considered action (because $\mathbf{M} = \mathbf{T}^1 + \gamma^1 \mathbf{T}^2 + \dots = \mathbf{T}^1$ when $\gamma = 0$, Figure 1b). Upon retrieving the first memory and updating the temporal context, the second retrieved memory is an immediate successor of the first sample (Figure 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; Figure 3a and b).

Note that since each retrieved item promotes the retrieval of successor states, this regime explains only part of the contiguity effect: It predicts the recall of items encoded *after*, but not *before*, the just-recalled item (Figure 3d). This is caused by the assumption that $\mathbf{c}_i^{\text{IN}} = \mathbf{x}_i$. During encoding, this assumption means that \mathbf{x}_k only contributes to the temporal contexts associated with item(s) presented at a later time (i.e., \mathbf{x}_k only contributes to the context \mathbf{c}_t for $t > k$ if $\gamma > 0$; and for $t = k + 1$ if $\gamma = 0$). During retrieval, then, the temporal context previously updated by the last retrieval, $\mathbf{c}_i = \mathbf{x}_i$, will only drive the retrieval of items encoded after, but not before, the just-recalled item. This leads to a unilateral contiguity effect that differs from the bilateral effect found empirically and predicted by the original TCM (compare Figure 3d and Figure 1c, right). The original TCM predicts the bilateral contiguity effect because it assumes that the information retrieved from episodic memory, \mathbf{c}_i^{IN} , includes not only the preexperimental item representation \mathbf{x}_i but also the contextual state associated with that item and learned during the encoding phase (Howard & Kahana, 2002a). With this more general

Figure 3
Recall-Dependent Context Updates Lead to Rollouts



Note. (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). s_i shows the i th stimulus sampled. Grayscale 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} (.05, .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 Equation 3 and then selects the action with the larger estimated value. Rewards are uniformly placed between the second and the seventh rows (inclusive) and are reachable from either action, except that at least one reward is placed on the second row from the top when $p_{\text{stop}} = 1$. The image shows the fraction of maximum rewards (y -axis) expected as more samples are drawn (x -axis, shown in log-scale), setting $p_{\text{stop}} = .05$ 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: $\rho = 0$, $\beta = 1$, $\gamma = 0.5$. Rewards are uniformly placed between the second and the seventh rows (inclusive) and are reachable from either action. S = sequence of states.

formulation, which we study in the last section, items encoded either before or after the just-recalled item can be recalled.

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

$$\hat{q}_{\beta=1}(a) \propto \sum_{i=1}^N \mathbf{r}^\top \mathbf{x}(S_i), \quad (3)$$

where S_1, S_2, \dots, S_N are samples from the normalized SR with $p(S_1 = s) = \frac{\mathbf{x}_1^\top \mathbf{M}}{|\mathbf{x}_1^\top \mathbf{M}|} \mathbf{x}(s)$ representing the SR of the queried action, $p(S_2 = s) = \frac{\mathbf{x}_1^\top \mathbf{M}}{|\mathbf{x}_1^\top \mathbf{M}|} \mathbf{x}(s)$ representing the SR of the first sample, and so on. Note that each stimulus of the trajectory S_1, S_2, \dots, S_N is drawn from a different distribution (see Lemma 3 in the Method section).

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 estimate of 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 to note that such rollouts can be used to estimate action values.

In Equation 3, each rollout incorporates all future rewards with equal weight. The action value estimated in this way, therefore, has an effective discount factor of one (because sooner and later rewards are weighted equally). This is a surprising result because the sampling distributions specified by the normalized SR were encoded with a temporal context drift rate of $\gamma = 0$ (see Theorem 4 in the Method section). In other words, the rollouts during retrieval lead to an *effective* discount factor (denoted $\tilde{\gamma}$) of one, in stark contrast to the discount factor of the SR learned during encoding (which we assumed $\gamma = 0$). Note that we had no such mismatch in the previous section, where the effective discount factor obtained during retrieval was always identical to the discount factor of the SR learned during encoding (i.e., $\tilde{\gamma} = \gamma$). Here, the mismatch between $\tilde{\gamma} = 1$ and $\gamma = 0$ arises because retrieving n consecutive memories and summing the rewards according to Equation 3 amount to concatenating n one-step predictions (i.e., $\gamma = 0$), which is equivalent to performing a single n -step prediction (i.e., $\tilde{\gamma} = 1$).

In nonepisodic tasks, where an episodic does not end as in Plinko, weighting all future rewards equally would require each rollout to continue forever. In practice, a rollout that ends at a certain point includes all rewards sampled before the interruption with equal weight and none of the later rewards, effectively reducing the discount factor. Here, we posit a fixed probability of interrupting the retrieval process at any moment, denoted p_{stop} . This parameter is intended to capture the fact that the decision maker can choose how long to recall for (note that previous models in the TCM family similarly posited a stopping rule to terminate recall, e.g., Sederberg et al., 2008). The larger the interruption probability, the less likely

the rollout is to continue far into the future. In other words, the interruption probability leads to a larger probability of sampling states following closely from the queried action in comparison to states distant from the queried action, enabling an overweighting of imminent rewards in comparison to distant rewards that results in exponential discounting. The effective discount factor in this case is given by $\tilde{\gamma} = 1 - p_{\text{stop}}$, where p_{stop} is the interruption probability (see the Method section for details, especially Proposition 4.1).

All this raises a potentially confusing but important 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, which we assume fixed at retrieval. Sampling i.i.d. from this encoded SR (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 because the corresponding states are less likely to be retrieved), regardless of the duration of the sampling process. In contrast, by performing iterative sequential retrieval from the same model, it is possible to extend this timescale at retrieval time to give more or less weight to later rewards, that is, to estimate values reflecting a larger discount factor than the encoding γ . 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.

Leveraging this sampling strategy, the reliability of a value estimate is again proportional to the number of samples and rollouts performed. As in the previous section, 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; Figure 3c). This suggests that, again, surprisingly few samples are needed for effective decisions in bandit problems (Vul et al., 2014).

Going beyond the extreme case of $\gamma = 0$ studied above, we now consider the case of a general encoding timescale $\gamma > 0$. Here, the first retrieved item is a sample from the normalized SR of the candidate action, and each subsequent recall is a sample from the SR of the previous sample (Figure 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 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 Equation 3. 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 Proposition 4.1 in the Method section).

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 et al., 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. The decoupling of the timescale at retrieval from the timescale at encoding also enables control over the sampling scheme. In the extreme case of $p_{\text{stop}} = 1$, only one sample is drawn on each rollout, resulting in an i.i.d. sampling scheme with the nominal discount factor $\tilde{\gamma} = \gamma$. In the other extreme case of $p_{\text{stop}} \rightarrow 0$, a rollout continues indefinitely, resulting in an effective discount factor of $\tilde{\gamma} \rightarrow 1$. Intermediate values of p_{stop} results in intermediate discount factors $\gamma < \tilde{\gamma} < 1$. Overall, by controlling the interruption probability, the agent can control both the discount factor and the sampling scheme.

Similarly to the strict rollout case seen previously, the efficiency of generalized rollouts is also high: As before, 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; Figure 3g). The efficiency of the generalized rollout is slightly lower than the efficiency of the strict rollout (compare Figure 3c and 3g). This is because, for a prespecified number of samples, the generalized rollout performs more rollouts than the strict rollout, resulting in a slightly higher chance of repeatedly sampling the same state.

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, represents 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, 2002a). Indeed, the best fitting parameters describing context update in free-recall experiments usually fall between the two extremes (i.e., $0 < \beta < 1$ in Equation 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 gives rise to a mixture of sampling distributions. For instance, immediately after the first retrieval, the context mixture enables sampling from either the SR of the queried action (the original sampling distribution) or from the SR of the first sample (the updated sampling distribution). 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 in the previous sections. Each action can be evaluated according to:

$$\hat{q}_\beta(a) \propto \beta \sum_{i=1}^N \mathbf{r}^\top \mathbf{x}(S_i), \quad (4)$$

where $\beta > 0$ and $S_1, S_2, \dots, S_N \sim p(s)$ 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 (see Theorem 6 in the Method section) and otherwise an underestimate of the true action value (see Proposition 7.1 in the Method section); however, a relatively large number of samples is sufficient for an estimate that’s close to the truth (Figure 4c and f).

The same insights gained in the previous sections apply here, including extension of the effective discount factor with a larger β (Figure 4b and e) and the sample efficiency during decision making (Figure 4c and f). Notably, due to the partial updating implemented by setting $\rho = \beta = 0.5$, the effective discount factors 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 Figure 4b and e) no

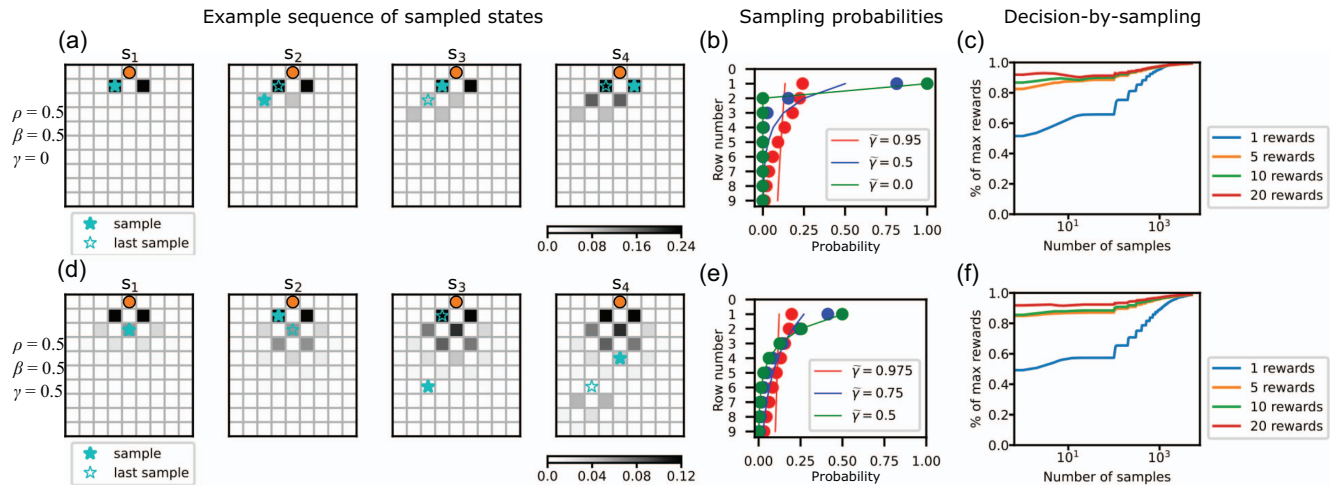
longer capture the empirical sampling distributions under the same p_{stop} unless $p_{\text{stop}} = 1$ (dots in Figure 4b and e). Recall that the larger the β , the larger the effective discount factor, which facilitates sampling of states further away from the action. That is, β controls the degree to which the timescale at retrieval is extended (Figure 4a and 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 Figure 4b).

Similar to generalized rollouts, a nonzero discount factor during retrieval results in slower convergence due to state skipping (Figure 4c vs. Figure 4f). However, unlike generalized rollouts (Figure 3c vs. Figure 3g), the difference between zero and nonzero discount factors is smaller. This is because the drift rate β is less than 1 in the intermediate regime, and the agent may occasionally jump back to visit a (rewarding) state that was previously skipped over.

In sum, 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 versus 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 et al., 2005; Keramati et al., 2011; Nicholas et al., 2022).

All simulations so far only consider the case of unlimited experience (i.e., multiple rounds of encoding; sampling from a converged SR).

Figure 4
An Intermediate Regime Between i.i.d. Sampling and Rollouts



Note. (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. s_i shows the i th stimulus sampled. Grayscale 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 Figure 3b) given the same discount factors. We illustrate these distributions for three values of p_{stop} (.05, .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. Rewards are uniformly placed between the second and the seventh rows (inclusive) and are reachable from either action. The image shows the fraction of maximum rewards (y -axis) expected as more samples are drawn (x -axis, shown in log-scale), setting $p_{\text{stop}} = .05$ as a function of different numbers of rewards placed on the Plinko board. (d–f) As in a–c, but using parameters: $\rho = 0.5, \beta = 0.5, \gamma = 0.5$. S = sequence of states. i.i.d. = independent and identically distributed.

The next section extends our predictions to settings when only limited experience is available.

