UNIVERSITY OF CALIFORNIA SAN DIEGO

Adaptive Control with Episodic Mechanisms

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Cognitive Science

by

Corey Zhou

Committee in charge:

Professor Anastasia Kiyonaga, Chair
Professor Marcus Benna
Professor Marcelo Mattar
Professor Eran Mukamel
Professor John Serences

2024

The Dissertation of Corey Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

Table of Contents

List of Figures

List of Tables

Acknowledgements

> You can't connect the dots looking forward; you can only connect them looking backward. So you have to trust that the dots will somehow connect in your future.
>
> *Steve Jobs*

I'd like to think that my scientist identity is largely a result of precious coincidences, made up of numerous people I could not even imagine meeting had I not onboarded this expedition.

I thank David Danks, my undergraduate honors thesis advisor, for the incredible patience, candid honesty, and exceptional incisiveness that I still aspire to achieve one day.

I thank Marcelo Mattar, my graduate advisor, for the tremendous generosity, unconditional support, and relentless inspiration of which I am absolutely honored to have the first-hand experience.

I thank all the brilliant collaborators I have had the pleasure to work with over the years (in chronological order): Angela Yu, Dalin Guo, Nathaniel Daw, Deborah Talmi, Ji-An Li, and Marcus Benna. Each of you taught me a unique lesson about what it means to be a good scientist and do good science.

To the former and current members of the Mattar Lab – thank you for showing me the multitude of ingenious thinking and empathatic camaraderie. Hopefully we cross paths again in the future.

I would also like to thank the rest of my committee – Anastasia Kiyonaga, Eran Mukamel, and John Serences – for their insightful feedback and unwavering trust in me.

In particular, I would like to thank Anastasia Kiyonaga and Jonathan Nicholas for their invaluable advice to develop the experiment paradigms in Chapter 4. Chapter 3 was made possible by illuminating discussions with Tianyi Zheng and Jason Schweinsberg.

Additionally, Chapter 5 benefited greatly from the input of Sebastian Michelmann.

I thank everyone in the Cognitive Science Department for building such an inclusive, caring, and vibrant community.

A shoutout to friends who stayed with me throughout my graduate school journey as well as those who I made along the way (in alphabetical order): Anjie, Dayoung, Irene, Joyce, Kyounghwa, Lu, Lucy, P, Yaqian, Yixiu, Yue, Ziyun, and Zoe.

A special thank you to Rosalind and moresci.sale. You keep me sane.

Finally, I thank my mother and father, for having my back all the time without an ounce of doubt. Thank you for listening to my frustrated rants and over-excited rambling, even though most of the time you probably have no idea what I am actually doing.

Chapter 3, in full, has been accepted for publication of the material as it may appear in Psychological Review. Zhou, Corey Y.; Talmi, Deborah; Daw, Nathaniel D.; Mattar, Marcelo G., American Psychological Association, 2024. The dissertation author was the primary author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Proceedings of the Annual Meeting of the Cognitive Science Society. Zhou, Corey Y.; Talmi, Deborah; Daw, Nathaniel D.; Mattar, Marcelo G., eScholarship University of California, 2024. The dissertation author was the primary author of this paper.

Chapter 5, in full, is currently being prepared for submission for publication of the material. Zhou, Corey Y.; Mattar, Marcelo G. The dissertation author was the primary author of this material.

Vita

VITA

2019	Bachelor of Science in Computer Science, Carnegie Mellon University
2019	Bachelor of Science in Cognitive Science, Carnegie Mellon University
2019-2024	Graduate Teaching Assistant, University of California San Diego
2024	Doctor of Philosophy in Cognitive Science, University of California San Diego

PUBLICATIONS

Ji-An, L.*, **Zhou, C. Y.***, Benna, M. K., Mattar, M. G. (2024). Linking In-context Learning in Transformers to Human Episodic Memory. arXiv (under review).

Zhou, C. Y., Talmi, D., Daw, N. D., Mattar, M. G. (in press). Episodic retrieval for model-based evaluation in sequential decision tasks. *Psychological Review*.

Zhou, C. Y., Talmi, D., Daw, N. D., Mattar, M. G. (2024). Temporally extended decision-making through episodic sampling. In *CogSci 2024*.

Zhou, C. Y., Talmi, D., Daw, N. D., Mattar, M. G. (2022). Computing values through episodic sampling. In *RLDM 2022*.

Zhou, C. Y., Guo, D., Yu, A. J. (2020). Devaluation of Unchosen Options: A Bayesian Account of the Provenance and Maintenance of Overly Optimistic Expectations. In *CogSci 2020*.

Zhou, Y., Danks, D. (2020). Different "Intelligibility" for Different Folks. In *Proceedings of the 2020 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

*co-first author

FIELD OF STUDY

Major Field: Cognitive Science

Studies in Computational Cognitive Science

Professor Marcelo Mattar

Abstract of the Dissertation

Adaptive Control with Episodic Mechanisms

by

Corey Zhou

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2024

Professor Anastasia Kiyonaga, Chair

What is memory for? Our daily experiences suggest that remembering serves adaptive purposes. Rather than simply storing information, memory actively selects, connects, and organizes it to predict the future and guide decisions. However, despite the suggestive links between memory and adaptive behavior, the mechanistic basis of this connection remains unclear. Memory research often studies memory in contexts with limited requirements for adaptive control, while the decision-making literature leaves out detailed memory mechanisms.

This dissertation aims to improve the existing memory-for-decision-making theory by grounding model-based sequential decision-making in classic mechanistic models

of episodic memory. Bridging across a body of previous results, I propose a formal framework for memory-informed decision-making that is cognitively plausible. In a series of three projects, I first show how a phenomenological account of episodic memory suggests methods of model-based evaluation and choice in sequential tasks. I then empirically test several key predictions implied by this framework, revealing remarkable decision patterns resulting from episodic memory biases. Finally, I extend the core framework to further explain how episodic memory operates over long timescales and event structures to enable continual learning and control. These three projects lay the groundwork for reverse engineering the hidden cognitive processes behind adaptive decision-making by leveraging well-studied episodic mechanisms.

Chapter 1

Introduction

What is memory for? Remembering information from one's past is clearly beneficial for survival, especially for rare but critical events such as "fire burns" or "some crocodiles attack humans." More generally, memory is crucial for adaptive behavior in a world filled with uncertainty: it is preferrable to get coffee from a shop that consistently serves better coffee than other closeby shops; it is also wise to pack an extra jacket for a weekend getaway because it snowed out of nowhere last time. Memory thus goes beyond the act of mere *remembering* to actively *selecting*, *connecting*, and *organizing* information. By encoding the temporal relationship between observations, actions, and outcomes, memory tracks causality, which informs future decisions for more rewarding experiences.

A process-level model offers unique insights into understanding the mind. David Marr's three levels of analysis identified process-level models as those primarily concerned with representations and algorithms (Marr, 1982). These models sit between computational models (which address the purpose of computation) and implementational models (which describe the physical realization). Specifically, process-level models attempt to answer questions like

> How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?

Similarly, a mechanistic theory about memory and decision making should explain the following:

> How can this *memory-for-decision-making* theory be implemented? In particular, what is the representation for the *memory samples* and *decision variables*,

and what is the algorithm for the transformation?

By explicitly defining the assumptions about representations and formally postulating the algorithm, such a framework generates testable hypotheses, such as manipulation of choice via manipulation of memory encoding or retrieval, as well as identification of possible neural substrates of memory that, if impaired, systematically change people's choice behavior.

The connection between memory and decision-making is of great interest, yet its mechanistic basis remains unclear. Empirical evidence suggests that people use a subset of past experiences to make choices in simple tasks, but computational models often fail to generalize to sequential decision problems or oversimplify human memory. Thus, a cognitively plausible and generalizable mechanistic account of memory's role in decision-making is lacking.

The potential interplay between memory and decision making has drawn interest from various fields, yet its mechanistic basis remains unclear. On one hand, empirical evidence suggests that people do use (a small subset of) their past experience to make choices in relatively simple (e.g., one-step) decision tasks (Bornstein & Norman, 2017; Bornstein et al., 2017; Duncan & Shohamy, 2016; Lieder et al., 2018; Nicholas et al., 2022; Plonsky et al., 2015). On the other hand, existing computational models often fail to generalize to sequential decision problems (Bornstein et al., 2017; Nicholas et al., 2022), or oversimplify the defining features of human memory (Blundell et al., 2016; Lengyel & Dayan, 2007; Ritter et al., 2018). Thus there is a lack of cognitively plausible and generalizable mechanistic accounts for ***how*** and ***what*** memory is recruited in decision-making.

One of the main aims of this dissertation is to improve the existing memory-for-decision-making theories, offering a mechanistic model that better explains the adaptive purpose of memory in sequential decision tasks and is better informed by decades of research in human memory.

1.1 Human episodic memory and its properties

Tulving (1972) proposed a dichotomy of declarative memory systems: *semantic* and *episodic* memory. Semantic memory consists of organized information that is relatively robust to accidental alteration and can be used as "cognitive reference." It resembles a summary of experiences and/or knowledge without perceptual or temporal details. In contrast, episodic memory is autobiographical, encoding details like *what*, *when*, and *where* – in other words, the temporal relationship between observations and actions, as well as the relevant spatial information. Tulving hypothesized that episodic memory is more susceptible to alteration with each retrieval. This dissertation focuses on episodic memory but also posits the role of semantic memory in decision-making.

Episodic memory has been studied since the inception of experimental psychology. Ebbinghaus (1885) tested memory retention by memorizing a list of nonsense syllables and tested himself after different study-test intervals. Although self-experimentation is now at best questionable, the experiment paradigm Ebbinghaus invented is quite similar to what episodic memory studies often uses – the free recall task (Fig. 1.1a). In this task, participants study a sequence of words and then freely recall them in any order (Kahana, 1996; Murdock, 1972; Ratcliff & McKoon, 1981).

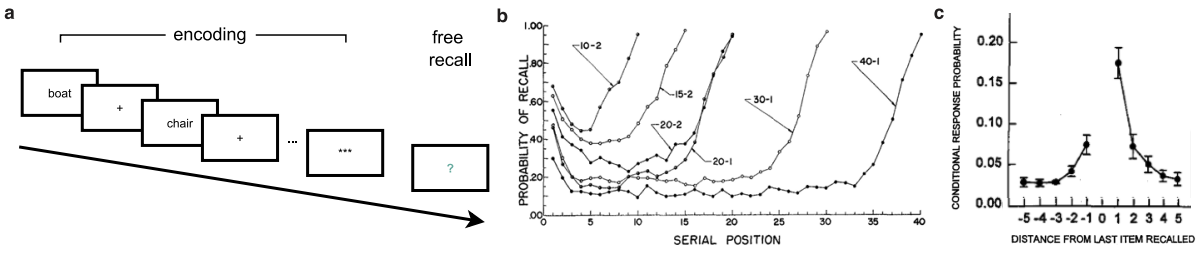

Figure 1.1: **The free recall task and episodic memory effects in humans. (a)** Schematic of a basic free recall task with two phases: encoding and (free) recall. During the encoding phase, words are presented (either visually on a screen or auditorily) in a randomized order with a fixation in between. During the recall phase, the subject writes down the studied words in any order they wish. **(b)** Serial position effect cruves from Murdock (1962). The two numbers annotated on each curve correspond to the length of the word list (first number) and the presentation duration in seconds (second number). **(c)** Conditional response probability curves from one of the studies in Murdock (1962) (plot obtained from Kahana, 1996). Specifically, the lists had 30 words and each word was presented for 1 second auditorily.