With Limited Experience, Retrieval Is Based on Trajectories

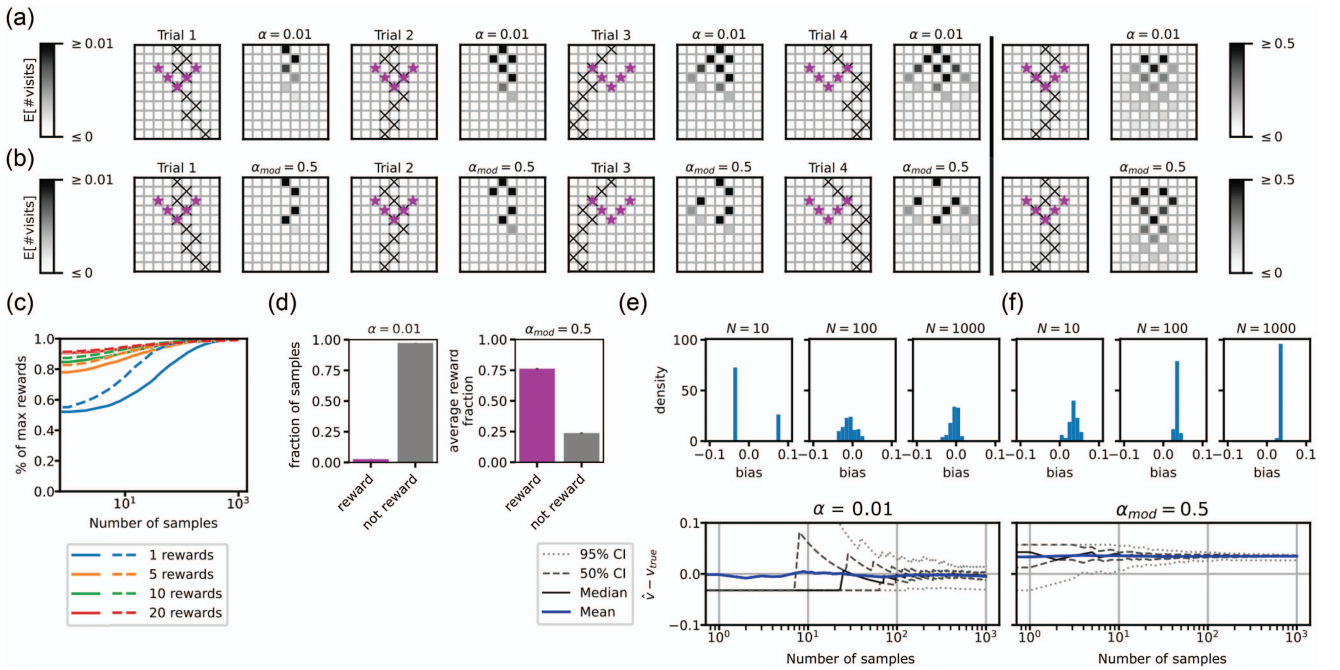
Our simulations thus far assumed that the retrieval simulations we describe are preceded by an extensive encoding phase in which each state (location on the Plinko board) is encoded a large number of times. With repeated exposure, the associations formed between stimuli and contexts converge to the true 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 (Gershman & Daw, 2017; Lengyel & Dayan, 2007; 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 TD 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 (Figure 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 or rollouts 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

Figure 5
Retrieval With Limited Experience and With Emotional Modulation



Note. (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 stars) 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 $\rightarrow \infty$, presented in the five pairs of panels going from left to right, all without emotional modulation ($\alpha = .01$). SR is updated after each trial using Equation 11. (b) The same as (a) but now with emotional modulation ($\alpha = .01$ for unrewarded states and $\alpha = .5$ for rewarded states). The same rule Equation 11 is used with different α s depending on the state. (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 Equation 2, with (dashed lines) and without emotional modulation (solid lines). Rewards are uniformly placed between the second and the seventh rows (inclusive) and are reachable from either action. 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 $p_{stop} = .05$ as a function of different numbers of rewards placed on the Plinko board. (d) Average fraction of sampled states with and without a reward. 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, 1,000 samples. Bottom: mean discrepancy between the true value and the estimated value as a function of number of samples on a log scale. SR = successor representation; CI = confidence interval; mod = emotionally modulated.

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 (Figure 5a, Trial 2). The learned SR comes to represent a composite of possible trajectories as experiences expand, eventually converging to the steady-state SR (Figure 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 (Gershman & Daw, 2017; Lengyel & Dayan, 2007). 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. For example, given trajectories ABDE and ACDF, our model predicts that rollouts along ABDF or ACDE are possible. Furthermore, states in the beginning of an experienced trajectory (predictions of the near future vs. distant future) are prioritized for retrieval, resulting in evaluations that are subject to an effective (and adjustable) 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 et al., 2018). Even when reward is not signaled overtly, signals that some stimuli should be prioritized promotes their retrieval (Mather et al. 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 (Cohen & Kahana, 2019; Talmi et al., 2019). Formal models of emotionally enhanced memory have attributed the effect either to a differential learning rate (Cohen & Kahana, 2019; Talmi et al., 2019) or differential information decay (Zhou et al., 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. Given that stimuli that are emotionally arousing, salient, and goal relevant typically increase memory, especially when measured through recall, we gloss over the many differences between emotional stimuli, prioritized stimuli, and rewards versus punishments with varied magnitude, by referring to all of them as “emotionally salient” or “important” states. We speak generally about “emotional modulation” to refer to their (often similar) effects on memory, especially in the free-recall setting most relevant to decision by sampling (Talmi et al., 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. Clearly, it can be wasteful to retrieve a large number of memories with no affective value. 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 (Figures 2c and f; 3c and g; and 4c and 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 conduct bias sampling toward the most relevant states (here, the goal), even if biasing the sampling procedure might lead to biases of the estimated payoff $q(a)$ (Lieder et al., 2018). We suggest that such favorable biasing can be accomplished by (and conversely helps to justify) emotionally modulated retrieval, which preferentially retrieves emotionally salient states. Here, we operationalize emotionally salient states as those with unusually large rewards or punishments.

Computationally, an emotionally modulated retrieval results in a *bias–variance trade-off*: preferential retrieval of emotionally-salient stimuli disproportionately influences the final evaluation, resulting in an *estimation bias*, that is, either an over- or an underestimation 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 toward emotionally salient stimuli, the larger the estimation bias and the 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 α_{mod} to encode emotionally salient stimuli into memory according to Equation 10 (Horwath et al., 2023; Talmi et al., 2019) as opposed to the normal learning rate α . An agent learns at rate α when the encoding is not affected by emotional arousal (due to rewards). With emotional modulation, it encodes rewards with a different learning rate α_{mod} , where $\alpha_{\text{mod}} > \alpha$ to reflect the enhanced encoding effect of emotional arousal. This means that the learned SR will be skewed

toward the rewarded states (Figure 5b). Consequently, in the Plinko game, states associated with rewards are sampled more frequently during retrieval (Figure 5d, right). Without emotional modulation, rewarded states would have been sampled only rarely (Figure 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 (Figure 5f, compare with Figure 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 (Figure 5c), suggesting that biased encoding can lead to a more efficient sampling process.

When the board contains exactly one reward, the agent has to discern between two options (locations at the top of the board) to drop the Plinko ball, where by design only one of them can possibly reach the reward with a small but nonzero probability while the other option has value zero. If an agent fails to sample any rewarding board location from a given option, it is unclear whether the value of the option is truly zero, or it simply had bad luck since the reward is so sparse. In this case, emotional modulation significantly increases the chance the (only) reward will be recalled, so the agent may differentiate the options much faster (Figure 5c, solid blue line vs. dashed blue line). The advantage continues, albeit to a diminishing extent, when the board contains more and more rewards, since even without emotional modulation the agent might sample many rewards to inform its decision. Nonetheless, overrepresentation of rewarding samples preserves the value difference between options, making it more salient at a low-sample scheme to aid faster decision making. Of course, when the sample size is large enough, the unmodulated agent can do just as well as its emotionally modulated counterpart.

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, the contiguity effect is bidirectional: Stimuli are also more likely to be recalled if they were experienced *before* as well as after the just-recalled stimulus (Figure 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 (Figure 3d and h) is because of one final simplification that has not yet been reexamined. We have assumed that when a memory is retrieved, it directly updates the temporal context with a static, task-independent, one-hot representation of the retrieved stimulus (\mathbf{x}_t in Equation 1; Figure 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 the stimulus is experienced. In particular,

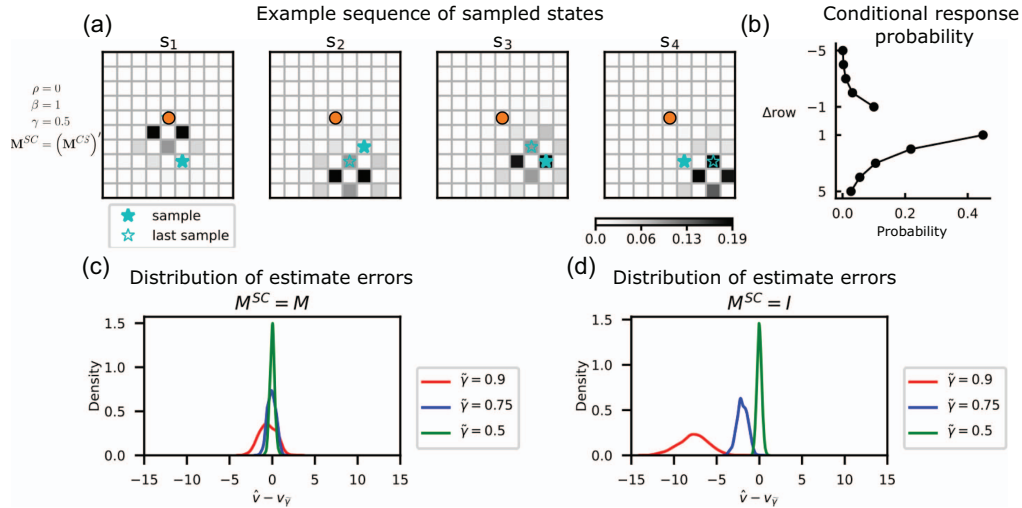
TCM assumes that the temporal context is updated by a retrieved *context* associated with a given stimulus, instead of being updated by the stimulus representation \mathbf{x}_t itself. Formally, the temporal context is updated during retrieval according to $\mathbf{c}_t = \rho\mathbf{c}_{t-1} + \beta\mathbf{c}_i^{\text{IN}}$, where $\mathbf{c}_i^{\text{IN}} = \mathbf{M}\mathbf{x}_i$, that is, \mathbf{c}_i^{IN} is the column of the SR indexed by the stimulus.

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, that is, in 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 $A \rightarrow B \rightarrow A$, $C \rightarrow B \rightarrow C$), which in turn generalizes action evaluation. Ideally, the relative strength of forward versus reverse contiguity (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 ball can move diagonally in any direction and 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 (Figure 6d). A retrieved context aids the agent to go beyond unidirectional experience and correctly estimate the values for the reversible Plinko (Figure 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.

Figure 6
Retrieving a Learned Context Allows Backward Sampling



Note. (a) An example sequence of memory retrieved when initiating the temporal context with the state shown as an orange circle and using $\gamma = 0.5$. s_i shows the i th stimulus sampled. Grayscale 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 successor representation 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 Markov decision process (i.e., a grid world rather than a Plinko game). Rewards are uniformly placed between the second and the seventh rows (inclusive). (d) As in (c), but using the identity matrix as the feature-to-context association matrix (as in the previous simulations). S = sequence of states; SC = stimulus-to-context matrix; CS = context-to-stimulus matrix.

Discussion

Summary of Findings

In this article, we proposed TCM-SR, a novel model of decision making that grounds model-based evaluation in the recall of episodic memories. 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 via model-based evaluation. 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 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 establishes a formal mapping between the well-studied TCM of episodic recall and the normative concept of the SR, a model of the world that is widely studied in RL. From TCM, our model inherits a drifting temporal context that integrates the agent's recent experience during memory encoding and guides retrieval. The agent then evaluates actions by retrieving memories from the SR, corresponding to task's states and the rewards expected to

result, as predicted by TCM. Such recursive retrieval implements a parameterized family of sampling algorithms that, when applied to sequential decision problems, enables action values to be straightforwardly estimated. 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 (Liu et al., 2021; Mattar & Daw, 2018; Momennejad et al., 2017, 2018; E. M. Russek et al., 2017, 2021; Stachenfeld et al., 2017). 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 but turn out to be adaptive for choice.

Implications for Decision Making

Our goal in this study was to investigate, in computational detail, the suggestion that episodic memory contributes to decision making. Despite being an ubiquitous view, most of the previous research on this topic is based on a relatively shallow analogy identifying "episodic memory" with a stylized memory store (essentially a perfect record of individual per-trial experiences) that is otherwise uninformed by research into the actual properties of episodic memory. While we acknowledge that the view of episodic memory we incorporate is still quite abstract and stylized, we believe that it is a large step forward in this respect from previous work and points the way toward additional advances. Furthermore, most of the previous models also treat only a simplified (single-step) decision

problem, whereas the signature problem facing neural decision mechanisms is how they address credit assignment over time in multistep problems, a problem we begin to address here.

Of particular relevance to our work is the model class known as “decision-by-sampling,” which posits that decision variables are constructed by integrating a handful of selective memory samples (Bornstein et al., 2017; Bornstein & Norman, 2017; Lieder et al., 2018; Plonsky et al., 2015). These models offer a parsimonious explanation to a number of empirically measured decision biases, yet they have only been examined in the single-step bandit case. Building upon previous studies of episodic memory and decision making, TCM-SR extends models of banditlike evaluation by sampling (Bornstein et al., 2017; Bornstein & Norman, 2017; Duncan & Shohamy, 2016; Nicholas et al., 2022; Rouhani et al., 2018; Zhao et al., 2021) into the sequential realm—a broader, more realistic, and more challenging class of problems. Banditlike evaluation arises as a special case in TCM-SR, 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; Lieder et al., 2018) to the sequential domain. By explicitly integrating the effects of contiguity from episodic memory research, we also move beyond the simplified relationship between decision and memory assumed by previous models (Braun et al., 2018; Duncan & Shohamy, 2016), laying out a new territory to systematically formulate and test the role of episodic memory in decision making down to the process level.

The crucial ingredient that allowed the TCM-SR to generalize from one-step bandits to sequential decision problems is the SR. Prior work suggests that the SR explains numerous patterns in human behavior (Momennejad et al., 2017; E. M. Russek et al., 2017) and in the activity of hippocampal neurons (Brea et al., 2016; Garvert et al., 2017; Stachenfeld et al., 2017). Our model leverages a previously established equivalence between TCM encoding and SR learning (Gershman et al., 2012). The observation that memory encoding gives rise to a representation useful for decision making is highly suggestive of an actual role in guiding decisions, yet the precise instantiation of this process remained unexplored thus far. Our model builds upon this foundation to address the retrieval and choice side of the problem. Through simulations and derivations, we show that the mechanisms of TCM predict a completely new role for the SR in evaluation. In particular, TCM predicts a temporally extended process of model-based evaluation via sampling from the SR. Our analytical derivations and simulations show that, when equipped with SR, TCM retrieval and update could give rise to an unbiased value estimator that corresponds to a family of well-known algorithms—i.i.d. sampling and rollouts—and their interpolation. The connection between sampling-based mechanisms and the SR unifies this family of sampling approaches under a common framework.

The prediction of a temporally extended sampling process is a departure from the canonical view in SR models from both neuroscience and artificial intelligence where state values are instead computed instantaneously via a dot product $\mathbf{v} = \mathbf{M}_t \mathbf{r}$ (Dayan, 1993). Due to this temporally extended sampling process, the model we propose here may seem strictly worse than the commonly used SR formulation, due to requiring more time to compute what ultimately are less precise value estimates. We note, however, that our goal in this article was not to improve upon the canonical SR formulation nor to engineer a novel machine learning algorithm balancing flexibility and speed. Instead, our goal was to determine how memory retrieval

supports value computation in humans, for whom empirical data align more with iterative sampling than with a parallel-dot product formulation. The model we proposed, TCM-SR, is an important step toward satisfying this goal, given the fact that each feature of TCM-SR can be converted into an experimental prediction, regardless of how advantageous to behavior those features are. Having said that, TCM-SR may also have some advantages over the vanilla SR formulation. For example, because vanilla SR bases choices on cached, long-term, on-policy state occupancy, it often fails to replan without additional trial-and-error relearning (Momennejad et al., 2017; Piray & Daw, 2021; E. M. Russek et al., 2017). In contrast, TCM-SR can accomplish some degree of replanning by controlling the rollout length and choosing how much caching—and thus, policy dependency—to allow by interpolating between SR-like long-term caching and MB-like step-by-step rollout.

TCM-SR also sheds light onto the particular situations in which episodic control is more or less useful. For example, episodic control has been framed as particularly useful in decision tasks when experience is limited or when the task involves extended dependencies (Gershman & Daw, 2017). In the latter case, generalized rollouts allow the agent to plan and act on a potentially much longer timescale than experienced, a prediction supported by the control of temporal abstraction in TCM-SR. On the other hand, TCM-SR also sheds light onto the converse situations where episodic control is less useful than other forms of control. This is expected, given that people have multiple systems to carry out decision making given different contexts, amount of experience, and resource constraints. For example, model-free processes are likely more suitable given ample experience or if simple one-to-one mapping between behavior and outcome exists. We view such task specificity of episodic control as a strength, not a weakness, of our model. To the extent that episodic memory is used for some task, we hypothesize that the parameters governing it should be adaptable to the circumstances, and our analysis makes clear, testable, directional predictions for future experiments about how memory might change in different task variants or even different cover stories in the free-recall memory setting. Further, but not unrelated, not all decisions must be based on episodic memory—indeed, amnesic patients are not incapable of making decisions. We consider it plausible that episodic memory is used primarily when it is most useful, much like how most people currently view the arbitration of model-free and model-based RL mechanisms for decision making. Here again, our model provides specific hypotheses about the situations where episodic memory may be most or least helpful.

Last, TCM-SR extends the proposal that people overweigh extreme events in simple decision-making tasks to sequential decision problems. This is not a trivial extension since a sequential task is not merely a sequence of single-step tasks; rather, it introduces additional dependencies and, consequently, additional complexity. Multiple successor states could follow an action or a reward. The role emotion plays in sequential decision making is not identical either, for instance, because emotion modulates which successor states are sampled, as opposed to which terminal outcomes are sampled in a single-step task. The extension from single-step samples to a sequential setting thus holds a degree of intricacy that requires more than straightforward aggregation. For a model to capture the emotional effect in sequential decision making, it needs to have the machinery both to predict emotionally modulated recall and to perform systematic sampling that respects

the sequential nature of the task. It is notable that descriptive models of memory (e.g., Talmi et al., 2019) could produce patterns of choice consistent with the normative predictions of Lieder et al. (2018). Thus, our model not only extends the normative results of Lieder et al. (2018) into the sequential setting, but it also provides a mechanism by which memory encoding results in a distorted representation, which, when sampled, automatically leads to an overrepresentation of extreme events.

Additionally, while Lieder et al. (2018) demonstrated the benefit of oversampling extreme events, they did not explore which representation might be learned (at encoding) that results in the right biases at retrieval. In our model, these results emerge naturally by incorporating known results from known episodic memory effects. By modulating the learning rate with which transitions are learned at encoding, an agent learns a biased SR that not only oversamples extreme values but also overpredicts the future occurrence of extreme events and biases any of the subsequent learning. TCM-SR also exposes connections between abstract sampling mechanisms (e.g., the importance weighting scheme in Lieder et al., 2018) and empirically informed details of human episodic memory (e.g., the emotional modulation of episodic retrieval). Such convergence of findings across multiple distinct literatures is a distinctive advantage of our model.

Implications for Episodic Memory

By formalizing the link between episodic memory and decision making, TCM-SR provides normative interpretations for features of episodic memory that have so far been framed only at the process level. We review three examples. First and foremost, TCM and its derivatives do not propose any adaptive function for the contiguity effect. Under TCM-SR, however, the contiguity effect is important because it enables a rollout retrieval scheme that supports model-based evaluation. Crucially, the notion of directed temporal progression implied by the contiguity effect also enables the construction of simulations of future events (Schacter et al., 2015). We view this aspect of memory as key to connecting the hippocampus' role in episodic memory with its long-hypothesized involvement in constructing cognitive maps that enable flexible model-based decisions in spatial and other sequential tasks (Daw et al., 2005; Gershman et al., 2012; O'Keefe & Nadel, 1978; Tolman, 1948; Wimmer & Shohamy, 2011).

Besides contiguity, our model suggests that the preferential retrieval of emotionally salient stimuli, well-characterized empirically and computationally in TCM extensions (Talmi et al., 2018), offers the agent a speed-accuracy tradeoff. Last, the ubiquitous human tendency to recall stimuli in the opposite order than experienced may allow a more efficient use of one's limited experience. With regard to the latter, our simulations show that the increased probability of backward recalls is a consequence of retrieving the encoding temporal context associated with each stimulus, in line with analyses in the original TCM work. Curiously, temporal context reinstatement is disrupted in amnesic patients, who display a preference for recalling items that were after the latest recalled item, but not before (Palombo et al., 2019). Since the same phenomenon is reproduced in TCM-SR by setting $\mathbf{c}_t^{\text{IN}} = \mathbf{x}_t$, we can speculate that our simulations of this regime are a reasonable model of decision making in amnesia. Broadly speaking, we expect a general impairment in episodic evaluation due to the overall decrease in memory performance seen in

amnesic patients. However, we also expect that these patients will retain some ability to perform forward rollouts (and, thus, some degree of model-based evaluation) while completely losing the ability to generalize transition dynamics shown in our last simulations. In sum, we envision these novel interpretations to be critical stepping stones for further inquiries about properties of episodic memory, including what is considered "optimal" in terms of memory dynamics.

In addition to normative interpretations of episodic memory, TCM-SR helps clarify an important distinction between the representation learned during memory encoding and the learning mechanisms that enable this representation to be acquired. Of particular relevance is the argument that, if each stimulus is only seen once, the learning rule in TCM is equivalent to a TD algorithm for learning the SR (Gershman et al., 2012). This equivalence means that the empirical data used to support TCM are also consistent with learning the SR, a possibility favored by the authors for normative reasons. More generally, one cannot know the learning rule simply from knowing the underlying representation, as multiple learning mechanisms (e.g., TD learning, Hebbian learning) can give rise to the same representations (e.g., the SR). This is true even when stimuli are repeated: For a given stimulus sequence, a Hebbian-learning rule with an appropriate decay term will converge to the same representation (the SR) as a TD-learning rule. This leads to two conclusions. First, the learning rule in TCM (Hebbian rule) can be consistent with learning the SR even when stimuli are repeated, so one should not expect a TD-learning rule just because they believe that the SR is learned. Second, inferring learning rules from behavioral data is rather difficult.

In TCM-SR, we followed the assumption in Gershman et al. (2012) that the representation resulting from repeated exposure to stimulus sequences is the SR. The learning algorithm is left unspecified in most the simulations, where we assume that the SR has been entirely (and correctly) learned by whatever learning rule that converges on the SR. The only exception in our article is the simulation of one-shot learning, where we use a TD-learning algorithm (as in Gershman et al., 2012) to model a partially learned SR, though all of the arguments in that section would remain unchanged had we used a different algorithm for learning the SR. We hope that future studies will shed light onto learning algorithms by employing a varying number of repetitions for each stimulus.

Empirical Predictions

By combining TCM and SR, two classes of theories well supported by a large body of experiments and simulations, our model automatically inherits all predictions of either model, including many with substantial empirical support. For example, our model inherits TCM's account of a panoply of list-learning phenomena (e.g., primacy, recency, and contiguity effects; Howard & Kahana, 2002a; Polyn et al., 2009a; Sederberg et al., 2008; 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 multistep MDP; 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 (Mattar & Daw, 2018; Momennejad et al., 2018; Pfeiffer & Foster, 2013).

In addition to inheriting these predictions, TCM-SR also makes 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).

An example property of episodic memory that might depend on task requirements is the strength of backward retrieval. When states or study items reflect nonreversible environmental dynamics—for example, playing a card in a poker game, capturing a piece in a chess game, falling under gravity, or consuming a nonreplenishable reward—a rational RL agent would be expected to dial back the reversibility assumption when learning an SR. In such scenarios, an action value (sum of future rewards) should be computed based only on rewards that the agent has experienced after executing the action. This contrasts with reversible environments (like 2D and real-world spatial navigation), where an action value can be computed based on rewards experienced not only after executing the action but also before it because the latter could still be obtained after the action and should therefore affect its value. Note that, in free-recall experiments, subjects are instructed to recall as many words from a list as possible, *in any order*—that is, subjects can move through the word list in any order. Since the recall order is irrelevant, such tasks can be viewed as having reversible dynamics.

In another example, the usefulness of emotional memory enhancement (Figure 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., 2018). Emotional memory modulation also leads to value estimation biases that can be measured empirically. For example, agents may be asked to choose between two equivalent options in terms of their expected (state-action) value, where Option A only has moderate rewards (or penalties), but Option B presents occasional large rewards (or penalties). The bias can then be quantified as the extent to which Option B is favored. In itself, this sounds similar to classic tests of risk sensitivity, but in this case, the bias may also be modulated by varying aspects of encoding/recall (e.g., list length, recency and primacy effects, adding a distractor task) to encourage more or less episodic sampling.

Relation to Existing Models

TCM Extensions

TCM has been extended in a series of successor models, each focused on explaining additional properties of episodic memory by augmenting TCM with novel processes (Cohen & Kahana, 2022; Lohnas et al., 2015; Polyn et al., 2009a; Sederberg et al., 2008). Our focus on the original TCM was purely didactic: By focusing on the simplest model of this class, we were able to tease apart the effect of each feature of TCM. However, this also meant that we ignored many facts about episodic memory that we hope to incorporate in the future. Much like the evolution of TCM, our initial model provides the scaffolding over which extensions can be added so that the model accounts for increasingly larger bodies of data.

Our model differs from prior models that incorporate effects of repeated learning, such as CMR2 (Lohnas et al., 2015), in terms of the specific encoding mechanisms, but effectively achieves the same outcome. The original TCM as formulated by Howard and Kahana (2002a) uses Hebbian learning, which weighs all past experiences equally (i.e., without temporal decay) and resets the context–stimulus associations upon a new list. Unsurprisingly, this design fails to explain intrusions in recall: That is, if two or more lists are studied, people may recall words that appeared in a list before the most recent one, even though they do prioritize recalling from the latest list (i.e., interlist recency effect).