Free recall studies consistently reveal intriguing patterns like serial position effects and contextual recall effects. *Primacy effect*, for instance, refers to the tendency that

people exhibit higher recall rates for words near the beginning of a list (Atkinson & Shiffrin, 1971; Murdock, 1962; Rundus, 1971; Tan & Ward, 2000) (Fig. 1.1b), and is likely driven by verbal rehearsal (Howard & Kahana, 1999; Marshall & Werder, 1972). A slightly different measure of primacy effect is by the probability of first recall (i.e., the probability distribution of the serial position of first recalled word in the studied list; see Howard and Kahana, 1999; Kahana et al., 2008; Laming, 2010).

Subjects also exhibit a *recency effect* under certain experiment conditions, such that they recall words near the end of a list more often than those in the middle (Fig. 1.1b; see Atkinson and Shiffrin, 1971; Murdock, 1962; Rundus, 1971; Tan and Ward, 2000). Again, this effect may be measured alternatively with the probability of first recall (Howard & Kahana, 1999; Kahana et al., 2008). It is attributed to the short-term (working) memory and disappears with distractor tasks (Greene, 1986; Howard & Kahana, 1999; Kahana et al., 2008).

The most well established contextual recall effect is the *temporal contiguity effect* (Kahana, 1996; Murdock, 1962; Murdock & Okada, 1970): words studied close in time are likely recalled together (temporal contiguity) in the same order (foward asymmetry). This effect is quantified by lag conditional response probability (lag-CRP) and shows a pattern centered around zero but elevated in the positive lag direction, indicating forward asymmetry (Fig. 1.1c; Kahana (1996)). Specifically, lag-CRP is computed as the conditional probability that, given the most recently recalled stimulus and its serial position i during encoding, the subsequently recalled stimulus comes from serial position $i + j$, where j is a signed integer representing the lag. The stronger the asymmetry, the higher temporal clustering the recall exhibits. In addition to temporally adjacent items, people also tend to recall semantically related words together (i.e., semantic clustering; Howard and Kahana (2002b) and Patterson et al. (1971)).

Both recency and contiguity effects are approximately scale-invariant, while the primacy effect decreases in longer lists (Fig. 1.1b; see Howard and Kahana, 1999; Howard et al., 2008; Murdock, 1962; Polyn et al., 2011). The consistency of the temporal contiguity effect, even when the task doesn't require a specific order of recall, is especially striking.

Moreover, this effect is robust across individuals (Healey et al., 2014) and categories (Polyn et al., 2011), although semantic structures may reduce the size of the contiguity effect (Healey & Uitvlugt, 2019). Temporal clustering even predicts performance in free recall tasks (Q. Zhang et al., 2022), suggesting that temporal contexts govern multiple timescales, and that the *temporal structure* is a key feature of human episodic memory and a decisive factor in memory performance.

However, the function of this temporal structure, beyond facilitating memory recall, is unclear. Since memory studies have primarily (if not exclusively) focused on recall performance in isolation, there is a gap between existing memory research and mechanistic theories of memory for adaptive decision-making. To truly answer "what is episodic memory for," it's essential to explain the role of temporal contiguity in memory and its impact on adaptive decision-making. It is vital to not only acknowledge the participation of episodic memory in adaptive choice, but also ground decision making in the dynamics of episodic encoding and retrieval.

1.2 Mechanistic models of episodic memory

Naively, one might think of episodic memory as a storage system where memories are placed in slots during encoding, and recalling involves applying retrieval rules to find the correct slot. However, to account for scale-invariance, adjustments are needed to either the probability of retrieval or mechanisms to track the latest observation. This is because observations are often encoded without knowing when they end (e.g., participants in free recall studies memorize words without knowing the list length). The latter – tracking the latest observation with specific memory mechanisms – is usually assumed in mechanistic models of episodic memory, such as the temporal context model (TCM; Howard and Kahana, 2002a).

TCM is an influential model of human episodic memory, providing a general framework to model memory recall as association retrieval. It replicates various episodic retrieval patterns in free recall tasks, including primacy[1], recency, and temporal contiguity effects. Over the past two decades, TCM has been augmented with additional mechanisms

[1]An additional primacy bias is needed; see Polyn et al. (2009b) and Lohnas et al. (2015).

to explain more episodic memory phenomena, such as the context maintenance and retrieval model (CMR; Polyn et al., 2009a) accounting for clustering effects, CMR3 for memory effects (Cohen & Kahana, 2019), and many more (Healey & Kahana, 2016; Lohnas et al., 2015; Sederberg et al., 2008; Talmi et al., 2019). Prior studies have further suggested the entorhinal cortex and the hippocampus as the underlying biological infrastructures (Howard et al., 2005; Kragel et al., 2020; Sakon & Kahana, 2021).

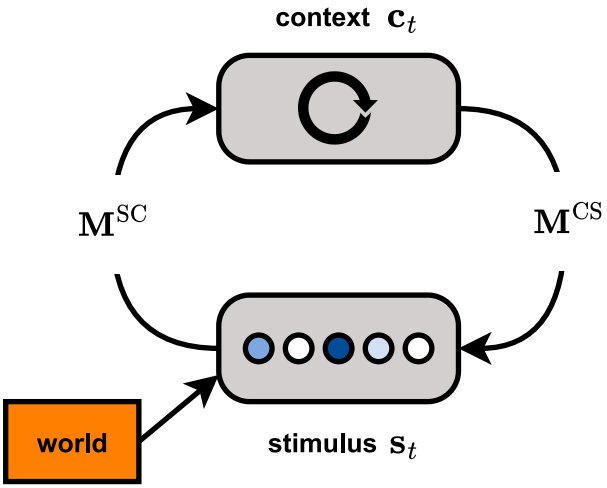

Figure 1.2: **A schematic of the temporal context model (TCM).** When a new stimulus arrives from the external environment (encoding phase) or is recalled from memory (retrieval phase), its features update an evolving temporal context via an associative matrix $\mathbf{M}^{\mathrm{SC}}$. The context representation integrates the update while maintaining information about the previous observations. During memory retrieval, stimuli bearing more similarity to the active temporal context are more likely to be recalled, the extent of which is mediated by another associative matrix $\mathbf{M}^{\mathrm{CS}}$. This process repeats as the agent encodes or recalls information.

TCM assumes that episodic encoding and recall are both mediated by an evolving *temporal context* (Fig. 1.2). It is through this temporal context that items experienced close in time become associated during encoding, driving subsequent recall and giving rise to the temporal contiguity effect. Essentially, at any point in time, the context is a fuzzy representation of the agent's past experience weighted by recency.

At time t, TCM centrally posits that the temporal context, $\mathbf{c}_t$, evolves according to

$$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{c}_t^{\mathrm{IN}}. \tag{1.1}$$

Here, the temporal context is updated by an input context $\mathbf{c}_t^{\mathrm{IN}}$. The content of $\mathbf{c}_t^{\mathrm{IN}}$ is typically determined by information arriving either externally through the senses during

encoding or internally through memory during retrieval. The extent of this update is controlled by β, commonly referred to as the *drift rate*. Simultaneously, ρ determines how much of the previous context is retained. To make sure the system is stable (e.g., the context does not grow without bound), $\mathbf{c}_t$ is often constrained to a unit vector, and ρ and β are chosen accordingly. Thus if β is small (i.e., ρ is large), temporal contexts close in time are more similar to each other; if β is large (i.e., ρ is small), temporal contexts close in time are likely different.

During encoding, each observed stimulus s_t becomes associated with the temporal context $\mathbf{c}_t$ present at that moment. Formally,

$$\mathbf{M}^{\mathrm{CS}} \leftarrow \mathbf{M}^{\mathrm{CS}} + \alpha \mathbf{x}_t \mathbf{c}_t^{\intercal}, \tag{1.2}$$

where α is the learning rate, $\mathbf{x}_t$ (shortform for $\mathbf{x}(s_t)$) is the representation of stimulus s_t in terms of features, and $\mathbf{M}^{\mathrm{CS}}$ stores the associations between item and context. During retrieval, the probability of retrieving (sampling) an item is proportional to how well the context associated with that item matches the current temporal context, or

$$p(s_k) \propto \mathbf{M}^{\mathrm{CS}} \mathbf{c}_t \cdot \mathbf{x}_k,$$

with retrieval influencing the temporal context through Eq. 1.1 and, consequently, subsequent recalls (Fig. 1.2).

TCM explains the recency effect by maintaining the temporal context at the end of encoding. Because temporal contexts drifts continuously, the last context is more strongly associated with the stimuli encountered towards the end. It also captures the temporal contiguity effect through the 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 is shared around a point in time with both previous and subsequent contexts, but the latter is only incorporated in ensuing contexts. As a result, TCM predicts lag-CRP to be asymmetrical,

with higher probability to recall subsequent stimuli than preceding ones, and close-by stimuli than remote ones.

The temporal context is akin to a moving spotlight with a fuzzy edge, carrying recency-weighted historical information relevant to the present, with the degree of information decay controlled by ρ. Larger β values result in sharper lag-CRPs, stronger forward asymmetry, and stronger temporal clustering – core features of human episodic memory (Q. Zhang et al., 2022). Neuroimaging studies suggest the hippocampus – in particular, CA1, CA2, CA3[2], and CA23DG[3] – as candidate substrates for representing the temporal context (Dimsdale-Zucker et al., 2018; Nielson et al., 2015; Schapiro et al., 2013; F. Wang & Diana, 2017).

In summary, TCM provides an algorithmic hypothesis of how human episodic memory, especially in free recall paradigms, manifests specific retrieval dynamics contingent on the temporal order. It is worth noting that TCM is a phenomenological (rather than normative) model. Unlike normative models, which aim to explain how cognition *should* happen – for example, optimizing with respect to an objective function such that the maximum value corresponds to the best performance, phenomenological models only seek to faithfully describe experimental data. i.e., how cognition *is*. Thus TCM aims to reproduce empirically observed patterns rather than to rationalize them. It neither prescribes the optimal solution nor implies that human episodic mechanisms are inherently "correct" or "standard".

1.3 Memory-informed decision-making

The relationship between memory and decisions is most evident in procedural memory, a type of non-declarative memory. The role of procedural memory in decision making is often modeled using a neurocomputational account where dopamine signals reward prediction errors, and stimulus-response associations underpin habits. This is closely related to model-free (MF) control in reinforcement learning (RL), where temporal

[2]The involvement of CA3 in temporal context representation is controversial and is postulated to be limited to cases where similar memories need to be differentiated; see F. Wang and Diana (2017).

[3]Coarse temporal contexts, especially in conjunction with cognitive contexts; see Dimsdale-Zucker et al. (2022).

difference (TD) prediciton errors drive the incremental update of choice proxies (Sutton, 1988). Intuitively, both procedural memory and MF control adopt a trial-and-error approach to directly learn the value of an action. For instance, the preference of a coffee shop might result from MF learning, such that repeated experiences update the expected value, or "goodness," of each shop based on the actual reward. i.e., for coffee shop s,

$$\hat{V}(s) \leftarrow \hat{V}(s) + \alpha\delta,$$

where $\delta = R(s) - \hat{V}(s)$ is the prediction error between actual reward $R(s)$ and the previous value estimate $v(s)$ and α is the learning rate. The choice proxies lead to a preference for the coffee shop with the highest value. Formally, the value $V(s)$ is called a *decision variable*. Dopamine activity is believed to track TD prediction errors and drive MF learning (Chang et al., 2015, 2017).

However, MF controllers usually lack the ability to learn the nuanced dynamics of the world. For instance, if coffee shop s' hires a top barista, a naive MF algorithm would slowly adjust the aggregated statistic $\hat{V}(s')$ to reflect a broad-brush "averaged" view of past experiences, without realizing the fundamental change. Similarly, because procedural memory (e.g., habits) does not typically retain spatiotemporal details during encoding, the agent likely needs to relearn the action sequence to adapt to the changed environment or task.

In contrast, model-based (MB) control, another RL method, uses world model to estimate values. Recent studies suggest that dopamine may play a key role in MB control (Sharpe et al., 2017) and associative learning (Sharpe et al., 2020). An MB algorithm may learn *transition probabilities* as part of its world model, facilitating quick recomputation of values if the agent believes the world has changed. For instance, it may have two world models, one corresponding to the world where s' consistently serves mediocre coffee, and

the other where s' consistently serves amazing coffee. e.g.,

$$\mathbb{P}(\{s' \text{ has B-grade coffee}\}|W = 1) = 0.8$$
$$\mathbb{P}(\{s' \text{ has A-grade coffee}\}|W = 1) = 0.2$$
$$\mathbb{P}(\{s' \text{ has B-grade coffee}\}|W = 2) = 0.2$$
$$\mathbb{P}(\{s' \text{ has A-grade coffee}\}|W = 2) = 0.8$$

Here, $W \in \{1, 2\}$ denotes the true state of the world. To make a decision, the MB controller computes the *expected return*, which is the expected reward of a particular option (e.g., the shop s). In the coffee shop example, denoting the reward associated with a certain grade of coffee as $R_{x\text{-grade coffee}}$, the MB value estimate is

$$\hat{V}(s') = \mathbb{E}\left[R(s')\right] = \sum_{w\in\{1,2\},x\in\{\text{fair, good}\}} \mathbb{P}(W = w)\,\mathbb{P}(\{s' \text{ has } x\text{-grade coffee}\}|W = w)\,R_{x\text{-grade coffee}}.$$

Given the connection between procedural memory and MF control, there is a growing interest in exploring the relationship between declarative memory and MB control, which represents a form of deliberative evaluation in goal-directed behavior (Doll et al., 2015). This is because unlike procedural memory, which is formed through trial-and-error of reward prediction, declarative memory consists of explicit knowledge of the environment or task. In other words, instead of directly learning an action sequence for a specific problem, declarative memory may facilitate learning of the structure of the problem to inform its action plan, which corresponds to MB control. The cognitive map (Tolman, 1948) is a potential world model that encodes structured information from experience for adaptive inference (Eichenbaum, 2001). A semantic memory system, subserved by cortical learning mechanisms, is suggested to support the formation of cognitive maps. However, its slow, incremental learning is more MF-like and cannot account for the fast, one-shot decisions often made by people (e.g., pack an extra jacket because it snowed out of nowhere in the past). A promising alternative thus lies in the other declarative memory system: episodic memory.

Episodic memory has been proposed to guide MB evaluations and decisions by

scaffolding the construction of hypothetical future scenarios (Schacter & Addis, 2007; Schacter et al., 2015). Prior research has shown that subject choices often reflect the use of selective samples of past experience (Bornstein & Norman, 2017; Duncan & Shohamy, 2016; Lieder et al., 2018), with a preference for memory-based decisions over MF-like decisions, especially in volatile environments (Nicholas et al., 2022). Some even argued that the main purpose of episodic memory is not to remember the past, but rather to anticipate the future (Klein, 2013, 2016).

This has led to interest in a class of *decision-by-sampling* algorithms. These algorithms loosely resemble episodic memory in that decisions are achieved by sampling a small number of past events and their outcomes (Bornstein & Norman, 2017; Bornstein et al., 2017; Lieder et al., 2018; Plonsky et al., 2015). The sampling dynamics is posited to be implemented through episodic encoding and retrieval, such as the one described in TCM (Howard & Kahana, 2002a) and CMR (Polyn et al., 2009a). In line with these predictions, patients with episodic memory deficits show impairments in decision making tasks (Bakkour et al., 2019; Gupta et al., 2009; Gutbrod et al., 2006), and episodic memory is strongly modulated by rewards and emotionally salient information (Clewett et al., 2019; Horwath et al., 2023; Mather et al., 2015; Talmi et al., 2019).

Despite the theoretical and empirical support for episodic memory in decision-making, it remains unclear why decision-relevant information must come from episodic memory rather than other types of memory. Most research so far has focused on one-step tasks (Bornstein & Norman, 2017; Bornstein et al., 2017; Duncan & Shohamy, 2016; Nicholas et al., 2022; Rouhani et al., 2018), viewing episodic memory in decisions as merely a veridical record of past events (Braun et al., 2018; Duncan & Shohamy, 2016; Nicholas et al., 2022). For example, the multi-armed bandit problem, a classic RL task used in these studies, requires a decision maker to repeatedly choose one of the options ("arms") and maximize the total reward over many trials. While the reward associated with a option may change over time due to environmental volatile, these changes are independent of the participant's choice – that is, the choice at an earlier time does not cause changes to the environment in a way that affects choices at later times. Other

studies that require more extensive integration of episodic memory also face this limitation (Shadlen & Shohamy, 2016). These tasks do not pose any significant temporal structure across successive time steps. To maximize the total reward, it suffices to maximize the expected reward at each time step (or a handful of time steps) independently.

1.4 Models of memory and sequential decision-making

Unlike bandit tasks, actions in the real world often have longstanding consequences. Myopic decisions that may solve a bandit task do not play out well in the long term, as future states of the world are stochastically influenced by the previous world state and the action. For example, drinking bad coffee might boost alertness temporarily but cause heartburn later; an extra jacket might only prove useful during an abrupt snowstorm days later. Sequential decision tasks like spatial navigation or chess further suggest that the brain engages in constructive, deliberate evaluation, akin to mental simulation informed by map- or model-like information (Pfeiffer & Foster, 2013; van Opheusden et al., 2021). Rather than independently optimizing the action at each step, people likely optimize *over time*. They may sample their episodic memory across an extended temporal horizon and strategically combine these episodic samples to compute decision variables. However, we still understand little about the mechanisms through which deliberative sequential decisions are made, particularly how they draw on specific memory processes long-established in memory laboratories. Studying episodic memory in one-step decision tasks offers a restricted view of memory-informed decision-making, leaving the purpose of episodic memory dynamics unclear.

Similarly, algorithms that enhance RL agents with memory capabilities improve performance in sequential decision problems but abstract away key aspects of episodic memory due to their focus on normative principles (Blundell et al., 2016; Gershman & Daw, 2017; Lengyel & Dayan, 2007; Pritzel et al., 2017; Ritter et al., 2018). For instance, Lengyel and Dayan (2007) proposed a three-system architecture where episodic control constitutes a distinct "third way" apart from MF and MB control. Gershman and Daw (2017) introduced an episodic RL algorithm that computes action values by considering all relevant past trajectories. Pritzel et al. (2017) and Ritter et al. (2018) modeled long-term

memory using a differentiable neural dictionary (DND) to look up the value associated with a particular action and compare across actions. Nonetheless, these RL methods have limited empirical support for their stylized approach to model episodic memory and incorporate few details about episodic memory beyond one-shot learning.

Another approach to model long-term memory and sequential decision-making makes use of deep neural networks. For instance, recurrent neural network (RNN) models have been shown to support efficient deliberation (X. Zhang et al., 2020) and capture human-like generalization in sequential tasks (Giallanza et al., 2024). Feedforward neural network (FNN) models can infer latent causes in a way that produces context-dependent predictions consistent with human participants (Lu et al., 2023), while Transformer models with episodic history solve sequential tasks with a compositional structure (i.e., consisting of several subtasks; Pashevich et al., 2021). As these models often stem from machine learning research with the goal of improving task performance, their primary goal is not to explain the detailed mechanisms of human memory in decision-making. As the result, their conceptualization of "episodic memory" is superficial. Moreover, while some of them are designed to mirror classic episodic memory models such as CMR (e.g., Giallanza et al., 2024), their black-box nature limits interpretability and insights into the underlying process.

In summary, the theory linking episodic memory and decision-making has gained attention in decision-making and machine learning research. Yet, just like memory research has resulted in limited explanations regarding the adaptive purpose of memory in complex decision scenarios, decision-making models have largely overlooked well-established features of episodic memory. To formally bridge this gap – in particular, model-based choice – it is thus crucial to integrate insights from both fields, including a rational explanation of episodic memory features and a decision-making model that aligns with episodic memory dynamics. To address this, Chapter 3 proposes a novel mechanistic model positing how effective decision-making can arise from episodic encoding and retrieval by elaborating and extending a theoretical link between TCM and the successor representation (SR) in RL (see Section 2.3 for details). This model encompasses a family of decision-by-sampling

algorithms, some revisiting classic RL algorithms, while others present a new "middle ground". It also makes quantitative predictions about action evaluation in the small sample regime, emotional modulation effects, and experience generalization. Chapter 4 subsequently presents a set of preliminary results that support the theory of sequential decision making through episodic sampling.

1.5 Event representations

Thus far, I have described phenomenological findings and mechanistic models of episodic memory, alongside the growing research on how episodic memory influences decision-making in both cognitive science and machine learning. However, a noteworthy gap remains – a cognitively-informed process-level account of episodic memory's role in sequential decision-making. A key question arises: *what exactly is an "episode" in episodic memory?*

We live in a continuum, yet our living experience is rarely an uninterrupted stream. Movies are made of sequences of images presented at a constant rate; walking from the office to the closest coffee shop is a continuous sequence of steps. But we recount the story of a movie as discrete subplots, and plan the coffee break in terms of chunks of actions (e.g., going down the building, walking to a landmark). We naturally perceive events and recognize tasks despite a dearth of instructions. Each identified event likely constitutes an "episode" in episodic memory (Ezzyat & Davachi, 2011; Sargent et al., 2013).

Although terms like "event" and "event model" are relatively new (Radvansky & Zacks, 2017; Zacks et al., 2007), the idea is rooted in cognitive science, underpinning the early schema theory (Bartlett, 1932), the later structural schema theories (Iran-Nejad & Winsler, 2000; Rumelhart, 1980), situation models (Johnson-Laird, 1983; Zwaan & Radvansky, 1998), and subgoals (McGovern & Barto, 2001). While the terms differ, the ideas and evidence converge. Understanding streams of experiences in terms of structured representations allowed us to break down complex situations and tasks, and to begin to study the representations individually and interatively with each other.

Event and task structures systematically interact with episodic memory. Event segmentation theory (EST; Zacks et al., 2007) suggest that humans maintain event models

tracking causal structures in the stream of experience (Radvansky, 2012; Radvansky & Zacks, 2017). These abstract representations are generally stable, such that large prediction errors signal boundaries (Axmacher et al., 2010; Reynolds et al., 2007; Zacks et al., 2011). More recent event segmentation theories posit that temporal structures also predict perceived event boundaries (Baldwin & Kosie, 2020; Baldwin et al., 2008; Schapiro et al., 2013). Even when the transition between two observations is predictable, people appear to rely on statistical learning to align event boundaries with clusters of observations that occur close together in time. Events are thus identified by both bottom-up clues and top-down information such as goals and expectations (Dubrow & Davachi, 2016). In the latter case, clustered experiences may represent (sub)tasks that an agent actively engages in rather than events that they passively observes.

Interestingly, event types might not correspond to semantic clusters or share salient features that are deemed "meaningful" in the conventional sense. Schapiro et al. (2013) showed that participants grouped arbitrary fractal stimuli based on temporal proximity, which suggests that temporal statistics are sufficient to form structured knowledge, possibly through associative learning, without any prior knowledge. These event types are created and updated in a dynamic manner, consistent with what Bartlett (1932) predicted about schemata. Conversely, event perception driven by prediction error reflects an alternative pathway where cognitive templates or schemata scaffold memory organization.