In CMR2, the Hebbian-learning procedure is inherited from previous formulations of TCM, and additional parameters are introduced to make up for this limitation and account for memory intrusions. TCM-SR, in contrast, builds onto the key observation made by Gershman et al. (2012) that when stimuli are not repeated, Hebbian learning is equivalent to TD learning. Unlike Hebbian updates, however, TD learning in RL naturally incorporates decay of experiences in the distant past without a need for any additional parameters as in CMR2. In the abstracted formal setting of our example task, the repeated learning setting is simulated by observing sequences of Plinko ball positions on the *same* board under the same transition dynamics, where the TCM-SR agent updates the same underlying representation (SR) using TD learning. This departure from Hebbian learning equips the agent with two properties similar to those that are also enabled in somewhat different ways by CMR2: First, episodic memory across multiple experiences are aggregated in a shared representation, analogous to the item-to-context associations in CMR2; second, the exponential discounting of past observations in each TD update naturally imposes a bias toward the most recent “list” (trajectory), corresponding to the interlist recency effect.

While TCM-SR appears to lack the mechanisms present in the successors of TCM (and while we do not directly model recall in list-learning studies), our view is that it may point to a different approach to explaining interlist effects. Moreover, TCM-SR urges us to reconsider memory intrusion from the decision-making perspective: Although recalling words from nontarget lists is generally regarded as a “failure” in free-recall tasks, failure to distinguish experiences between episodes may actually be advantageous for behavior generalization, for example, because it enables integrating experiences from different episodes (different situations, contexts, or settings: stylized in free-recall experiments as lists and in our formal setting as game trajectories) into a broader world model to enable more flexible

inference at retrieval. Figure 5a specifically illustrates how a handful of experiences may be interpolated to facilitate sampling of trajectories that does not exactly correspond to any specific observation but is instead composed of many “intrusions” of observations from past episodes and gives rise to efficient choice. We now sketch these points in the article and also point out that issues of generalization across contexts (as incompletely captured by cross-list effects in list learning and cross-episode effects in RL) are important areas for future empirical and theoretical work.

Last, the artificial task of Plinko involves limited semantic (or, in general, nontemporal) information, and this is not usually the case for real-world decisions. Semantic association and clustering are widely observed and affect memory recall in ways that may interact with temporal organization (Howard & Kahana, 2002b; Polyn et al., 2009b). While TCM lacks the machinery to capture the effects of semantics during encoding and recall, many model descendants of TCM do and do so on top of its foundational framework. For instance, CMR (Polyn et al., 2009a) already incorporates semantic layers, and CMR3 (Cohen & Kahana, 2022) further expands into the domain of emotional effects. We view this as perhaps the major open issue in further connecting these two areas, but to be clear, we also view it as a huge area requiring major conceptual advances to treat properly. Accordingly, we note that future work should look into mapping semantic similarity into the decision domain, especially in a sequential setting. The groundwork of TCM-SR could then be extended to explore even more refined predictions about episodic control.

Besides the extensions in the CMR family, TCM has also been extended to explain the temporal dynamics of free recall. For example, TCM-A posited a retrieval rule based on a leaky-accumulator decision model (Sederberg et al., 2008). We did not incorporate this feature into TCM-SR, as it would have complicated our analyses and derivation of sampling-based value estimation. However, the leaky-accumulator dynamics could be incorporated into future versions of our model. While this means that TCM-SR is unable to explain the temporal dynamics of individual retrievals, the interruption probability parameter enables TCM-SR to explain variability in the total number of recalls in a single rollout, akin to the time of continuous recall. In the memory literature, the decision to stop recalling is a widely studied topic (e.g., Dougherty & Harbison, 2007; J. F. Miller et al., 2012; Murdock & Okada, 1970). A criterion for recall termination is also present in TCM-A (Sederberg et al., 2008), modeled as the probability that none of the leaky accumulators in TCM-A reach the threshold within the prespecified number of time steps. While our implementation of recall termination is much simpler, it can nonetheless explain some empirical findings, such as the exponential growth of interresponse time during free recall (Murdock & Okada, 1970; J. F. Miller et al., 2012). Future work that aims to provide more process-level details regarding the stopping criterion of sampling for decision purposes may extend the interruption probability with TCM-A-like mechanisms or incorporate other retrieval strategies (Badre et al., 2014; Naim et al., 2020).

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. Lengyel and Dayan (2007)

envisioned a three-system architecture (model-free RL, model-based RL, and episodic control), with the episodic controller used primarily in the low-data regime. This controller executes the action that, across all past experiences, has led to the maximum reward. In Gershman and Daw (2017), on the other hand, the episodic RL algorithm computes action values by considering all relevant trajectories the agent has experienced. TCM-SR contrasts with both models by assuming that trajectories are only encoded indirectly via state–context associations while maintaining the ability to simulate rollouts during retrieval. TCM-SR achieves similar action values to episodic RL in most cases. A notable exception is that only TCM-SR is able to combine value estimates across trajectories. Importantly, TCM-SR also describes the *process* of retrieval, predicting that (a) individual states rather than trajectories are retrieved, even if those states are themselves components of a trajectory, and (b) successive retrieved states need not be temporally adjacent to one another. Future work should investigate whether these predictions better describe how humans evaluate actions.

Episodic Versus 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 planninglike 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 last, inactivating the hippocampus in rats causes their behavior to shift from model based to model free (K. J. Miller et al., 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 (Kumaran et al., 2016; O’Reilly et al., 2014). 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 evidences 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.

Gamma Models

The drifting temporal context in TCM-SR resembles the learning of a bootstrapped target distribution in gamma models (Janner et al., 2020). Both models facilitate sampling and model-based rollouts based on an SR, and both exhibit a hybrid of model-free and model-based mechanisms. In particular, when a TCM-SR agent engages in generalized rollouts (i.e., $\rho = 0$, $\beta = 1$), it is equivalent to sampling from the optimal gamma model. Like gamma models, TCM-SR can be understood as incorporating the discriminative SR into a continuous generative model (though we focus more on establishing exact or approximate connections with the traditional MDP model rather than adopting an alternative generative model of events), allowing potentially infinite-horizon predictions and distinct timescales of learning versus control. The interruption probability in TCM-SR, which decouples the retrieval process from the discount factor at encoding, is effectively equivalent to the model-based value expansion (Feinberg et al., 2018).

Crucially, TCM-SR further generalizes the gamma model, showing that the two regimes explored there (sampling from a fixed gamma model vs. rolling these samples out) are in effect special cases of our more general decision-by-sample scheme, as the intermediate sampling regime ($0 < \beta < 1$) we introduce could be used for predictions with more complex (e.g., nongeometric) time step weighting. Most notably, of course, our descriptive model is grounded in a computational model of episodic memory, which makes the parallel with the gamma model even more striking.

Alternative Theories of Episodic Memory

Although we chose to focus on the retrieval dynamics posited by TCM, other theories of episodic memory could also be interpreted in the context of decision making. For example, associative chaining models of episodic memory also allow rollouts to some extent although they differ from TCM in two major ways. First, without significant extensions, early chaining models fail to explain contiguity effects over long timescales, which have been observed in free recall (Howard et al., 2008; Kahana et al., 2008). In contrast, temporal abstraction (i.e., representation of actions and states at different timescales) in TCM can happen across tasks. Thus, while within a compact task, such as Plinko, TCM and chaining models may produce similar predictions about sample retrieval, we expect TCM-SR to capture behavior better when the task spans an extended time. Additionally, chaining models assume rehearsal is required to build associations between nonneighboring stimuli, yet TCM forms such associations at time of encoding without rehearsal. Therefore, TCM-SR predicts generalized rollouts (particularly those skipping over states) in the absence of rehearsal. Considering the formal mappings we have established in this article, these two important characteristics are preserved in future extension of the current model to more complex TCM-based models (e.g., CMR).

Other mechanistic models of episodic memory may also predict rollouts. Dual-store theories predict the asymmetric contiguity effect from a random walk on a one-dimensional context-state space with an added forward bias (Davelaar et al., 2005) and may be used to generate rollouts on a short timescale. Chunking models group temporally adjacent stimuli together and retrieve by chunk, where recall proceeds in the forward direction within a chunk (Farrell, 2012). Similar to associative chaining models, however, both types of models need additional and separate machinery to simulate contiguity over longer time scales.

Extensions and Future Directions

A key simplification of our model is that it treats decisions as a single choice at the start, with subsequent events unfolding passively like a Plinko ball. In contrast, many real-world tasks (like navigating mazes) require actions to be chosen at every step, with these choices influencing the value of the initial action (Sutton & Barto, 2018). While our model does not fully solve this broader class of tasks, the action evaluation problem it addresses is a key subproblem in the more general policy optimization problem (known as “policy evaluation” in the RL literature). If combined with a “policy improvement” process that relearns, recomputes, or readjusts the SR to reflect improved policies as learning proceeds, TCM-SR can lead to optimal policies even in tasks with multiple steps of decisions. However, future work should consider alternative approaches to multistep decision making, 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), akin to traditional value iteration algorithms.

An advantage of our model is that it provides the scaffolding over which additional details about episodic memory can be added. For example, inspired by TCM-A, our model can be augmented with a leaky-accumulator model to capture the temporal dynamics of recall (Sederberg et al., 2008). Similarly, inspired by CMR, our model can be augmented to model the semantic similarity between stimuli, which should account for the empirically observed tendency toward semantically related recall. Last, inspired by CMR2, our model can be augmented to integrate context from encoding to retrieval, which, over repeated learning, can affect sequential recall because of blended contexts. These three examples highlight the fact that TCM-SR can be extended with features incorporated in TCM extensions. The advantage of this approach is that these features have been both incorporated into the TCM framework and validated by empirical data. By incorporating these features into TCM-SR, we will be able to examine their role in decision making and invite their interpretation in normative terms, as we did in this article.

Besides extensions of TCM, future directions include capturing reward effects on memory (Mason et al., 2017) and consolidation (Braun et al., 2018; Mattar & Daw, 2018), equipping TCM with generalization (e.g., over categories, similar to Kumaran & McClelland, 2012, on REMERGE), and incorporating event boundaries (Clewett et al., 2019), where TCM-SR may be extended to offer normative explanations and to produce novel experimental predictions (Wen & Egner, 2022). In particular, eCMR (Talmi et al., 2019) adds value layers on top of CMR, where both models are close successors of TCM such that the modifications are also relevant to

TCM-SR. While we have not yet pursued all these directions, the connection between the TCM family and the full RL formalism in the present work offers a foundation for pursuing these additional avenues.

In TCM-SR, retrieval is driven solely by temporal associations. While this aligns with the view of episodic memory as forecasting the future, the retrieval process in TCM-SR is completely independent of the agent’s goals. While our simulations of emotional modulation reproduce the well-characterized modulation of retrieval by rewards (Mather et al., 2015; Stefanidi et al., 2018; Yonelinas & Ritchey, 2015), empirical data suggest that the agent’s goals also affect the content and order of recall (Aka & Bhatia, 2021). Future work should augment TCM-SR with these findings to account for these data. Our model could also leverage corpus statistics to represent items as nonorthogonal vector embeddings. When used in conjunction with CMR dynamics (but fixed representations), these representations have been shown to account for behavioral patterns in free association tasks, where subjects generate a sequence of words that come to mind in response to a cue word (Richie et al., 2022). It would be interesting to investigate how free associations influence choice in a decision-making task with words (e.g., natural language).

Our simulations suggest that retrieving a temporal context at decision time improves generalization in time-reversible environments. However, we have not examined the implications of context retrieval during encoding. In such cases, TCM predicts that the representation of each stimulus becomes similar to its predecessors (Howard et al., 2011). This may lead to unwanted associations between stimuli if temporal or spatial proximity is not predictive of rewards (similar to the case, for instance, of sentence understanding; Howard et al., 2011). On the other hand, when proximity in the state space is predictive of similar reward outcomes, such as a Plinko game, the features encoded in \mathbf{e}^{IN} could help the agent generalize across novel (potentially continuous) states and improve knowledge transfer. Future work should examine these scenarios in detail.

Relatedly, the encoding phase of TCM-SR bears one notable difference from TCM and CMR, namely, that the stimulus–context associations \mathbf{M}^{FC} do not incorporate task-dependent representations (e.g., predecessors, captured by $\mathbf{M}_{\text{exp}}^{\text{FC}}$ and weighted by γ_{FC} in CMR; see Polyn et al., 2009a, for details) during encoding. This mainly impacts the last set of results where \mathbf{M}^{FC} should be updated to reflect $\mathbf{M}_{\text{exp}}^{\text{FC}}$ at each encoding step (i.e., $\gamma_{\text{FC}} > 0$ throughout encoding), yet Gershman et al.’s (2012) result only applies if $\mathbf{M}^{\text{FC}} = \mathbf{I}$ (i.e., $\gamma_{\text{FC}} = 0$ throughout encoding). Thus the theoretical results we derived require the assumption that the learned task-dependent representations are not incorporated until retrieval and the subsequent evaluation.

Incorporating a nonzero proportion of $\mathbf{M}_{\text{exp}}^{\text{FC}}$ in \mathbf{M}^{FC} no longer leads to the conventional SR. Instead, it captures both the successors and predecessors of a queried action. The cost of the additional predecessor information, however, is a slight underrepresentation of intermediate states that are neither closely following the candidate option nor predictive of its final placement. The agent will therefore underestimate the value of each action in proportion to the amount of rewards available in the middle of the board.