Event and task boundaries are not static organizational tricks like labeled folders; they are dynamic, offering a range of behavioral (dis)advantages. They improve item recall (Heusser et al., 2018; Pettijohn et al., 2016), disrupt memory of temporal order (DuBrow & Davachi, 2013; Dubrow & Davachi, 2016; Heusser et al., 2018) or enhance such order memory (Wen & Egner, 2022) depending on the specific retrieval context. Long-term memory also impacts event structure encoding (Gershman et al., 2014; Zacks & Tversky, 2001), forming nested knowledge hierarchies (Baldassano et al., 2016; Hasson et al., 2015). The hippocampus is posited to rapidly learn and represent relational structures within episodes (McClelland et al., 1995), guiding memory reconstruction (Moscovitch et al., 2005; Norman & O'Reilly, 2003). Statistically learned event representations are

thought to complement this by exploiting statistical and temporal regularities in an unsupervised manner (Austerweil & Griffiths, 2011; Baldwin et al., 2008; Schapiro et al., 2013), enabling hierarchical event cognition (Fukai et al., 2021). Furthermore, effective knowledge representation likely entails both relational and statistical mechanisms (Kemp et al., 2010; Tenenbaum et al., 2011; Tversky et al., 2008).

Structured knowledge is likely reused in similar situations. From grocery trips to strategic video games, intuitive and effective solutions frequently adopt a divide-and-conquer approach. We combine actions within each event as learned skills in our behavioral repertoire to solve novel problems. So, how does our cognitive architecture represent and choose past experiences to subserve adaptive composition and generalization?

1.6 Adaptive control with event representations

McClelland et al. (1995) hypothesized that long-term memory balances between fast, adaptive episodic memory with slow, stable semantic memory, a tradeoff formulated in meta-learning (Bengio et al., 1991; Ritter et al., 2018). To facilitate efficient decision-making, memories should be maximally similar within an episode but maximally different across events. This allows an agent to exploit relevant experience and existing policies based on shared task structures while exploring heterogeneous representations to distinguish different tasks (Musslick & Cohen, 2021; Musslick et al., 2019). In hierarchical RL, discovering options (Sutton et al., 1999) or established sequences of actions enriches an agent's behavioral repertoire, enabling the composition of learned skills and fast adaptation (Barreto et al., 2021; Machado et al., 2023). More recently, RNN modeling work suggests that episodic encoding improves representation learning efficiency, while episodic retrieval biases the agent to exploit learned task representations (Lu et al., 2024).

In support of these normative theories, event/task representations have been suggested to contribute to cognitive control and adaptive behavior. Segmenting sequential tasks helps problem-solving (Anderson & Fincham, 2014) by creating structured representations (Franklin et al., 2020; Goodman et al., 2011; Kemp & Tenenbaum, 2008; Kemp et al., 2010) that can be reinstated during inference (Baldassano et al., 2016; Gershman & Niv, 2012; Graesser et al., 1994) and support knowledge transfer (Zadbood et al., 2017).

This minimizes task interference (Botvinick et al., 2001; E. K. Miller & Cohen, 2001) and promotes efficient planning (Cushman & Morris, 2015; Tomov et al., 2018). Event representations thus not only help us understand our personal experience, but also support active control (Flesch et al., 2023; Giallanza et al., 2024; Rougier et al., 2005).

However, distinct event representations can also have adverse effects. Specifically, abruptly changing the contingency between the unconditioned and condition stimuli in classical conditioning accounts for failure to eliminate fear memories (Gershman et al., 2010; Redish et al., 2007) because animals leveraged the learning context to decide when (not) to generalize its behavior (Bouton, 2004). Instead of blindly consolidating all past experiences with a unified representation, animals compartmentalize episodic memories by inferring events from context. Thus even after successfully learning a new association, old memories remain intact. When the environment reverts (e.g., being placed in the old cage), behavior also "regresses.". On the other hand, gradual changes (e.g., incrementally reducing the frequency of the aversive stimuli) reduce the likelihood of fear recurrence (Gershman et al., 2013), which aligns with the idea that prediction errors drive event segmentation, and sudden changes increase the likelihood of event boundaries (i.e., the animal predicts a significantly different future experience than the actual experience *under dynamics of the assumed event*). But what determines whether the current experience belongs to the onging event?

One prominent hypothesis is that each event corresponds to a different latent state of the world (Gershman & Niv, 2012; Redish et al., 2007). As experiences are encoded by episodic memory, the agent engages in *latent cause inference* to organize them into higher-order representations. The learned dynamics of each hidden state allow the agent to predict future rewards, as the same action may have different outcomes depending on the true world state (e.g., packing an extra jacket for a spring trip to New York vs. San Diego – the former is wise, the latter is nuissance). Surprise consequently signals a possible change in the latent state and prompts a re-evaluation of optimal choices. Chapter 5 addresses this challenge by formalizing the three-way connection between memory, events, and intelligent control with a computational model that delineates the underlying mechanism.

In closing, this dissertation sets out to answer the overarching question "what is memory for" by positing the role of episodic mechanisms in adaptive sequential decision-making. To formalize this hypothesis, Chapter 3 theorizes that the temporal dynamics of TCM offers an unexpected but built-in way to perform model-based evaluation and control over extended time. Chapter 4 empirically tests the proposed theory that adaptive decisions recruit sequential episodic retrieval – specifically, it examines three important hypotheses implied by the model: (1) memories with higher recall have a larger weight on memory-based decisions, (2) but such effect may be modulated using temporal contiguity effect, and (3) that the choice between options composed of temporally extended events is best predicted by what is recalled. Finally, Chapter 5 postulates the basis of meta-learning (i.e., learning to learn) to lie in hierarchical episodic mechanisms that take event structure into account.

Chapter 2

Formalism

This chapter introduces the key elements of the framework that will carry through the rest of the dissertation. Specifically, Section 2.1 lists the abbreviations and notations used for the rest of the dissertation; Section 2.2 and Section 2.3 covers the core computational framework of reinforcement learning; Section 2.4 and Section 2.5 describes the foundational theories of episodic memory that Chapter 3 and Chapter 5 build on; finally, Section 2.6 presents the latent cause inference framework used in Chapter 5 to model episodic dynamics and decision-making behavior given more complex event structures.

2.1 Notations and Abbreviations

2.1.1 Abbreviations

CMR	context maintenance and retrieval
CRP	conditional response probability
DM	decision making
DND	differentiable neural dictionary
EGO	Episodic Generalization and Optimization
EM	episodic memory
EST	event segmentation theory
FNN	feedforward neural network
HRL	hirarchical reinforcement learning
MAP	maximum a posteriori
MB	model-based
MDP	Markov Decision Process
MF	model-free
MRP	Markov Reward Process
RL	reinforcement learning
RNN	recurrent neural network
sCRP	sticky Chinese restaurant process
SEM	structure event memory
SF	successor features
SR	successor representation
TCM	temporal context model
TD	temporal difference

2.1.2 Reinforcement Learning Notations

s	state
S_t	state experienced at time t
a	action
A_t	action executed at time t
r	reward
R_t	reward encountered at time t
γ	discount factor (also referred to as the temporal horizon)
α	learning rate
λ	eligibility trace decay rate
π	policy
$\mathbf{r}$	one-step reward function
R_t	reward obtained at time t
$\mathbf{v}$	state value function
$V(s)$	expected total reward from visiting state s
$\widehat{V}(s)$	estimated total reward in expectation by visiting state s
$Q(s, a)$	expected total reward by performing action a in state s
$\widehat{Q}(a)$	estimated total reward in expectation by performing action a in state s
$q(a)$	expected total reward by performing action a
$\hat{q}(a)$	estimated total reward in expectation by performing action a
$\mathbf{e}/\boldsymbol{\xi}$	eligibility trace
$\mathbf{T}$	one-step transition matrix
T_{ij}	the (i, j)-th entry of $\mathbf{T}$
$\mathbf{P}$	latent one-step transition matrix
$\mathbf{M}$	successor representation (SR)
M_{ij}	the (i, j)-th entry of $\mathbf{M}$
$\boldsymbol{\phi}_s$	basis function of state s
$\boldsymbol{\Psi}$	successor features (SF)
$\boldsymbol{\psi}_s$	successor features of state s

2.1.3 Temporal Context Model Notations

γ	proportion of previous temporal context used in updating the current context during *encoding*
$\widetilde{\gamma}$	effective time horizon
ρ	proportion of previous temporal context used in updating the current context during *retrieval*
β	proportion of the retrieved temporal context used in updating the current context during *retrieval*
p_{stop}	interruption probability during retrieval
α	learning rate during encoding
α_{mod}	emotionally modulated learning rate during encoding
$\mathbf{c}_t$	temporal context at encoding time step t
$\mathbf{c}_i$	temporal context at retrieval time step i
$\mathbf{c}_t^{IN}$	retrieved temporal context at encoding time step t
$\mathbf{c}_i^{IN}$	retrieved temporal context at retrieval time step i
$\mathbf{x}(s)$	feature vector of stimulus s
$\mathbf{x}_t$	feature vector of stimulus experienced at encoding time step t
$\mathbf{x}_i$	feature vector of stimulus experienced at retrieval time step i
T	total encoding time
N	total number of samples retrieved (retrieval steps)

2.1.4 Latent Cause Inference Notations

K	latent cause/state/event/task
$\hat{K}$	inferred event
NK	total number of inferred events
α	rate at which new events germinate
κ	rate at which high-frequency events reoccur
λ	stickiness of the sCRP process
w_{hist}	integration window of history
pe_{thres}	prediction error threshold

Solving sequential decision tasks frequently entails adaptive control to maximizing the total reward over an extended temporal horizon (e.g., an entire chess game) through careful action selection. More importantly, each choice and the corresponding state of the world (stochastically) determines the subsequent state of the world and the "goodness" of a future action. This problem is formalized as RL. When time is assumed to be discrete, the dynamics are modeled using a Markov decision process (MDP; Bellman, 1957). In this framework, an agent must choose among different actions at different points in time, each of which would lead to different future experiences and varying amounts of reward. To evaluate an action, they estimate a decision variable, a proxy to the expected future reward as the result of choosing that action.

2.2 Markov decision & reward processes

In an MDP, a task is formalized by a 5-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ denotes the set of states. A state may correspond to a specific observation or experience, such as a word token, a visual scene, or a coffee shop. $\mathcal{A}$ denotes the set of possible actions. $\mathcal{P} : \mathcal{S} \mapsto \mathcal{S}$ is the Markov transition function that defines the probability distribution $\mathcal{P}(s'|s)$ of transitioning from state s to state s'. $\mathcal{R} : \mathcal{S} \mapsto \mathbb{R}$ (or $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$) is the reward function $\mathcal{R}(s)$ (or $\mathcal{R}(s, a)$) specifying the reward magnitude received upon visiting state s (or executing the action a in s). Finally, $\gamma \in [0, 1)$ is the discount factor that controls the temporal horizon of computations by reducing the importance of rewards in distant future.

The goal of the agent is to choose the action that maximizes the cumulative discounted return $G = \sum_{t=1}^{\infty} \gamma^t \mathcal{R}(S_t)$, where S_t is a random variable denoting the state visited 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 $\mathbb{P}(S_{t+1} = s' \mid S_t = s) = \mathcal{P}(s'|s)$. This gives rise to a "trajectory" given by $S_1, R_1, S_2, R_2, S_3, R_3, \ldots, S_H, R_H$, where H is the length (number of time steps) of the full trajectory.