Last, a widely recognized aspect of episodic memory is that retrieval modifies the existing memory traces through the process of consolidation. Repeated retrievals, in particular, can result in more abstract representations and, in some accounts, to the formation of semantic memory (McClelland et al., 1995). While we acknowledge

the evidence from empirical and modeling work supporting retrieval-induced learning, in TCM-SR, we consider a simplified setting where only real experience (external stimuli presented to the subject) causes learning. Accounting for retrieval-induced learning in TCM-SR would require a number of additional assumptions (e.g., the amount and content of episodic retrieval and, thus, learning happening between the encoding of an experience and the later use of the corresponding memories for decision making) that are outside the scope of the current model. While this may be viewed as a significant limitation, we note that we can view our simulations as describing the first bout of retrieval after the encoding phase—that is, before any retrieval-induced learning takes place. It is also plausible to assume that, in the regime of extensive “real” experience assumed in our simulations, any additional learning induced by retrieval mechanisms is negligible and unlike to modify the learned steady-state SR (which itself can be viewed as a learned abstraction). In either case, we make these assumptions to preserve the simplicity and interpretability of our model.

Method

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 (Figure 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, then it falls diagonally to the right with Probability 1. Similarly, if the ball is at the right edge, then 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 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 that extends a Markov chain by

Table 1
Summary of Notations

Notation	Description
γ	Proportion of previous temporal context used in updating the current context during <i>encoding</i> (equivalently, the time horizon of SR)
$\tilde{\gamma}$	Effective time horizon
ρ	Proportion of previous temporal context used in updating the current context during <i>retrieval</i>
β	Proportion of the retrieved temporal context used in updating the current context during <i>retrieval</i>
p_{stop}	Interruption probability during retrieval
α	Initial learning rate during encoding
α_{mod}	Initial emotionally modulated learning rate during encoding
S_t	Stimulus encountered at time t
R_t	Reward encountered at time t
$q(a)$	Expected reward by performing action a
$\hat{q}(a)$	Estimated reward in expectation by performing action a
\mathbf{T}	One-step transition matrix
T_{ij}	The (i, j) -th entry of \mathbf{T}
\mathbf{M}	Successor representation (SR)
m_{ij}	The (i, j) -th entry of \mathbf{M}
r	One-step reward function
\mathbf{v}	State-value function
\mathbf{c}_t	Temporal context at encoding time step t
\mathbf{c}_r	Temporal context at retrieval time step t
\mathbf{c}_i^N	Retrieved temporal context at retrieval time step i
a	Action
s	Stimulus/state
$\mathbf{x}(s)$	Feature vector of stimulus s
\mathbf{x}_t	Feature vector of stimulus at encoding time step t
\mathbf{x}_i	Feature vector of stimulus at retrieval time step i
T	Total encoding time
N	Total number of samples retrieved

adding a reward to each state. Unlike in a 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 article to remain consistent with the decision-making literature.

The task is formalized by a five-tuple $\langle S, A, P, R, \gamma \rangle$. $S = \{s_1, s_2, \dots, s_{|S|}\}$ denotes the set of states, and A denotes the set of actions corresponding to each state in the top row, that is, $A = \{s_{a=1}, s_{a=2}, \dots, s_{a=|A|}\}$. $P: S \mapsto S$ is the Markov transition function that defines the probability distribution $P(s'|s)$ of transitioning from state s to state s' . $R: S \mapsto \mathbb{R}$ is the reward function $R(s)$ 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 the distant future (Table 1).

The goal of the agent is to choose the action that maximizes the cumulative discounted return $G = \sum_{t=1}^{\infty} \gamma^{t-1} R(S_t)$, 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 $P(S_{t+1} = s' | S_t = s) = P(s'|s)$. This gives

rise to a “trajectory” given by $S_1, R_1, S_2, R_2, S_3, R_3, \dots, S_H, R_H$, where $H = \frac{|S|}{|A|}$ 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 : $v(s) = \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k R_{t+k} | S_t = s]$. 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: $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t | A = a]$ and can also be defined strictly in terms of states: We refer to $q(a)$ as “action value.”

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 TD algorithm, for example, iteratively updates the agent’s estimate $Q(a)$ as $Q(a) \leftarrow Q(a) + \alpha(R_1 + \gamma v(S_2) - Q(a))$ whenever action a is performed. In model-based methods, in contrast, the agent uses a model of the world (i.e., an estimate of P and R) to estimate $q(a)$. If both P and R are perfectly known, the agent can generate a plausible trajectory $S_1, R_1, S_2, R_2, S_3, R_3, \dots, S_T, R_T$ resulting from a , where S_1 is defined by the action selected, $S_{i+1} \sim P(\cdot | S_i)$ and $R_i = R(S_i)$. 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 simulated trajectory can then be used as a Monte Carlo estimate of the action value, that is, $Q(a) = \sum_{i=1}^H \gamma^{i-1} R_i$.

SR

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

$$\begin{aligned} \mathbf{v} &= \mathbf{T}^1 \mathbf{r} + \gamma^1 \mathbf{T}^2 \mathbf{r} + \gamma^2 \mathbf{T}^3 \mathbf{r} + \dots \\ &= \left(\sum_{k=0}^{\infty} \gamma^k \mathbf{T}^k \right) \mathbf{r} \\ &= (\mathbf{I} - \gamma \mathbf{T})^{-1} \mathbf{T} \mathbf{r}. \end{aligned} \quad (5)$$

The matrix $(\mathbf{I} - \gamma \mathbf{T})^{-1} \mathbf{T}$ is the SR, denoted by \mathbf{M} (Dayan, 1993). The (i, j) -th entry m_j 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. The value function can thus be estimated as $\mathbf{v} = \mathbf{M} \mathbf{r}$.

We note that our definition differs from the more traditional $(\mathbf{I} - \gamma \mathbf{T})^{-1}$. 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.

TCM

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, 2002a).

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 (Figure 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 to quantify such effect. The lag conditional response probability 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 preexperimental 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 conditional response probability to be asymmetrical, with higher probability to recall subsequent stimuli than preceding ones (Figure 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 et al., 2008; Howard & Kahana, 1999), suggesting maintenance of temporal contexts over multiple timescales and entities.

In summary, the temporal context and its evolution dynamics in TCM provide 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^{IN} denote the preexperimental context at t , both as column vectors. Additionally, let $\mathbf{x}(S_t)$ be the feature representation of the stimulus encoded at time t , for example, 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 , that is, we use $t \in \{1, 2, \dots, T\}$ to index encoding time and $i \in \{1, 2, \dots, N\}$ to index retrieval steps. Additionally, TCM achieves associative learning via a context-to-stimulus matrix \mathbf{M}^{CS} and a stimulus-to-context matrix \mathbf{M}^{SC} . The learning and update rules are summarized in Table 2.

TCM posits that when a stimulus \mathbf{s}_t is experienced either in encoding or retrieval, the following sequence of events take place in the following order: First, presenting \mathbf{x}_t evokes its associated context \mathbf{c}_t^{IN} via the stimulus-to-context matrix according to Equation 7. If the stimulus is unique, then \mathbf{c}_t^{IN} is equivalent to the stimulus’ preexperimental context; if the stimulus is repeated, \mathbf{c}_t^{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 Equation 8. Note that ρ and β are chosen so that \mathbf{c}_t remains a unit vector. Last, \mathbf{M}^{CS} and \mathbf{M}^{SC} are updated as needed and the above sequence ensues. If Hebbian learning is assumed, for instance, \mathbf{M}^{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 Equation 11.

At the beginning of each new experiment, \mathbf{M}^{CS} may be reset for simplicity. Howard and Kahana (2002a) derived a learning rule for the stimulus-to-context matrix \mathbf{M}^{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 article.

The SR in the TCM

Consider the special case where \mathbf{M}^{SC} is the identity matrix. It follows that $\mathbf{c}_t^{\text{IN}} = \mathbf{x}_t$, that is, the associated context of a stimulus is exactly its corresponding features. Thus, Equation 8 is reduced to Equation 1.

Assuming one-hot encoding, we can use the delta function to map each retrieved \mathbf{x}_i to an abstract state vector indexed by time. Thus the j -th entry of \mathbf{c}_{t+1} satisfies

Table 2
Summary of the Temporal Context Model

Name	Expression	Equation number
Context-to-feature matrix	$\mathbf{M}^{\text{CS}} = \sum_t \mathbf{x}_t \mathbf{c}_t^{\text{T}}$	(6)
Input context	$\mathbf{c}_t^{\text{IN}} = \mathbf{M}^{\text{SC}} \mathbf{x}_t$	(7)
Context update	$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{c}_t^{\text{IN}}$	(8)

Note. CS = context-to-stimulus matrix; IN = incoming; SC = stimulus-to-context matrix.

$$c^{(j)}_{t+1} = \begin{cases} \rho c^{(j)}_t & \text{if } S_t \neq s_j \\ \rho c^{(j)}_t + \beta & \text{if } S_t = s_j \end{cases}, \quad (9)$$

which is analogous to the eligibility trace over the sequence of feature vectors. Consider also the following temporal difference learning algorithm for learning the SR matrix:

$$\mathbf{M}_{t+1} \leftarrow \mathbf{M}_t + \alpha \mathbf{c}_{t+1} (\mathbf{x}_{t+1}^\top + \gamma \mathbf{x}_t^\top \mathbf{M}_t - \mathbf{x}_t^\top \mathbf{M}_t), \quad (10)$$

where we have used the temporal context vector \mathbf{c}_t to represent the eligibility trace.

Gershman et al. (2012) showed that if stimuli are presented only once and $\rho = \lambda\gamma$, the context-to-stimulus matrix learned according to Equation 6 is equivalent to the (transpose of the) SR matrix learned via temporal difference, that is, $\mathbf{M}^{\text{CS}} = \mathbf{M}^\top$. Indeed, as long as the SR is initialized with zeros, the second and third terms inside the parentheses in Equation 11 are zero, leading to $\mathbf{M}_{t+1} \leftarrow \mathbf{M}_t + \alpha \mathbf{c}_{t+1} \mathbf{x}_{t+1}^\top$ —or, equivalently, $\mathbf{M}_{t+1}^{\text{CS}} \leftarrow \mathbf{M}_t^{\text{CS}} + \alpha \mathbf{x}_{t+1} \mathbf{c}_{t+1}^\top$, which is exactly the learning rule described in Equation 6.

Following Gershman et al. (2012), we assume this correspondence between the SR and the context-to-stimulus matrix to also hold when stimuli are presented more than once, assuming a temporal difference update rule. This leads to the following learning rule for \mathbf{M}^{CS}

$$\mathbf{M}_{t+1}^{\text{CS}} \leftarrow \mathbf{M}_t^{\text{CS}} + \alpha (\mathbf{x}_{t+1} + \gamma \mathbf{M}_t^{\text{CS}} \mathbf{x}_{t+1} - \mathbf{M}_t^{\text{CS}} \mathbf{x}_t) \mathbf{c}_{t+1}^\top. \quad (11)$$

Value Computation in TCM-SR

Setting \mathbf{M}^{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 state s_k , $v(s_k)$. Let $\mathbf{m}_{k,*}$ denote the row in \mathbf{M}_γ corresponding to s_k and $m_{k,j}$ the entry corresponding to the expected discounted number of future visitations to s_j starting from state s_k . 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_k can consequently be rewritten as

$$v(s_k) = \mathbf{m}_{k,*} \mathbf{r} = \sum_j m_{k,j} r(s_j). \quad (12)$$

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

$$\tilde{v}(s_k) = \frac{1}{N(1-\gamma)} \sum_{i=1}^N r(S_i). \quad (13)$$

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

However, in general, TCM draws are not i.i.d., because a nonzero β would cause the temporal context to drift toward 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 toward successors of the previous sample. In particular, \mathbf{x}_t may be obtained via

$$\mathbf{x}_t \sim \frac{1}{Z} \mathbf{M}^{\text{CS}} \mathbf{c}_t, \quad (14)$$

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

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 SR. 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 nonzero 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 $p(1-\gamma)m_{S,S'}^\pi$. The 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., 2018) or faster decay of less salient outcomes (Zhou et al., 2020). From the RL perspective, both accounts effectively lead to overrepresentation 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 that overrepresentation 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 overrepresentation creates a bias in estimation, fewer

samples are required for a confident estimate. We give a formal derivation in the next section.

Last, 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., Lehnert et al., 2017; Momennejad et al., 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}^{SC} to be the identity matrix—that is, the context associated with a state is exactly its feature vector. Alternatively, \mathbf{M}^{SC} could encode some backward transitions such as the transpose of \mathbf{M}^{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 . For example, 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, that is, $\mathbf{c}_0 = \mathbf{x}_0$. 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. For example, using \mathbf{M} as a shorthand for \mathbf{M}^π .

Independent Samples From Memory Yield Unbiased Value Estimates

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

Lemma 1. Recall that 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

$$\mathbb{P}(S_i) = (1 - \gamma)\mathbf{x}'_0\mathbf{M}\mathbf{x}_i. \quad (15)$$

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

$$\begin{aligned} \mathbb{P}(S_1) &= \frac{1}{1/(1 - \gamma)}(\mathbf{M}^{\text{CS}}(\rho\mathbf{c}_0 + \beta\mathbf{M}^{\text{SC}}\mathbf{x}_0))'\mathbf{x}_1 \quad (\text{Equation 9}) \\ &= (1 - \gamma)(\mathbf{M}^{\text{CS}}\mathbf{x}_0)'\mathbf{x}_1 \\ &= (1 - \gamma)\mathbf{x}'_0\mathbf{M}\mathbf{x}_1. \end{aligned} \quad (16)$$

Now consider arbitrary time step $i > 1$. By Equation 8, $\mathbf{c}_i = \mathbf{c}_{i-1} = \dots = \mathbf{c}_0 = \mathbf{x}_0$. 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, \dots, S_N , the value of state S_0 , $v(S_0)$, satisfies

$$v(S_0) = \frac{1}{N(1 - \gamma)}\mathbb{E}\left[\sum_{i=1}^N \mathbf{r}(S_i)\right]. \quad (17)$$

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

$$\begin{aligned} \mathbb{E}[\mathbf{r}(S_i)] &= \sum_{k=1}^{|S|} \mathbb{P}(S_i = s_k)\mathbf{r}(s_k) \\ &= \sum_{k=1}^{|S|} (1 - \gamma)\mathbf{x}'_0\mathbf{M}\mathbf{x}(s_k)\mathbf{r}(s_k) \quad (\text{Lemma 1}) \\ &= (1 - \gamma)\mathbf{x}'_0\mathbf{M}\sum_{k=1}^{|S|} \mathbf{x}(s_k)\mathbf{r}(s_k) \\ &= (1 - \gamma)\mathbf{x}'_0\mathbf{M}\mathbf{r} \\ &= (1 - \gamma)\mathbf{x}'_0\mathbf{v}. \end{aligned} \quad (18)$$

By linearity of expectation,

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^N \mathbf{r}(S_i)\right] &= \sum_{i=1}^N \mathbb{E}[\mathbf{r}(S_i)] \\ &= N(1 - \gamma)\mathbf{x}'_0\mathbf{v} \\ &= N(1 - \gamma)v(S_0). \end{aligned} \quad (19)$$

Rearranging the terms, we have

$$v(S_0) = \frac{1}{N(1 - \gamma)}\mathbb{E}\left[\sum_{i=1}^N \mathbf{r}(S_i)\right]. \quad (20)$$

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 $\rho = 0$, $\beta = 1$, $\mathbf{M}^{\text{CS}} = \mathbf{M}'$, and $\mathbf{M}^{\text{SC}} = \mathbf{I}_{|S|}$, 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

$$\mathbb{P}(S_i) = (1 - \gamma)^i\mathbf{x}'_0\mathbf{M}^i\mathbf{x}_i. \quad (21)$$

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 $i > 0$, $\mathbb{P}(S_i) = (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{x}_i$.

$$\begin{aligned} \mathbb{P}(S_{i+1}|S_i) &= \frac{1}{Z} (\mathbf{M}^{\text{CS}}(\rho \mathbf{c}_i + \beta \mathbf{M}^{\text{SC}} \mathbf{x}_i))' \mathbf{x}_{i+1} \\ &= \frac{1}{Z} (\mathbf{M}^{\text{CS}}(\mathbf{M}^{\text{SC}} \mathbf{x}_i))' \mathbf{x}_{i+1} \\ &= \frac{1}{Z} \mathbf{x}'_i \mathbf{M} \mathbf{x}_{i+1}, \end{aligned} \quad (22)$$

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

$$\begin{aligned} \mathbb{P}(S_{i+1}) &= \sum_{s_k} \mathbb{P}(S_i = s_k) \mathbb{P}(S_{i+1}|S_i = s_k) \\ &= \sum_{s_k} (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{x}(s_k) \cdot (1 - \gamma) \mathbf{x}(s_k)' \mathbf{M} \mathbf{x}_{i+1} \\ &= (1 - \gamma)^{i+1} \mathbf{x}'_0 \mathbf{M}^i \sum_{s_k} (\mathbf{x}(s_k) \mathbf{x}(s_k)') \mathbf{M} \mathbf{x}_{i+1} \\ &= (1 - \gamma)^{i+1} \mathbf{x}'_0 \mathbf{M}^{i+1} \mathbf{x}_{i+1}. \end{aligned} \quad (23)$$

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

$$v_{\tilde{\gamma}=1}(S_0) = \frac{1}{(1 - \gamma)} \mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right]. \quad (24)$$

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

$$\begin{aligned} \mathbb{E}[\mathbf{r}(S_i)] &= \sum_{k=1}^{|\mathcal{S}|} P(S_i = s_k) \mathbf{r}(s_k) \\ &= \sum_{k=1}^{|\mathcal{S}|} (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{x}(s_k) \mathbf{r}(s_k) \quad (\text{Lemma 3}) \\ &= (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \sum_{k=1}^{|\mathcal{S}|} \mathbf{x}(s_k) \mathbf{r}(s_k) \\ &= (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{r}. \end{aligned} \quad (25)$$

By linearity of expectation,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right] &= \sum_{i=1}^{\infty} \mathbb{E}[\mathbf{r}(S_i)] = \sum_{i=1}^{\infty} (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{r} \\ &= (1 - \gamma) \mathbf{x}'_0 (\mathbf{T} + \gamma \mathbf{T}^2 + \gamma^2 \mathbf{T}^3 + \dots) \mathbf{r} \\ &\quad + (1 - \gamma)^2 \mathbf{x}'_0 (\mathbf{T}^2 + 2\gamma \mathbf{T}^3 + 3\gamma^2 \mathbf{T}^4 + \dots) \mathbf{r} \\ &\quad + \dots = (1 - \gamma) \mathbf{x}'_0 \mathbf{T} \mathbf{r} + (\gamma(1 - \gamma) \\ &\quad + (1 - \gamma)^2) \mathbf{x}'_0 \mathbf{T}^2 \mathbf{r} + (\gamma^2(1 - \gamma) + 2\gamma(1 - \gamma)^2 \\ &\quad + (1 - \gamma)^3) \mathbf{x}'_0 \mathbf{T}^3 \mathbf{r} + \dots = (1 - \gamma) \mathbf{x}'_0 \mathbf{T} \mathbf{r} \\ &\quad + (1 - \gamma) \mathbf{x}'_0 \mathbf{T}^2 \mathbf{r} + (1 - \gamma) \mathbf{x}'_0 \mathbf{T}^3 \mathbf{r} + \dots \\ &= (1 - \gamma) \mathbf{x}'_0 (\mathbf{T} \mathbf{r} + \mathbf{T}^2 \mathbf{r} + \mathbf{T}^3 \mathbf{r} + \dots) \\ &= (1 - \gamma) \mathbf{x}'_0 \mathbf{v}_{\tilde{\gamma}=1}. \end{aligned} \quad (26)$$

Rearranging the terms, we have

$$v_{\tilde{\gamma}=1}(S_0) = \frac{1}{(1 - \gamma)} \mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right]. \quad (27)$$

Now, consider a fixed probability p_{stop} that interrupts the sampling process of the generalized rollout regime at any moment, that is, 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 predefined 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 4.1. Given $\rho = 0$, $\beta = 1$, $p_{\text{stop}} \in [0, 1]$, and arbitrary encoding γ , the effective discount factor $\tilde{\gamma}$ of the estimated value satisfies $\tilde{\gamma} = \gamma p_{\text{stop}} - p_{\text{stop}} + 1$.

Proof. Consider retrieval at some time i . Let A_i denote the event that the sampling process is not 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

$$\begin{aligned} \mathbb{E}[\mathbf{r}(S_i)] &= \mathbb{P}(A_i) (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{r} + \mathbb{P}(A_i^c) \cdot 0 \\ &= (1 - p_{\text{stop}})^{i-1} (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{r}. \end{aligned} \quad (28)$$

By linearity of expectation,

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right] &= \sum_{i=1}^{\infty} \mathbb{E}[\mathbf{r}(S_i)] = \sum_{i=1}^{\infty} (1 - p_{\text{stop}})^{i-1} (1 - \gamma)^i \mathbf{x}'_0 \mathbf{M}^i \mathbf{r} \\ &= (1 - \gamma) \mathbf{x}'_0 (\mathbf{T} + \gamma \mathbf{T}^2 + \gamma^2 \mathbf{T}^3 + \dots) \mathbf{r} \\ &\quad + (1 - p_{\text{stop}}) (1 - \gamma)^2 \mathbf{x}'_0 (\mathbf{T}^2 + 2\gamma \mathbf{T}^3 + 3\gamma^2 \mathbf{T}^4 \\ &\quad + \dots) \mathbf{r} + \dots = (1 - \gamma) \mathbf{x}'_0 \mathbf{T} \mathbf{r} + (\gamma(1 - \gamma) \\ &\quad + (1 - p_{\text{stop}})(1 - \gamma)^2) \mathbf{x}'_0 \mathbf{T}^2 \mathbf{r} + (\gamma^2(1 - \gamma) \\ &\quad + 2(1 - p_{\text{stop}})\gamma(1 - \gamma)^2 \\ &\quad + (1 - p_{\text{stop}})^2(1 - \gamma)^3) \mathbf{x}'_0 \mathbf{T}^3 \mathbf{r} \\ &\quad + \dots = (1 - \gamma) \mathbf{x}'_0 \mathbf{T} \mathbf{r} \\ &\quad + (1 - \gamma)(\gamma p_{\text{stop}} - p_{\text{stop}} + 1) \mathbf{x}'_0 \mathbf{T}^2 \mathbf{r} \\ &\quad + (1 - \gamma)(\gamma p_{\text{stop}} - p_{\text{stop}} + 1)^2 \mathbf{x}'_0 \mathbf{T}^3 \mathbf{r} + \dots \\ &= (1 - \gamma) \mathbf{x}'_0 (\mathbf{T} \mathbf{r} + (\gamma p_{\text{stop}} - p_{\text{stop}} + 1) \mathbf{T}^2 \mathbf{r} \\ &\quad + (\gamma p_{\text{stop}} - p_{\text{stop}} + 1)^2 \mathbf{T}^3 \mathbf{r} + \dots). \end{aligned} \quad (29)$$

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

$$\mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right] = (1 - \gamma) \mathbf{x}'_0 \mathbf{v}_{\tilde{\gamma} = \gamma p_{\text{stop}} - p_{\text{stop}} + 1}. \quad (30)$$

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 Suggest an Intermediate Regime

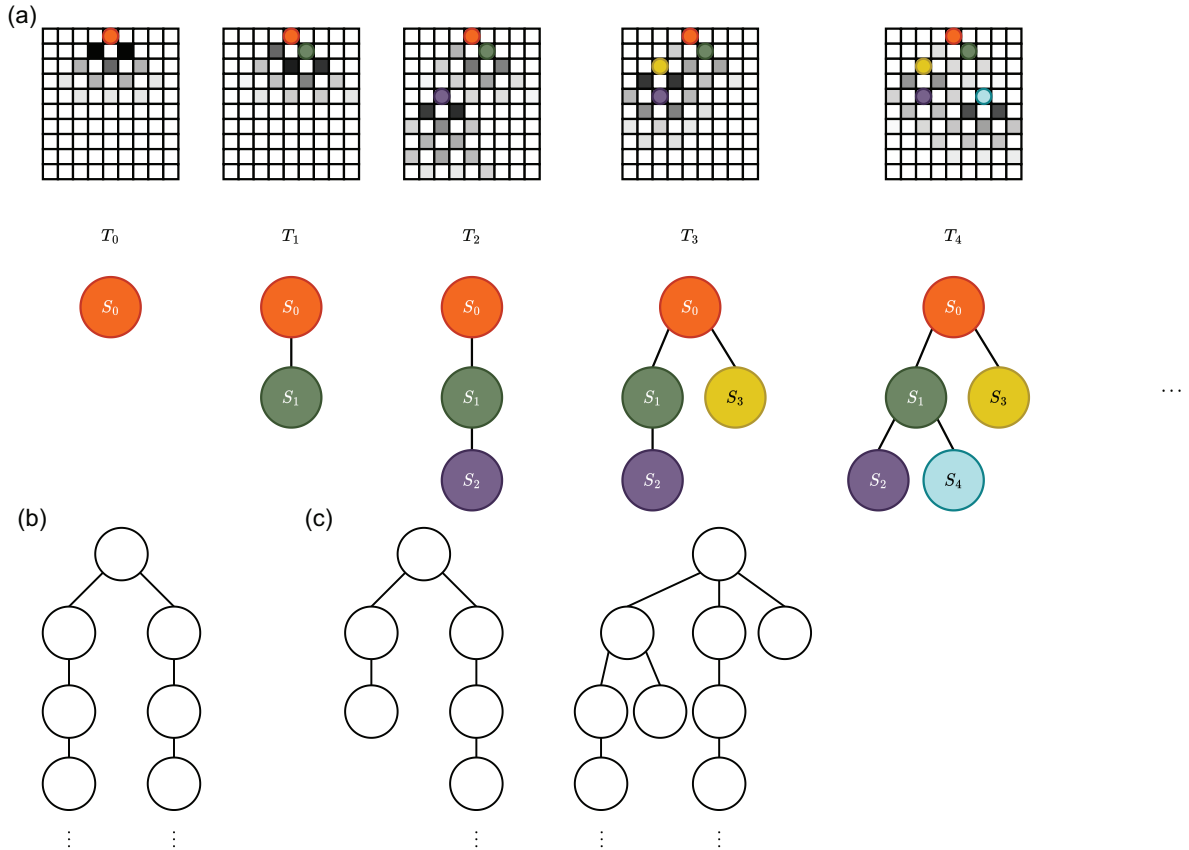
Observe that the sequentially obtained samples can be conceptualized as a random tree with root at S_0 (Figure 7a). At