The value of state s, denoted $V(s)$, is defined as the expected return when starting

in s:

$$V(s) = \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k \mathcal{R}(S_{t+k}) \mid S_t = s\right]$$

The state-action value $Q(s, a)$ is defined as the expected return when taking action a in state s:

$$Q(s, a) = \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k \mathcal{R}(S_{t+k}, A_{t+k}) \mid S_t = s, A_t = a\right].$$

State values are related to state-action values such that

$$V(s) = \mathbb{E}_A\left[Q(s, a) \mid a \in A\right]$$

In order to select an action in a fixed state s, the agent estimates $Q(s, a)$ for each candidate action. The field of RL describes various methods for estimating $Q(s, a)$, broadly divided into model-free and model-based methods. Model-free methods are those where the agent learns to estimate $Q(s, a)$ directly from experience. The classic temporal difference (TD) algorithm, for example, iteratively updates the agent's estimate $Q(s, a)$ as

$$Q(s, a) \leftarrow Q(s, a) + \alpha\left(\gamma \mathcal{R}(s, a) + \gamma^2 \max_{a'} Q(s', a') - Q(s, a)\right)$$

whenever action a is performed (Sutton, 1988). The learning rate α affects the rate of convergence. In model-based methods, in contrast, the agent uses a model of the world (i.e., an estimate of $\mathcal{P}$ and $\mathcal{R}$) to estimate $Q(s, a)$. If both $\mathcal{P}$ and $\mathcal{R}$ are perfectly known, the agent can generate a plausible trajectory $S_1, R_1, S_2, R_2, S_3, R_3, \ldots, S_T, R_T$, where $S_{i+1} \sim \mathcal{P}(.|S_i)$ and $R_i = \mathcal{R}(S_i)$. Each such trajectory is called "rollout", alluding to the fact that states (and rewards) are sampled recursively (Tesauro & Galperin, 1996). The total discounted reward along a rollout trajectory is a Monte Carlo estimate of the action value, i.e., $Q(a) = \sum_{i=1}^{H} \gamma^i R_i$.

Given a set of states, the distribution over the available actions is called the *policy*

of the agent. Formally, policy is a probability mass function of the form

$$\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{R}^+ \cup \{0\}.$$

It takes in a state-action pair (s, a) and outputs the probability that the agent selects action a at state s. State and action values are both calculated with respect to a particular policy.

Related to MDP is the Markov Reward Process (MRP), which is a discrete-time stochastic process that extends a Markov Chain by adding a reward to each state. Unlike in an MDP, the state dynamics in an MRP are not under control of the agent. This is equivalent to 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 action) and not *control* (e.g., which action to select).

2.3 Successor Representation and Features

Assuming states are one-hot encoded (e.g., state s is represented by $\mathbf{1}_s$), consider the policy-dependent one-step state-transition matrix $\mathbf{T}^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ whose (i, j)-th entry T^π_{ij} corresponds to the probability of transitioning from state i to state j: $T^\pi_{ij} = \mathbb{P}^\pi(S_{t+1} = s_j \mid S_t = s_i)$. Consider also the one-step reward vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}|}$ whose k-th entry r_k corresponds to the reward present in state k. Assuming rewards are only obtained *after* a transition (e.g., recalling a word), the value function under the policy π can be expressed in vector form as:

$$\begin{aligned}
\mathbf{v}^\pi &= \mathbf{T}^\pi\mathbf{r} + \gamma(\mathbf{T}^\pi)^2\mathbf{r} + \gamma^2(\mathbf{T}^\pi)^3\mathbf{r} + \cdots \\
&= \left(\sum_{k=0}^{\infty} \gamma^k(\mathbf{T}^\pi)^k\right)\mathbf{T}^\pi\mathbf{r} \\
&= (\mathbf{I} - \gamma\mathbf{T}^\pi)^{-1}\,\mathbf{T}^\pi\mathbf{r}.
\end{aligned} \tag{2.1}$$

The successor representation (SR; Dayan, 1993) is defined as

$$\mathbf{M}^\pi_\gamma = (\mathbf{I} - \gamma\mathbf{T}^\pi)^{-1}\,\mathbf{T}^\pi, \tag{2.2}$$

where γ is the temporal discount factor. A lower γ discounts future occurrences more and therefore corresponds to a more myopic agent who prefer immediate observations. The (i, j)-th entry M_{ij}^{π} corresponds to the expected sum of future visitations to state j from state i, discounted according to γ. The SR can be learned directly from experience using TD learning. If the "true" SR is available to the agent, all state values can be estimated simultaneously by $\mathbf{v}^{\pi} = \mathbf{M}_{\gamma}^{\pi}\mathbf{r}$.

Without access to the ground truth $\mathbf{T}$, SR can be iteratively learned using TD-learning. This TD-SR algorithm minimizes the prediction error between the predicted and actual visitations over time. Specifically, at each time point t, the raw prediction error is

$$\mathbf{1}'_{S_t} + \gamma \mathbf{1}'_{S_t}\mathbf{M}_{t-1}^{\pi} - \mathbf{1}'_{S_{t-1}}\mathbf{M}_{t-1}^{\pi}.$$

To break it down, $\mathbf{1}'_{S_t} + \gamma \mathbf{1}'_{S_t}\mathbf{M}_{t-1}^{\pi}$ counts the number of actual visitations by adding the current visit ($\mathbf{1}_{S_t}$) and all expected future visitations with temporal discount ($\gamma \mathbf{1}'_{S_t}\mathbf{M}_{t-1}^{\pi}$). $\mathbf{1}'_{S_{t-1}}\mathbf{M}_{t-1}^{\pi}$ counts the number of predicted visitations.

In addition, upon observing the transition $S_{t-1} \to S_t$, we can use an eligibility function to capture the state that leads to the current state S_t:

$$\mathbf{e}_t = \gamma\lambda \mathbf{e}_{t-1} + \mathbf{1}_{S_{t-1}}. \tag{2.3}$$

where $\lambda \in [0, 1]$ controls the trace decay rate. Large λ values thus propagate the effect of a transition back in time. This is analogous to the eligibility trace over the set of states $\mathbf{X}$.

Finally, the TD-learned SR at time step t is updated according to

$$\mathbf{M}_t^{\pi} \leftarrow \mathbf{M}_{t-1}^{\pi} + \alpha \mathbf{e}_t(\mathbf{1}'_{S_t} + \gamma \mathbf{1}'_{S_t}\mathbf{M}_{t-1}^{\pi} - \mathbf{1}'_{S_{t-1}}\mathbf{M}_{t-1}^{\pi}). \tag{2.4}$$

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 rewards are collected upon *entering*, but not *exiting* a state.

Behavioral and neural evidence suggests that SR is a predictive representation that humans and animals appear to learn and use (Momennejad et al., 2018; Piray & Daw, 2021; Russek et al., 2017, 2021; Stachenfeld et al., 2017). In particular, SR constitutes a reusable (cached) world model that separates the transition structure of the world from the reward structure. For instance, with SR, state values can be computed by a linear operation $V(s) = \mathbf{1}_s'\mathbf{M}\mathbf{r}$. This avoids recursively solving the Bellman equation over multiple time steps, making SR a temporally abstracted representation. Thus SR effectively "flattens" temporally extended trajectories, transforming multi-step dependencies into a one-step bandit-like structure.

Successor features (SF; Barreto et al., 2017) further generalize SR. Instead of the expected number of future visits to each *state* as in SR, SF captures how frequently each *feature* will be encountered in expectation. To arrive at this generalized temporal abstraction, states are no longer assumed to be one-hot encoded; instead, they can be represented by arbitrary representations (e.g., real-value vectors representing the visual features of a movie scene). In an SR, each entry encodes the identity of a state; in an SF, each entry encodes a latent feature. Crucially, these state representations are value-predictive just like in SR. Concretely, for an arbitrary state s with a basis function $\boldsymbol{\phi}_s$, assume that there exists a real-valued vector $\mathbf{w}$ such that

$$\mathcal{R}(s) = \boldsymbol{\phi}_s'\mathbf{w}.$$

The successor feature $\boldsymbol{\psi}_s^\pi$ of state s under policy π is defined as

$$\boldsymbol{\psi}_s^\pi = \mathbb{E}^\pi\left[\sum_{k=1}^{\infty}\gamma^k\boldsymbol{\phi}_{t+k} \mid S_t = s\right],$$

such that the state value of s can be computed as

$$V^\pi(s) = (\boldsymbol{\psi}_s^\pi)'\mathbf{w}.$$

The SF obtained from some policy π – denoted as $\boldsymbol{\Psi}^{\pi}$ – satisfies

$$\boldsymbol{\psi}_s^{\pi} = \boldsymbol{\Psi}^{\pi} \boldsymbol{\phi}_s.$$

When the stochastic one-step transition function $\mathbf{P}^{\pi}$ *over the latent feature space* is known, the SF $\boldsymbol{\Psi}^{\pi}$ can be analytically obtained:

$$\boldsymbol{\Psi}_{\gamma}^{\pi} = \left(\mathbf{I} - \gamma \mathbf{P}^{\pi}\right)^{-1} \mathbf{P}^{\pi}. \tag{2.5}$$

In reality, the latent transition dynamics is rarely accessible, so the agent has to estimate $\boldsymbol{\Psi}$ using its own experience. Again, using TD-learning, the SF can be iteratively learned as

$$\boldsymbol{\Psi}_t^{\pi} \leftarrow \boldsymbol{\Psi}_{t-1}^{\pi} + \alpha \boldsymbol{\xi}_t (\boldsymbol{\phi}'_{S_t} + \gamma \boldsymbol{\phi}'_{S_t} \boldsymbol{\Psi}_{t-1}^{\pi} - \boldsymbol{\phi}'_{S_{t-1}} \boldsymbol{\Psi}_{t-1}^{\pi}) \tag{2.6}$$

after experiencing the transition $S_{t-1} \to S_t$. The eligibility trace $\boldsymbol{\xi}$ at each time step t is

$$\boldsymbol{\xi}_t = \gamma \lambda \boldsymbol{\xi}_{t-1} + \boldsymbol{\phi}_{S_{t-1}}. \tag{2.7}$$

This algorithm is similar to the SF-learning algorithm (Lehnert & Littman, 2019) and the linear TD-PF algorithm (Bailey & Mattar, 2022). Specifically, setting $\lambda = 0$ (i.e., learning with TD(0)) reduces Eq. 2.6 to the SF-learning algorithm. The representation learned here is the transpose of the result of TD-PF, which learns the *predecessor features* instead. Furthermore, SF is a strict generalization of SR: if $\boldsymbol{\phi}$ is one-hot encoded, Eq. 2.6 is reduced to Eq. 2.4, and so $\boldsymbol{\Psi} = \mathbf{M}$.

2.4 Temporal Context Model

This section provides a formal description of TCM, a phenomenological mechanistic model of human episodic memory introduced in Section 1.2. It has four core elements: state representations $\mathbf{x}$, temporal contexts $\mathbf{c}$, and two associative matrices $\mathbf{M}^{\mathrm{CS}}, \mathbf{M}^{\mathrm{SC}}$. $\mathbf{M}^{\mathrm{CS}}$ and $\mathbf{M}^{\mathrm{SC}}$ are commonly initialized as the zero matrix and the identity matrix

respectively. During the encoding phase, TCM updates the temporal context based on the encountered state and stores the associations between each pair of context and state. During the retrieval phase, it again evolves the temporal context based on its own recalls, which are driven by the learned associations. All models derived from TCM inherit these elements.

To make the different phases explicit, we use $t \in \{1, 2, \ldots, T\}$ to index encoding time and $i \in \{1, 2, \ldots, N\}$ to index retrieval steps in TCM. Let $\mathbf{x}(S_t)$ be the feature vector of the state encoded at time t, e.g., $\mathbf{x}(S_t) = \mathbf{1}_{S_t}$, the one-hot encoded state 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 $\mathbf{x}_i$. For context vectors, we use $\mathbf{c}_t$ and $\mathbf{c}_i$ to indicate the drifting experimental contexts during encoding and retrieval respectively. $\mathbf{c}_t^{\text{IN}}$ and $\mathbf{c}_i^{\text{IN}}$ are the contexts specifically associated with the state experienced at a particular time step, either due to external input (former) or self-initiated recall (latter). The learning and update rules during encoding are summarized in Tab. 2.1 (replacing all t with i gives the corresponding rules during retrieval).