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 mixture 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 Equation 8. In general, for any $i > j \geq 0$ we have

$$\mathbb{P}(pa(i) = j) = \begin{cases} \rho^{i-1} & \text{if } j = 0 \\ \rho^{i-j-1}\beta & \text{if } j > 0. \end{cases} \quad (31)$$

Note that the construction necessarily results in a tree because of the sequential nature of the sampling process, namely, a newly

Figure 7
Visualization of the Intermediate Sampling Regime



Note. (a) A possible sequence of samples obtained using the intermediate sampling regime ($\rho, \beta > 0$). The tree starts with a single root node in orange representing the state action to be evaluated. The other colored circles on the Plinko board indicate samples drawn, and they correspond to the nodes of the same color in the constructed tree. An edge is drawn between a pair of nodes (samples) if the child (at a lower level of the tree) is drawn from the successor representation-defined distribution defined at the parent (at a higher level of the tree). Grayscale colors indicate the sampling probabilities before the next sampling step. (b) Schematics of an ideal tree (in expectation) resulted from infinite sampling under the intermediate sampling regime when $\rho = \beta = 0.5$. For generality, nodes are not indexed. (c) Schematics of a few nonideal trees resulted from finite sampling under the intermediate sampling regime when $\rho = \beta = 0.5$. They both have “stubs” or short branches counting from the root node. T = Plinko boards label; S = sequence of states.

All rights, including for text and data mining, AI training, and similar technologies, are reserved.

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 5. Assume $\rho + \beta = 1$, $\rho, \beta > 0$. As $N \rightarrow \infty$, 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

$$\lim_{N \rightarrow \infty} \mathbb{E}[d_N(i)] = \begin{cases} 1/(1 - \rho) & \text{if } i = 0 \\ 1 & \text{if } i > 0. \end{cases} \quad (32)$$

An illustration of such a tree structure in expectation is shown in Figure 7b.

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, $\lim_{N \rightarrow \infty} \mathbb{E}[d_N(0)] = 1/(1 - \rho)$, $\lim_{N \rightarrow \infty} \mathbb{E}[d_N(j)] = 1$ for all positive j .

Corollary 5.1. 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 5 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 6. Given $\rho + \beta = 1$, $\rho, \beta > 0$,

$$v_{\gamma=1}(S_0) = \frac{\beta}{(1 - \gamma)} \mathbb{E} \left[\sum_{i=1}^{\infty} \mathbf{r}(S_i) \right]. \quad (33)$$

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 described above. Specifically, as $N \rightarrow \infty$, $\rho = 1$ 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_∞ , that is, sum along paths and average across paths from Theorem 2 and 4, respectively. The result then directly follows from Lemma 5 noting $1 - \rho = \beta$.

Lemma 7. Given $\rho + \beta = 1$, $\rho, \beta > 0$, $N < \infty$, there is a nonzero 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

$$\begin{aligned} \mathbb{P}(l_2) &= \prod_{i=3}^N (1 - \mathbb{P}(pa(i) = 2)) \\ \Rightarrow \log \mathbb{P}(l_2) &= \sum_{i=3}^N \log(1 - \rho^{i-3} \beta) = \sum_{i=0}^{N-3} \log(1 - \rho^i \beta) \approx - \sum_{i=0}^{N-3} \rho^i \beta < 0 \\ \Rightarrow \mathbb{P}(l_2) &> 0. \end{aligned} \quad (34)$$

A few examples of such possible tree structures are shown in Figure 7c.

Proposition 7.1. Given $\rho + \beta = 1$, $\rho, \beta > 0$, $N < \infty$, the value estimator in Theorem 6 is biased if all rewards have the same sign (i.e., if they are either all positive or all negative).

Proof. Lemma 7 implies that any random tree resulting from the finite sampling process likely has a short path (a ‘‘stub’’). According to Theorem 6, the estimate of the value of the root node S_0 (in the case of Plinko, a location to drop the ball), $\hat{v}(S_0)$, is unbiased only when all paths starting from the root is *infinite* in length. Thus, assuming all rewards are positive, the sum of rewards along a short (or, rather, finite) path is at most as big as the sum of rewards along an infinite path; when positive rewards are relatively ample, the difference is likely larger. For an arbitrary random tree constructed from N samples (N is finite), denote the total number of distinct paths starting from its root node as P and the number of nodes along path j as N_j such that $N = \sum_{j=1}^P N_j$. Further, let $S_i^{(j)}$ be the i -th sample along path j . Since the state-action value is estimated by averaging total rewards of each path starting from the root (Theorem 6), if one of the path underestimates, the overall estimate $\hat{v}(S_0) = \frac{\beta}{(1-\gamma)} \sum_{j=1}^P \sum_{i=1}^{N_j} \mathbf{r}(s_i^{(j)})/P$ will also be biased in the same direction. A similar argument can be made to show that if all rewards are negative, $\hat{v}(S_0)$ will overestimate $v(S_0)$.

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 downward 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. This intermediate sampling regime that is readily implemented by TCM-SR is also closely related to common random numbers and the vine sampling scheme (Schulman et al., 2015), which offers additional computational and behavioral advantage by reducing variance in value estimation than generalized rollouts given a fixed number of samples.

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 nonsalient stimuli to learn \mathbf{M}^{CS} . For clarity, a state s either contains nothing (i.e., $R(s) = 0$) or a small reward ($R(s) = 1$). All else being equal, the resultant, emotionally modulated TCM-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 $\bar{\mathbf{M}}' \neq \mathbf{M}'$. To compute an estimation of expectation, it needs *importance sampling* to translate distributions of \mathbf{M} to $\bar{\mathbf{M}}$.

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

$$w_{S_i} = \frac{P(S_i)}{Q(S_i)} = \frac{m_{S_0, S_i}}{\bar{m}_{S_0, S_i}}. \quad (35)$$

While exact importance weights are intractable, it has been suggested that people readily approximate them (Lieder et al., 2018; Schultz et al., 1997). As with other components of this algorithmic TCM-SR theory (e.g., Hebbian- vs. TD-learning rules of the SR), we are not committed to any specific implementation as long as they give rise to the same representation so as to limit assumptions beyond well-studied features of episodic memory. The decision weights could be implemented by an implicit process as in Lieder et al. (2018) or a more explicit self-correcting process. Nonetheless, since the goal of our article is to rationally predict how optimal decisions may be made given certain episodic memory constraints, we choose to assume theoretically “perfect” importance sampling and examine the resultant behavior.

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

$$\begin{aligned} \mathbb{E}[\mathbf{r}(S_i)] &= \sum_{k=1}^{|\mathcal{S}|} P(S_i = s_k) \mathbf{r}(s_k) \\ &= \sum_{k=1}^{|\mathcal{S}|} Q(S_i = s_k) \frac{P(S_i = s_k)}{Q(S_i = s_k)} \mathbf{r}(s_k) \\ &= \sum_{k=1}^{|\mathcal{S}|} w_{S_i} Q(S_i = s_k) \mathbf{r}(s_k). \end{aligned} \quad (36)$$

Theorem 2 can be then applied to estimate a specific state value. In general, $\tilde{\mathbf{v}}$ is biased if N is finite. Specifically, $\tilde{\mathbf{v}}$ demonstrates a bias-variance trade-off, such that extreme events are overrepresented 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 distribution of S_i is $Q(S_i) = (1 - \gamma)^i \mathbf{x}'_0 \tilde{\mathbf{M}}^i \mathbf{x}_i$. Denote the (S_0, S_i) -th entry of \mathbf{M}^i as $(\mathbf{M}^i)_{S_0, S_i}$ and that of $\tilde{\mathbf{M}}^i$ as $(\tilde{\mathbf{M}}^i)_{S_0, S_i}$. To correct for the difference between P and Q , each sample S_i should be reweighed by

$$w_{S_i} = \frac{P(S_i)}{Q(S_i)} = \frac{(\mathbf{M}^i)_{S_0, S_i}}{(\tilde{\mathbf{M}}^i)_{S_0, S_i}}. \quad (37)$$

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 10×9 (i.e., $H = 10$, $|\mathcal{S}| = 90$, excluding the absorbing state that is outside the main board). Binary rewards were randomly placed in locations between Rows 1 and 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 $\gamma = 0$ (Figure 2a–c) and $\gamma = 0.5$ (Figure 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}^{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 1,000 (independent) samples per trial (i.e., $N = 1,000$). At least one reward was placed within the agent’s temporal horizon. For example, given $\gamma = 0$, Row 2 contained at least one reward. The sampling distributions over rows (Figure 2b and e) reflect trial averages if starting from the top-center state (marked with an orange circle in Figure 2a and 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 one, five, 10, and 20 binary rewards accessible from either dropping location (Figure 2c and 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

$$\text{pmr} = \frac{v(S_{\text{chosen}})}{v(S^*)}, \quad (38)$$

where S_{chosen} is the state selected by the deterministic policy, S^* is the state with the highest expected total return, and $v(\cdot) : \mathcal{S} \mapsto \mathbb{R}$ is the state value function. Note an optimal choice implies $\text{pmr} = 1$. Figure 2c and f shows 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 $\gamma = 0$ (Figure 3a–d) and $\gamma = 0.5$ (Figure 3e–h) during encoding. For each γ , simulation were repeated using three different probabilities of interruption $p = .05$, $p = .5$, and $p = 1$, resulting in three different effective discount factors $\tilde{\gamma}$ s for each underlying true γ at retrieval (Figure 3b and 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 an orange circle in Figure 3a and e) and ended if either the ball hit the bottom of the board or the sampling process terminated due to the nonzero interruption probability. The stimulus-to-context associative matrix \mathbf{M}^{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 (Figure 3b and f) reflect averages across 1,000 trials per experiment if starting from the top-center state. The implied contiguity curves (Figure 3d and 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 the 5,000 trials and averaging across trials. One hundred 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 one, five, 10, and 20 binary rewards accessible from either dropping location (Figure 3c and g). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Figure 3c and g shows 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 $\gamma = 0$ (Figure 4a–c) and $\gamma = 0.5$ (Figure 4d–f) during encoding. For each γ , simulations were repeated using three different probabilities of interruption $p = .05$, $p = .5$, and $p = 1$, resulting in three different effective discount factors $\tilde{\gamma}$ s for each underlying true γ at retrieval (Figure 4b and 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 an orange circle in Figure 4a and d) and ended when the ball hit the bottom of the board or the sampling process terminated due to the nonzero interruption probability. The stimulus-to-context associative matrix \mathbf{M}^{SC} was equal to the identity matrix $\mathbf{I}_{|S|}$.

A total of 100 experiments were conducted for each combination of discount factor and interruption probability. The sampling distributions over rows (Figure 4b and 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 the 5,000 trials and averaging across trials. One hundred games were simulated, and each trial consists of a variable number of correlated samples. The interruption probability is fixed at .05. Simulations were repeated for games with one, five, 10, and 20 binary rewards accessible from either dropping location (Figure 4c and f). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Figure 4c and f shows 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}^{SC} was equal to the identity matrix $\mathbf{I}_{|S|}$.

The intermediate and converged SR matrices of the top-center state (four panels to the left in Figure 5a and 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 four complete trajectories. An intermediate SR was computed after observation of each complete trajectory. The unmodulated and modulated learning rates were initialized to .01 and .5, respectively, that is, $\alpha_0 = .01$, $\alpha_{\text{mod}, 0} = .5$. Both the unmodulated agent (Figure 5a) and the modulated agent (Figure 5b) were trained using the same exponential decay schedule such that the learning rates upon observing trajectory t was defined as

$$\alpha_t = \alpha_0 \times e^{-kt}, \quad (39)$$

$$\alpha_{\text{mod}, t} = \alpha_{\text{mod}, 0} \times e^{-kt}, \quad (40)$$

where decay rate $k = 0.001$. In both cases, the SR converged after observing 10,000 trajectories.

We used 100 random experiments (games) and drew 1,000 samples from the TD-learned SR after one observation (trajectory) in each experiment to compute the average fraction of samples that contained a reward (Figure 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 (Figure 5e and 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 1,000 i.i.d. samples and 50 trials. Simulations were repeated for games with one, five, 10, and 20 binary rewards accessible from either dropping location (Figure 5c). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Figure 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}^{SC} was equal to the SR matrix \mathbf{M} . Simulations used $\gamma = 0.5$ during encoding and three different interruption probabilities $p = .2$, $p = .5$, $p = 1$, resulting in three different effective discount factors $\tilde{\gamma}$ s at retrieval (Figure 6c and d). Each simulation consisted of 500 experiments and 1,000 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; Figure 6c and d).