Table 2.1: Summary of the Temporal Context Model

Name	**Expression**	
Context-to-Feature Matrix	$\mathbf{M}^{\text{CS}} = \sum_t \mathbf{x}_t \mathbf{c}_t'$	(2.8)
Input Context	$\mathbf{c}_t^{\text{IN}} = \mathbf{M}^{\text{SC}} \mathbf{x}_t$	(2.9)
Context Update	$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{c}_t^{\text{IN}}$	(1)
Feature Retrieval	$\mathbf{x}_i = \mathbf{M}^{\text{CS}} \mathbf{c}_i$	(2.10)

TCM posits that when a state S_t is experienced either in encoding or retrieval, the following sequence of events take place in order: first, presenting $\mathbf{x}_t$ evokes its associated context $\mathbf{c}_t^{\text{IN}}$ via the stimulus-to-context matrix (Eq. 2.9). If the stimulus is unique, $\mathbf{c}_t^{\text{IN}}$ is equivalent to the stimulus' pre-experimental context; if the stimulus is repeated, $\mathbf{c}_t^{\text{IN}}$ also contains the (weighted) experimental context where it was previously experienced. Next,

the retrieved context updates the current context $\mathbf{c}_t$ (Eq. 1.1). Note that $\rho \in [0, 1)$ and $\beta \in (0, 1]$ are chosen so that $\mathbf{c}_t$ remains a unit vector. ρ and β may be different during encoding and retrieval (i.e., $\rho_{\text{enc}}, \beta_{\text{enc}}$ during encoding; $\rho_{\text{rec}}, \beta_{\text{rec}}$ for recall). Finally, $\mathbf{M}^{\text{CS}}$ and $\mathbf{M}^{\text{SC}}$ are updated as needed and the above sequence ensues. If Hebbian learning is assumed, for instance, $\mathbf{M}^{\text{CS}}$ at time t during encoding is updated by the outer product of the recently encoded stimulus $\mathbf{x}_t$ and its temporal context $\mathbf{c}_t$ as shown in Eq. 2.8.

As a concrete example, consider the special case where states correspond to n unique one-hot encoded words $\{\mathbf{x}_i\}_{1 \leq i \leq n}$ and $\mathbf{M}^{\text{SC}}$ remains the identity matrix. It follows that $\mathbf{c}^{\text{IN}} = \mathbf{x}$. i.e. the associated context of a stimulus is exactly its corresponding features. Consequently, $\mathbf{M}^{\text{CS}} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'$, and Eq. 1.1 at retrieval is reduced to $\mathbf{c}_i = \rho \mathbf{c}_{i-1} + \beta \mathbf{x}_i = \sum_{j=1}^{i} \beta \rho^{i-j} \mathbf{x}_j$, which is a linear combination of previously recalled stimuli.

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

2.5 The Successor Representation in the Temporal Context Model

Gershman et al. (2012) showed that if stimuli are unique and presented only once during encoding, there exists a choice of β_{enc} such that the learning of $\mathbf{M}^{\text{CS}}$ according to Eq. 2.8 and the transpose of the SR over the state space of the presented stimuli. i.e., $\mathbf{M}^{\text{CS}} = \mathbf{M}'$.

To see this, note that because of the uniqueness assumption, the prediction error is always zero, and so Eq. 2.4 is reduced to $\mathbf{M}_t \leftarrow \mathbf{M}_{t-1} + \alpha \mathbf{c}_t \mathbf{x}_t'$ – or, equivalently, $\mathbf{M}_t' \leftarrow \mathbf{M}_{t-1}' + \alpha \mathbf{x}_t \mathbf{c}_t'$; this is exactly the Hebbian learning rule for $\mathbf{M}^{\text{CS}}$ in Eq. 1.2 or Eq. 2.8 if $\alpha = 1$. Note that the temporal context of TCM functions as the eligibility trace in TD(λ), where β_{enc} controls the degree of trace decay, and ρ_{enc} is the discount factor of the learned SR.

If the visited states are not unique, Eq. 2.8 predicts that context-to-stimulus

association will grow without bound, whereas Eq. 2.4 avoids this issue while maintaining the same functional form in the case of unique stimuli (Gershman et al., 2012). Alternatively, if the Hebbian learning rule of Eq. 2.8 incorporates an adaptive learning rate, i.e.,

$$\mathbf{M}_{t+1}^{\mathrm{CS}} \leftarrow \mathbf{M}_{t}^{\mathrm{CS}} + \alpha \left(\mathbf{x}_{t+1} \mathbf{c}_{t}' - \mathbf{M}_{t}^{\mathrm{CS}} \mathbf{x}_{t} \mathbf{x}_{t}' \right) \tag{2.11}$$

where α_t decreases over time, the result converges to the SR in the limit. In other words, Hebbian learning with decay leads to the same learning outcome as TD-learning. Since the current goal is to develop an *algorithmic* (rather than implementational) theory of memory-informed decision making and control, we do not commit to any particular implementation of learning rules in the current dissertation, acknowledging that in theory the results can be attained through more than one means.

Moreover, if the one-hot encoding assumption of $\mathbf{x}$ is relaxed to an arbitrary representation $\boldsymbol{\phi}$, the associative matrix converges to the true SF in the limit ($\mathbf{M}^{\mathrm{CS}} \to \boldsymbol{\Psi}$). To see this, note that the eligibility trace in Eq. 2.7 is analogously defined as in Eq. 2.3. Thus there again exists a choice of β_{enc} such that the temporal context is equal to $\boldsymbol{\xi}$ up to a scaling factor. In this case, the drifting temporal context assigns eligibility to the latent features as opposed to the observable states.

It is worth noting that TCM by itself offers poor structured knowledge over a long timescale: it does not structure experiences but instead learns a single representation (SR/SF) over the entire sequence of observations. While TCM is unlikely a complete account of complex event cognition and adaptive behavior in temporally extended settings, it is a promising candidate for acquiring event-specific knowledge. It also allows unsupervised option discovery, where "options" are sequences of actions that extend over multiple timesteps. This is made possible by the temporal abstraction of SR and SF (Machado et al., 2023).

2.6 Latent Cause Inference

Given TCM is limited in accounting for complex adaptive behavior, especially when multiple (and possibly ever-growing) events/tasks are present, additional mechanisms are

required. One way this may be achieved is by clustering experiences into discrete events based on semantic themes (Griffiths & Tenenbaum, 2006; Howard & Kahana, 2002b) or more generally, temporal statistics (e.g., Schapiro et al., 2013), such that different events or tasks learn divergent representations that facilitate episodic evaluation with respect to the event-specific environment dynamics. Latent cause inference (LCI) offers an algorithmic explanation how such clustering may happen in humans and animals (Gershman & Hartley, 2015; Gershman & Niv, 2012). Like the example in Section 1.3, LCI assumes that the agent segments their experience into "states" of the world, where each state captures regularities and predictable variabilities in its observations. If the prediction significantly deviates from the actual observation, the agent attempts to attribute the new observation to a different latent cause, which may be a previously experienced state or a completely new one.

To start with, we assume a generative model of events, captured by a hidden Markov model (HMM; Fig. 5.1a) – the observation $\boldsymbol{\phi}_t$ at time t is generated by event K_t, which is determined by a sticky Chinese restaurant process (sticky-CRP; Fox et al., 2009; Gershman et al., 2014). Specifically, sticky-CRP generates an event based on the past event frequency, identity of the most recent event, and some fixed probability that a new event starts. Formally, K_t is sampled from all available event k's according to

$$P\left(K_t = k|\mathbf{K}_{1:t-1}\right) \propto \begin{cases} \kappa C_k + \lambda \mathbb{1}\{K_{t-1} = k\} & \text{if } k \leq NK_t \\ \alpha & \text{if } k = NK_t + 1, \end{cases} \tag{2.12}$$

where $\mathbf{K}_{1:t-1}$ is the past event sequence, C_k is indicates the number of observations generated by event k in the past, $\mathbb{1}\{K_{t-1} = k\} = 1$ if the most recent event is k and 0 otherwise, and N_t is the number of unique events up to time t. κ modulates the past event frequency, λ modulates the temporal autocorrelation, and α determines degree of generalization (i.e., how often new events germinate). Intuitively, if κ is relatively large, high-frequency events tend to reoccur; if λ is relatively large, adjacent observations probably belong to the same event. Sticky-CRP has been used previously as the Bayesian prior to model nonparametric clustering (Fox et al., 2009) and event cognition (Franklin

et al., 2020), so is the case for our model.

Maximum a posteriori (MAP) estimates are used to infer the latent event, which is subsequently used to organize inter-event representations and learn intra-event dynamics. However, exact computation requires inference over all possible event assignments, corresponding to a discrete combinatorial space that's typically intractable. Previous works in temporal clustering (Collins & Frank, 2013; Franklin et al., 2020; L. Wang & Dunson, 2011) thus approximate the posterior distribution with a local MAP estimate, specifically the greedy segmentation that is the highest in probability:

$$P(K_t|\mathbf{\Phi}_{1:t}) = \sum_{\hat{\mathbf{K}}_{1:t-1}} P(K_t|\mathbf{\Phi}_{1:t}, \hat{\mathbf{K}}_{1:t-1}) \approx P(K_t|\mathbf{\Phi}_{1:t}, \hat{\mathbf{K}}_{1:t-1}),$$

where $\mathbf{\Phi}_{1:t}$ is the sequence of past observations. Applying Bayes' rule to the equation above, the posterior can be computed as

$$P(K_t|\mathbf{\Phi}_{1:t}) \propto P(\boldsymbol{\phi}_t|\mathbf{\Phi}_{1:t-1}, K_t)P(K_t|\hat{\mathbf{K}}_{1:t-1}). \tag{2.13}$$

Here, we make an explicit distinction between the model's inferred event $\hat{K}_t$ and the ground truth K_t. The inferred event at time t is therefore

$$\hat{K}_t = \underset{k}{\operatorname{argmax}}\, P(K_t = k|\mathbf{\Phi}_{1:t}, \hat{\mathbf{K}}_{1:t-1}).$$

Note that in practice, the whole history sequence $\mathbf{\Phi}_{1:t}$ may be too long to perform efficient inference; instead, only the most recent observations within a fixed window of size w_{hist} is considered. This aligns with our experience as observations tend to be smooth and autocorrelated, where distant experiences tend to be less informative about the present (e.g., predicting the next observation). To further alleviate the computational burden, the full posterior distribution is only computed at time t when the prediction error exceeds $\sum_{i=1}^{w_{\text{hist}}} pe_{t-i}/w_{\text{hist}} + pe_{\text{thres}}$, where $\sum_{i=1}^{w_{\text{hist}}} pe_{t-i}/w_{\text{hist}}$ is the simple moving average of the previous w_{hist} points, and pe_{thres} is the error threshold. In other words, it only considers a possible event boundary if the prediction error is sufficiently large given the recent history

of predictions. Otherwise, only the posterior probability of the currently active event is computed and updated.

Chapter 3

TCM-SR

Episodic retrieval for model-based evaluation in sequential decision tasks[1]

In this chapter, 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).