The simulation code is available at <https://github.com/corxyz/tcm-sr>. This study was not preregistered.

References

- Aka, A. & Bhatia, S. (2021). What I like is what I remember: Memory modulation and preferential choice. *Journal of Experimental Psychology: General*, 150(10), 2175–2184. <https://doi.org/10.1037/xge0001034>
- Badre, D., Lebrecht, S., Pagliaccio, D., Long, N., & Scimeca, J. (2014). Ventral striatum and the evaluation of memory retrieval strategies. *Journal of Cognitive Neuroscience*, 26(9), 1928–1948. https://doi.org/10.1162/jocn_a_00596
- Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, 8, Article e46080. <https://doi.org/10.7554/eLife.46080>
- Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., & Hassabis, D. (2016). *Model-free episodic control*. arXiv. <https://doi.org/10.48550/arXiv.1606.04460>
- 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(12), Article 1003387. <https://doi.org/10.1371/journal.pcbi.1003387>
- Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, 8, Article 15958. <https://doi.org/10.1038/ncomms15958>
- Bornstein, A. M. & Norman, K. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, 20, 997–1003. <https://doi.org/10.1038/nn.4573>
- Braun, E. K., Elliott, W. G., & Shohamy, D. (2018). Retroactive and graded prioritization of memory by reward. *Nature Communications*, 9, Article 4886. <https://doi.org/10.1038/s41467-018-07280-0>
- Brea, J., Gaál, A. T., Urbanczik, R., & Senn, W. (2016). Prospective coding by spiking neurons. *PLOS Computational Biology*, 12(6), Article e1005003. <https://doi.org/10.1371/journal.pcbi.1005003>
- Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories are constructed from experience. *Hippocampus*, 29(3), 162–183. <https://doi.org/10.1002/hipo.23074>
- Cohen, R. T. & Kahana, M. J. (2019). *Retrieved-context theory of memory in emotional disorders*. bioRxiv. <https://doi.org/10.1101/817486>
- Cohen, R. T. & Kahana, M. J. (2022). A memory-based theory of emotional disorders. *Psychological Review*, 129(4), 742–776. <https://doi.org/10.1037/rev0000334>
- Coulom, R. (2006). *Efficient selectivity and backup operators in Monte-Carlo tree search* [Conference session]. Proceedings of the 5th International Conference on Computers and Games.
- Davelaar, E., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, 112(1), 3–42. <https://doi.org/10.1037/0033-295X.112.1.3>
- Daw, N. D. & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655), Article 20130478. <https://doi.org/10.1098/rstb.2013.0478>
- 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. <https://doi.org/10.1038/nn1560>
- Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624. <https://doi.org/10.1162/neco.1993.5.4.613>
- Dolan, R. J. & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80, 312–325. <https://doi.org/10.1016/j.neuron.2013.09.007>
- 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. <https://doi.org/10.1016/j.nlm.2014.04.014>
- Dougherty, M. R. & Harbison, J. I. (2007). Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 1108–117. <https://doi.org/10.1037/0278-7393.33.6.1108>
- Duncan, K. & Shohamy, D. (2016). Memory states influence value-based decisions. *Journal of Experimental Psychology: General*, 145(11), 1420–1426. <https://doi.org/10.1037/xge0000231>
- Eichenbaum, H. (2001). The hippocampus and declarative memory: Cognitive mechanisms and neural codes. *Behavioural Brain Research*, 127(1–2), 199–207. [https://doi.org/10.1016/S0166-4328\(01\)00365-5](https://doi.org/10.1016/S0166-4328(01)00365-5)
- Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, 119(2), 223–71. <https://doi.org/10.1037/a0027371>
- Feinberg, V., Wan, A., Stoica, I., Jordan, M., Gonzalez, J., & Levine, S. (2018). *Model-based value estimation for efficient model-free reinforcement learning*. arXiv. <https://doi.org/10.48550/arXiv.1803.00101>
- 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, Article e17086. <https://doi.org/10.7554/eLife.17086>
- Gershman, S. J. & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68, 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>
- 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(6), 1553–1568. https://doi.org/10.1162/NECO_a_00282
- 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. <https://doi.org/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(7), 1686–1693. <https://doi.org/10.1016/j.neuropsychologia.2009.02.007>
- Gutbrod, K., Kroužel, C., Hofer, H., Müri, R., Perrig, W., & Ptak, R. (2006). Decision-making in amnesia: Do advantageous decisions require conscious knowledge of previous behavioural choices? *Neuropsychologia*, 44(8), 1315–1324. <https://doi.org/10.1016/j.neuropsychologia.2006.01.014>
- Healey, M. K. & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, 123(1), 23–69. <https://doi.org/10.1037/rev0000015>
- Horwath, E. A., Rouhani, N., DuBrow, S., & Murty, V. P. (2023). Value restructures the organization of free recall. *Cognition*, 231, Article 105315. <https://doi.org/10.1016/j.cognition.2022.105315>
- 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), 923–941. <https://doi.org/10.1037/0278-7393.25.4.923>
- Howard, M. W. & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3), 269–299. <https://doi.org/10.1006/jmps.2001.1388>
- Howard, M. W. & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval. *Journal of Memory and Language*, 46(1), 85–98. <https://doi.org/10.1006/jmla.2001.2798>
- Howard, M. W., Shankar, K. H., & Jagadisan, U. K. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science*, 3(1), 48–73. <https://doi.org/10.1111/j.1756-8765.2010.01112.x>
- 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. <https://doi.org/10.3758/PBR.15.1.58>
- Janner, M., Mordatch, I., & Levine, S. (2020). Gamma-models: Generative temporal difference learning for infinite-horizon prediction. In

- H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1724–1735). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2020/file/12ffb0968f2f56e51a59a6beb37b2859-Paper.pdf
- Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24, 103–109. <https://doi.org/10.3758/BF03197276>
- Kahana, M. J., Howard, M., & Polyn, S. (2008). Associative retrieval processes in episodic memory. *Psychology*, 3, 467–490. <https://surface.syr.edu/cgi/viewcontent.cgi?article=1002&context=psy>
- Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLOS Computational Biology*, 7(5), Article e1002055. <https://doi.org/10.1371/journal.pcbi.1002055>
- 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. <https://doi.org/10.1037/a0028681>
- Kumaran, D. & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, 119(3), 573–616. <https://doi.org/10.1037/a0028681>
- Lehnert, L., Tellex, S., & Littman, M. L. (2017). *Advantages and limitations of using successor features for transfer in reinforcement learning*. ArXiv. <https://doi.org/10.48550/arXiv.1708.00102>
- Lengyel, M. & Dayan, P. (2007). Hippocampal contributions to control: The third way. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20, pp. 889–896). Curran Associates.
- 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. <https://doi.org/10.1037/rev0000074>
- 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), Article eabf1357. <https://doi.org/10.1126/science.abf1357>
- Lohnas, L., Polyn, S., & Kahana, M. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, 122, 337–363. <https://doi.org/10.1037/a0039036>
- Mason, A., Farrell, S., Howard-Jones, P., & Ludwig, C. J. (2017). The role of reward and reward uncertainty in episodic memory. *Journal of Memory and Language*, 96, 62–77. <https://doi.org/10.1016/j.jml.2017.05.003>
- 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, Article e20. <https://doi.org/10.1017/S0140525X15000667>
- Mattar, M. G. & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11), 1609–1617. <https://doi.org/10.1038/s41593-018-0232-z>
- Mattar, M. G. & Lengyel, M. (2022). Planning in the brain. *Neuron*, 110(6), 914–934. <https://doi.org/10.1016/j.neuron.2021.12.018>
- Mattar, M. G., Talmi, D., & Daw, N. D. (2019). *Memory mechanisms predict sampling biases in sequential decision tasks* [Paper presentation]. Reinforcement Learning and Decision Making, Montréal, QC, Canada. <https://ccneuro.org/2018/proceedings/1164.pdf>
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419–457. <https://doi.org/10.1037/0033-295X.102.3.419>
- Miller, J. F., Weidemann, C., & Kahana, M. (2012). Recall termination in free recall. *Memory & Cognition*, 40, 540–550. <https://doi.org/10.3758/s13421-011-0178-9>
- Miller, K. J., Botvinick, M. M., & Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20, 1269–1276. <https://doi.org/10.1038/nn.4613>
- Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, 7, Article e32548. <https://doi.org/10.7554/eLife.32548>
- 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. <https://doi.org/10.1038/s41562-017-0180-8>
- Murdock, B. B. J. & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, 86(2), 263–267. <https://doi.org/10.1037/h0029993>
- Naim, M., Katkov, M., Romani, S., & Tsodyks, M. (2020). Fundamental law of memory recall. *Physical Review Letters*, 124, Article 018101. <https://doi.org/10.1103/PhysRevLett.124.018101>
- Nicholas, J., Daw, N. D., & Shohamy, D. (2022). Uncertainty alters the balance between incremental learning and episodic memory. *Elife*, 11, Article e81679. <https://doi.org/10.7554/eLife.81679>
- O'Keefe, J. & Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press.
- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive Science*, 38(6), 1229–1248. <https://doi.org/10.1111/j.1551-6709.2011.01214.x>
- Palombo, D. J., Di Lascio, J. M., Howard, M. W., & Verfaellie, M. (2019). Medial temporal lobe amnesia is associated with a deficit in recovering temporal context. *Journal of Cognitive Neuroscience*, 31(2), 236–248. https://doi.org/10.1162/jocn_a_01344
- Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place cell sequences depict future paths to remembered goals. *Nature*, 497, 74–79. <https://doi.org/10.1038/nature12112>
- Piray, P. & Daw, N. D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications*, 12(1), Article 4942. <https://doi.org/10.1038/s41467-021-25123-3>
- Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review*, 122(4), 621–647. <https://doi.org/10.1037/a0039413>
- Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009a). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1), 129–156. <https://doi.org/10.1037/a0014420>
- Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009b). Task context and organization in free recall. *Neuropsychologia*, 47(11), 2158–2163. <https://doi.org/10.1016/j.neuropsychologia.2009.02.013>
- Pritzel, A., Uribe, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., & Blundell, C. (2017). *Neural episodic control* [Conference session]. Proceedings of the International Conference on Machine Learning.
- Richie, R., Aka, A., & Bhatia, S. (2022). Free association in a neural network. *Psychological Review*, 130(5), 1360–1382. <https://doi.org/10.1037/rev0000396>
- 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* [Conference session]. Proceedings of the International Conference on Machine Learning.
- Rouhani, N., Norman, K., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(9), 1430–1443. <https://doi.org/10.1037/xlm0000518>
- 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*. <https://osf.io/preprints/psyarxiv/8j9zx>
- 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(9), Article e1005768. <https://doi.org/10.1371/journal.pcbi.1005768>
- 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. <https://doi.org/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. <https://doi.org/10.1016/j.nlm.2013.12.008>
- Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust region policy optimization. In F. Bach & D. Blei, editors, *Proceedings of the 32nd international conference on machine learning*, (Vol. 37, pp. 1889–1897). Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v37/schulman15.html>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275, 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- 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), 893–912. <https://doi.org/10.1037/a0013396>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>
- Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20(11), 1643–1653. <https://doi.org/10.1038/nn.4650>
- Stefanidi, A., Ellis, D. M., & Brewer, G. A. (2018). Free recall dynamics in value-directed remembering. *Journal of Memory and Language*, 100, 18–31. <https://doi.org/10.1016/j.jml.2017.11.004>
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3, 9–44. <https://doi.org/10.1007/BF00115009>
- Sutton, R. S. & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
- Talmi, D., Kavaliauskaitė, 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. <https://doi.org/10.1101/lm.053470.121>
- 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. <https://doi.org/10.1037/rev0000132>
- 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, pp. 1068–1074). MIT Press.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208. <https://doi.org/10.1037/h0061626>
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–402). Academic Press.
- van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., & Ma, W. J. (2021). *Revealing the impact of expertise on human planning with a two-player board game*. PsyArXiv. <https://doi.org/10.31234/osf.io/rhq5j>
- Vikbladh, O. M., Meager, M. R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., & Daw, N. D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron*, 102(3), 683–693. <https://doi.org/10.1016/j.neuron.2019.02.014>
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, 38(4), 599–637. <https://doi.org/10.1111/cogs.12101>
- Wen, T. & Egener, T. (2022). Retrieval context determines whether event boundaries impair or enhance temporal order memory. *Cognition*, 225, Article 105145. <https://doi.org/10.1016/j.cognition.2022.105145>
- Wimmer, G. E. & Shohamy, D. (2011). The striatum and beyond: Contributions of the hippocampus to decision making. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (pp. 281–310). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199600434.003.0013>
- Yonelinas, A. P. & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: An emotional binding account. *Trends in Cognitive Sciences*, 19, 259–267. <https://doi.org/10.1016/j.tics.2015.02.009>
- Zhao, W. J., Richie, R., & Bhatia, S. (2021). Process and content in decisions from memory. *Psychological Review*, 129(1), 73–106. <https://doi.org/10.1037/rev0000318>
- 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. *Cognitive Science*, 42, 1682–1688.

Received February 13, 2023

Revision received June 25, 2024

Accepted July 3, 2024 ■