As discussed in Section 1.3, many decisions benefit from the recall of one-off events, but most of the work coming from the decision making perspective lacks elaborate explanations about the role of episodic memory in model-based evaluation. While these algorithms show clear improvements in performance in sequential decision settings, they also abstract away the most intriguing aspects of episodic memory, namely the temporal dynamics involving previous memory retrieval. For instance, prior models often treat episodic memory as a lookup table and perform "episodic recall" by retrieving multiple values independently Lengyel and Dayan, 2007. Yet episodic recall in humans show clear temporal dependencies, as manifested by the temporal contiguity effect. 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 and Kahana, 2002a). TCM is a descriptive (phenomenological) model that reproduces various patterns of episodic retrieval in tasks like list-list learning. In the past

[1]This chapter is based on the following paper:
Zhou, C. Y., Talmi, D., Daw, N. D., & Mattar, M. G. (in press). Episodic retrieval for model-based evaluation in sequential decision tasks. *Psychological Review*.

two 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 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.

3.1 Results

3.1.1 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 a real-life game, in this chapter 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 Reinforcement Learning (RL; Sutton and Barto, 2018). On each trial, the agent chooses an action $A = a$

and receives the return $G = \sum_{t=1}^{\infty} \gamma^t R_t$, 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 G. This requires estimating, for each candidate action a, the expectation $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t R_t \mid A = a]$, 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 sub-problem for solving more elaborate choice tasks (e.g., Markov decision processes, in which actions can occur at every step).

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

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 — i.e., the agent understands the step-by-step transition structure of the game, a form of world model (Fig. 3.1b, 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 potentially time-consuming 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 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, $\mathbf{M} = \mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \gamma^2\mathbf{T}^3 + \cdots$, where each element M_{ij} of matrix $\mathbf{M}$ represents the discounted number of visits to state j from state i. $\mathbf{M}$ is known as Successor Representation (SR; Dayan, 1993), 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; 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 (Fig. 3.1b, 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 — i.e., $q(a) = \mathbf{x}_a'\mathbf{M}\mathbf{r}$, 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; 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 chapter, 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 rollout-like (iterative) state simulation process that differs from standard (non-iterative) uses of the SR described previously. Accordingly, we next describe the processes of memory encoding and retrieval that we will later link to value estimation.

3.1.2 Episodic retrieval via the Temporal Context Model

Our starting point is a standard model of memory encoding and retrieval, the Temporal Context Model (TCM; Howard and 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 (Fig. 3.1c). 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:

$$\mathbf{c}_t = \rho \mathbf{c}_{t-1} + \beta \mathbf{c}_t^{\text{IN}}. \tag{1.1}$$

At each moment t, the temporal context $\mathbf{c}_t$ is updated by input $\mathbf{c}_t^{\text{IN}}$, 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 $\mathbf{c}_t$ 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 $\mathbf{c}_t$ present at that moment (Fig. 3.1d-f). Formally, $\mathbf{M}^{\text{CS}} \leftarrow \mathbf{M}^{\text{CS}} + \mathbf{x}_t \mathbf{c}_t'$, where $\mathbf{x}_t$ (shortform for $\mathbf{x}(s_t)$) is the representation of stimulus s_t in terms of features, and

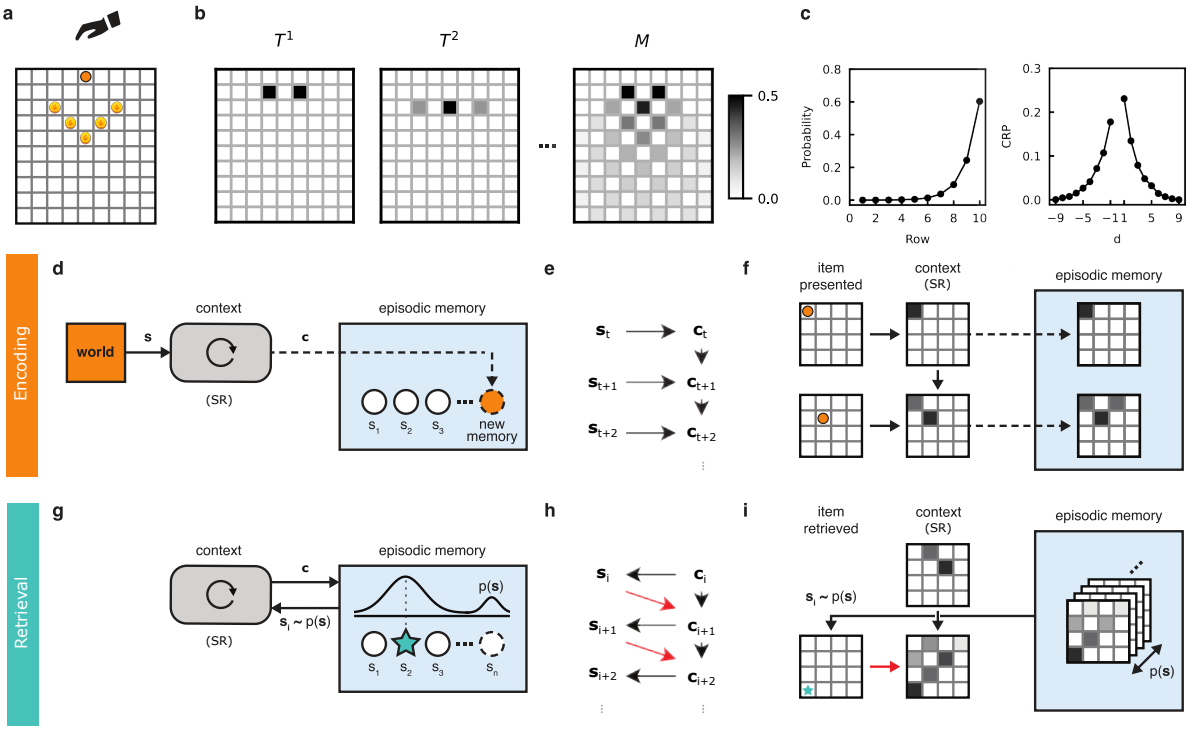

Figure 3.1: **Overview of the TCM-SR model. (a)** Our Plinko game has 10×9 states, each represented by a small square. The agent may take any of 9 possible actions, corresponding to the 9 locations on the top row where the Plinko ball (orange circle) may be dropped. The dropped ball follows a stochastic trajectory down the board, collecting scattered rewards (gold coins) 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 $\mathbf{T}^1$, $\mathbf{T}^2$, and $\mathbf{T}^3$ represent the probability distribution of the ball location 1, 2, and 3 time steps after the moment depicted in (a) respectively. The Plinko board labeled $\mathbf{M}$ represents the fully-learned Successor Representation (SR), given by $\mathbf{M} = \gamma^0\mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \gamma^2\mathbf{T}^3 + \cdots$. SR values correspond to the expected number of (discounted) visitations to each state on the board, starting from the action depicted in (a). **(c)** After each full trajectory is experienced and stored in memory, the recency effect (left) predicts that stimuli from the bottom rows, which have been experienced more recently, are more likely to be retrieved ($\beta = 0.7$, no backward sampling). 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 ($\beta = 0.1$, no backward sampling). **(d-f)** Encoding phase of TCM-SR. **(d)** Presentation of stimulus s_t at time t by the external world updates the temporal context $\mathbf{c}_t$. Memory encoding amounts to storing each temporal context present when a stimulus is seen. The first time each stimulus is presented, a new memory is stored (circle with dashed outline). Each subsequent time the same stimulus is presented, the associated memory is modified (not shown). **(e)** The temporal context $\mathbf{c}_i$ defines a distribution $p(s)$ over memories. It depends on the previous temporal context $\mathbf{c}_t$ and the current state s_{t+1}, corresponding to a recency-weighted representation of the stimuli (depicted in f). **(f)** Schematic of encoding two consecutive stimuli in the Plinko task. Stored memory of each stimulus (right box) includes a composite representation of temporal contexts present during each of the encoding situations. Dashed arrows indicate accumulative change to the stored episodic memory $\mathbf{M}^{\mathrm{SC}}$. **(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 $\mathbf{c}_{i+1}$ depends on the previous temporal context $\mathbf{c}_i$ and the retrieved stimulus s_{i+1}, which itself depends on the previous context $\mathbf{c}_i$. The red arrow illustrates how the temporal context is affected by each retrieved stimulus. **(i)** Schematic of retrieving a stimulus in the Plinko task. The temporal context is updated by a retrieved context, whose associated stimulus is sampled using the stored episodic memory $\mathbf{M}^{\mathrm{CS}}$(Eq. 3.7).

$\mathbf{M}^{\mathrm{CS}}$ 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 $p(s_k) \propto \mathbf{M}^{\mathrm{CS}}\mathbf{c}_i \cdot \mathbf{x}_k$ (Fig. 3.1g-i). Retrieval is thus determined by the current context and the agent's memory (Fig. 3.1h,i), i.e., 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 Eq. 1.1, which in turn affects the retrieval of subsequent items (Fig. 3.1g-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 (Fig. 3.1c). 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.

3.1.3 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_t 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 $S_1, S_2, \cdots, S_H$ 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, amounts 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, i.e., $\mathbf{M}^{\mathrm{CS}} \rightarrow \mathbf{M}'$ (see Gershman et al., 2012 or the Methods section for a formal demonstration; also, notice how the episodic memory representations in Fig. 3.1f share characteristics with $\mathbf{M}$ in Fig. 3.1b). This crucial equivalence is the basis for the remainder of this chapter, 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 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 where components deemed important are disabled or removed to investigate its effect). 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.

3.1.4 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, i.e., $\mathbf{M}^{\text{CS}} = \mathbf{M}'$.

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 Eq. 1.1, leading to $\mathbf{c}_i = \mathbf{c}_{i-1}$ (this is equivalent to removing the red arrows in Fig. 3.1g-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, i.e., they are i.i.d. (independently and identically distributed) 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 (Fig. 3.2a). 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:

$$\hat{q}_{\beta=0}(a) \propto \frac{1}{N} \sum_{i=1}^{N} \mathbf{r}'\mathbf{x}(S_i), \tag{3.1}$$

where $S_1, S_2, \ldots, S_N \sim p(s)$ are samples from the normalized SR, i.e., $p(s) = \frac{\mathbf{x}_a'\mathbf{M}}{|\mathbf{x}_a'\mathbf{M}|}\mathbf{x}(s)$. In this equation, $r(S_i) = \mathbf{r}'\mathbf{x}(S_i)$ is the reward present in state S_i. Thus, $\hat{q}(a)$ is obtained

by averaging reward samples (see Lemma 1 in Methods).

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 $p(s) \propto \left[\mathbf{M}^{\mathrm{CS}}\mathbf{c}\right] \cdot \mathbf{x}(s)$. 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., $\mathbf{M}^{\mathrm{CS}} = \mathbf{M}'$), TCM predicts that the probability of retrieving an item s is given by $p(s) \propto [\mathbf{M}'\mathbf{x}_a] \cdot \mathbf{x}(s) = [\mathbf{x}_a'\mathbf{M}] \cdot \mathbf{x}(s)$. For a one-hot representation of the current context $\mathbf{x}_a$, the first term $\mathbf{x}_a'\mathbf{M}$ 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 (Fig. 3.2a). Eq. 3.1 shows that the average reward across all sampled states is a proxy for the action value, as we originally defined it. Repeating such retrieval-based evaluation for each candidate action can thus inform the agent to select the highest-valued action. Note this procedure is not derived from normative considerations (i.e., what memories an agent ought to retrieve); rather, it is a direct prediction of TCM: given the assumptions in place, TCM predicts i.i.d. sampling from the SR, retrieving states whose average reward is the normative action value (see Theorem 2 in Methods). 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 $\beta_{\texttt{enc}}$ 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 occuring soon after the starting context (note that we distinguish the drift rate at encoding, $\beta_{\texttt{enc}}$, 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 $\gamma = 1 - \beta_{\texttt{enc}}$.

By affecting the temporal discount factor, the drift rate at encoding ultimately affect the overall value estimated during retrieval. Depending on the discount factor, the computed value ranges between (i) rewards sampled exclusively from imminent states ($\gamma = 0$, Fig. 3.2a,b), and (ii) rewards sampled from all future states, with a preference for earlier states ($\gamma > 0$, Fig. 3.2d,e). Notably, the former case ($\gamma = 0$) implements the evaluation required for bandit problems, in which action values depend only on instantaneous rewards. Indeed, a special case of the current model corresponds to a class of *decision-by-sampling* models that have been previously described and empirically tested in single-step problems like bandits (e.g., 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 MCTS (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 tree-like set of future situations in a sequential task to a set of individual future states weighted by their prevalence in the tree. Consequently, it transforms 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), it can be identified with many fewer samples than needed to estimate all action values accurately. Second, if no action value dominates the others, identifying the optimal action requires a large number of samples, but the extra computation will not lead to

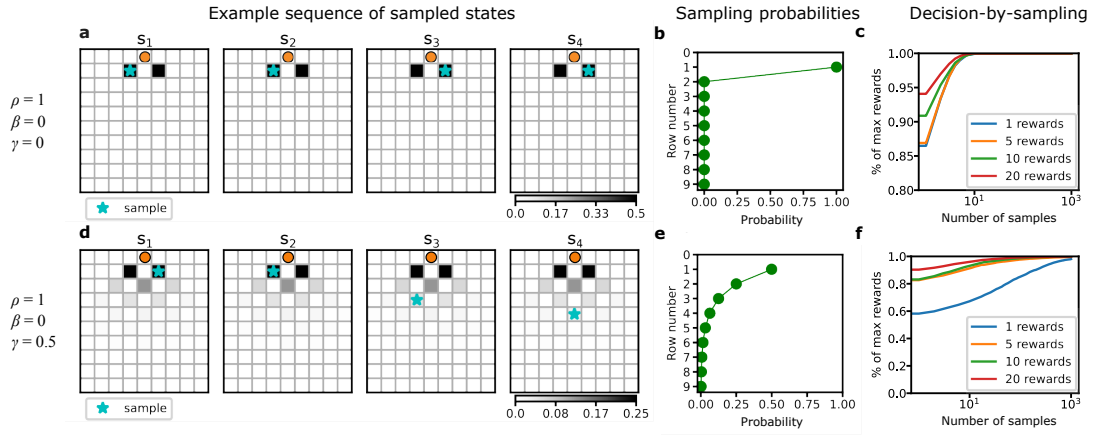

Figure 3.2: **Independent samples from memory yield unbiased value estimates. (a-c)** Sampling from a distribution with a short temporal horizon. Parameters: $\rho = 1$, $\beta = 0$, $\gamma = 0$. **(a)** An example of querying an action (orange circle) through memory recall (cyan stars). s_i shows the i^{th} stimulus sampled, where the same state can be sampled multiple times. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board. **(c)** We simulate an agent who evaluates two actions (at top-center state and the state immediately adjacent to the right) using the procedure from Eq. 3.1, 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 row (inclusive) and are reachable from either action. See the online article for the color version of this figure.

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; Fig. 3.2c). 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 (Fig. 3.2d). 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 (Fig. 3.2f), 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 recapituate them by retrieval at 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.

3.1.5 The contiguity effect enables value estimation via rollouts

The previous section considered a simplified setting in which the retrieval of a stimulus does not affect subsequent retrievals, giving rise to i.i.d. samples that the agent could average to obtain action values estimates. However, a prominent feature

of episodic memory is that consecutive retrievals are *not* independent. Indeed, the simplifying assumptions from the previous section eliminate the model's ability to explain the contiguity effect, ubiquitous in list learning experiments. Thus, we now consider a different parameter regime of TCM, in which stimulus retrieval *does* affect subsequent retrievals.

We focus initially on the extreme case where retrieval depends only on the immediately preceding retrieved stimulus — i.e., we set $\beta = 1$ and $\rho = 0$ in Eq. 1.1 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, i.e., $\mathbf{c}_i = \mathbf{x}_i$, 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 $p(s) \propto \left[\mathbf{M}^{\text{CS}}\mathbf{c}_i\right] \cdot \mathbf{x}(s)$. With the assumptions that $\mathbf{c}_i = \mathbf{x}_i$ and that $\mathbf{M}^{\text{CS}} = \mathbf{M}'$, this distribution is simplified to $p(s) \propto [\mathbf{M}'\mathbf{x}_i] \cdot \mathbf{x}(s) = [\mathbf{x}_i'\mathbf{M}] \cdot \mathbf{x}(s)$. 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 of $\gamma = 0$, the first retrieved memory is an immediate successor of the considered action (because $\mathbf{M} = \mathbf{T}^1 + \gamma^1\mathbf{T}^2 + \cdots = \mathbf{T}^1$ when $\gamma = 0$, Fig. 3.1b). Upon retrieving the first memory and updating the temporal context, the second retrieved memory is an immediate successor of the first sample (Fig. 3.3a). Repeating this sampling process recursively leads to a rollout (in Plinko, this process amounts to a simulation of a

trajectory through which the ball might plausibly fall; Fig. 3.3a,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 (Fig. 3.3d). 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 Fig. 3.3d and Fig. 3.1c, right). The original TCM predicts the bilateral contiguity effect because it assumes that the information retrieved from episodic memory, $\mathbf{c}_i^{\text{IN}}$, includes not only the pre-experimental 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 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 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:

$$\hat{q}_{\beta=1}(a) \propto \sum_{i=1}^{N} \mathbf{r}'\mathbf{x}(S_i), \tag{3.2}$$

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

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

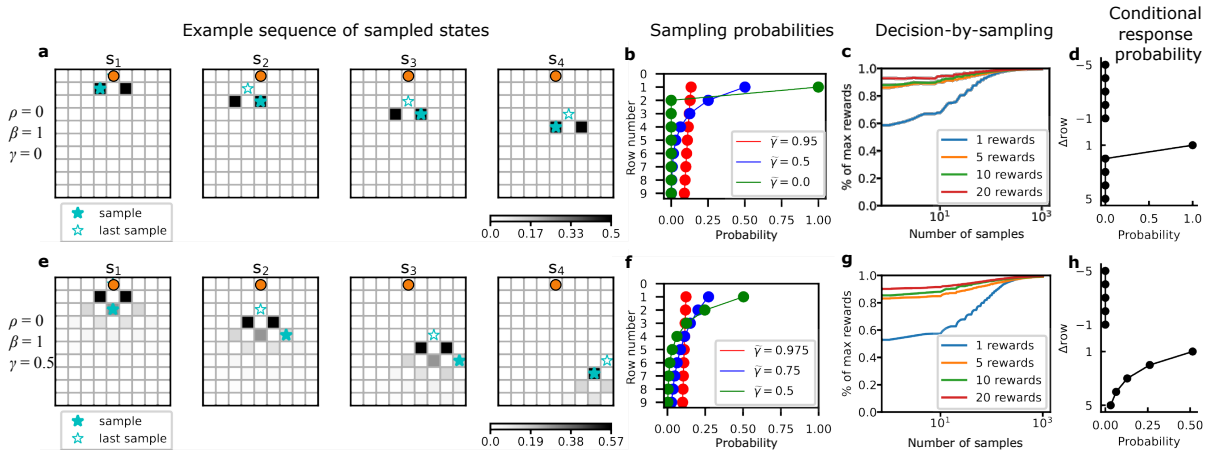

Figure 3.3: **Recall-dependent context updates lead to rollouts. (a-d)** Sampling from a distribution with a short temporal horizon. Parameters: $\rho = 0$, $\beta = 1$, $\gamma = 0$. **(a)** An example sequence of memory retrieved when initiating the temporal context as the top-center state (orange circle) through memory recall (cyan stars). $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board. We illustrate these distributions for three values of p_{stop} (0.05, 0.5, and 1), each leading to an effective temporal discount factor $\tilde{\gamma} = 1 - p_{\text{stop}}$. **(c)** We simulate an agent that evaluates two actions (at top-center state and the state immediately adjacent to the right) using the procedure from Eq. 3.2, and then selects the action with the larger estimated value. Rewards are uniformly placed between the second and the seventh row (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 $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d)** Probability that a sample is drawn from each row of the Plinko board, as a function of the distance to the previously sampled row. **(e-h)** As in **a-d**, but using parameters: $\rho = 0$, $\beta = 1$, $\gamma = 0.5$. Rewards are uniformly placed between the second and the seventh row (inclusive) and are reachable from either action. See the online article for the color version of this figure.

In Eq. 3.2, 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 theorem 4 in Methods). In other words, the rollouts during retrieval lead to an effective temporal context (denoted $\tilde{\gamma}$) of one,

in stark contrast to the temporal context 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 temporal context obtained during retrieval was always identical to the temporal context 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 Eq. 3.2 amounts to concatenating n one-step predictions (i.e., $\gamma = 0$), which is equivalent to performing a single n-step prediction (i.e., $\tilde{\gamma} = 1$).

Strictly speaking, however, an effective discount factor of $\tilde{\gamma} = 1$ would implausibly require each rollout to continue forever. In practice, a rollout that ends at a certain point includes all rewards sampled prior to 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 Methods for details, esp. 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, i.e. 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; Fig. 3.3c). 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 (Fig. 3.3e-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 Eq. 3.2. 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 Methods).

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

(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 $p_{\text{stop}} \to 0$, a rollout continues indefinitely, resulting in an effective discount factor of $\tilde{\gamma} \to 1$. 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; Fig. 3.3g). The efficiency of the generalized rollout is slightly lower than the efficiency of the strict rollout (compare Fig. 3.3c and Fig. 3.3g). This is because, for a pre-specified 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, represent two distinct modes of operation of episodic memory, which TCM formalized as extreme settings of the parameter space. Next, we consider intermediate, more general — and

likely more realistic — settings.

3.1.6 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 Eq. 1.1), 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:

$$\hat{q}_\beta(a) \propto \beta \sum_{i=1}^{N} \mathbf{r}'\mathbf{x}(S_i), \tag{3.3}$$

where $\beta > 0$ and $S_1, S_2, \ldots, S_N \sim p(s)$ are samples from the normalized SR $p(s)$ corresponding to some effective discount factor $\tilde{\gamma}$. Note that this estimator is only unbiased given an infinite number of samples (see Theorem 6 in Methods) and otherwise an underestimate of the true value (see Proposition 7.1 in Methods); however, a relatively large number of samples is sufficient for an estimate that's close to the truth (Fig. 3.4c,f).

The same insights gained in the previous sections apply here, including extension of the effective discount factor with a larger β (Fig. 3.4b,e) and the sample efficiency during decision making (Fig. 3.4c,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 Fig. 3.4b,e) no longer capture the empirical sampling distributions under the same p_{stop} unless $p_{\text{stop}} = 1$ (dots in Fig. 3.4b,e). Recall that the larger the β, the further into the future later samples reach: i.e., β controls the degree to which the timescale at retrieval is extended (Fig. 3.4a,d). Thus both increasing β and decreasing the interruption probability extend the agent's effective temporal horizon for action evaluation, with the exception that the resultant sampling distribution may not correspond to any specific $\tilde{\gamma}$ as it is not necessarily an exponential distribution (e.g. red dots in Fig. 3.4b).

Similar to generalized rollouts, a non-zero discount factor during retrieval results in slower convergence due to state skipping (Fig. 3.4c vs f). However, unlike generalized rollouts (Fig. 3.3c vs g), the difference between zero and non-zero 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 vs constructive, rollout-based simulation, allowing it to balance the costs and benefits of these evaluation regimes depending on circumstances. This is similar to other examples in which, it has been argued, the brain adjusts its decision computations due to similar cost-benefit tradeoffs (Daw 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). The next section extends our predictions to settings when only limited experience is available.

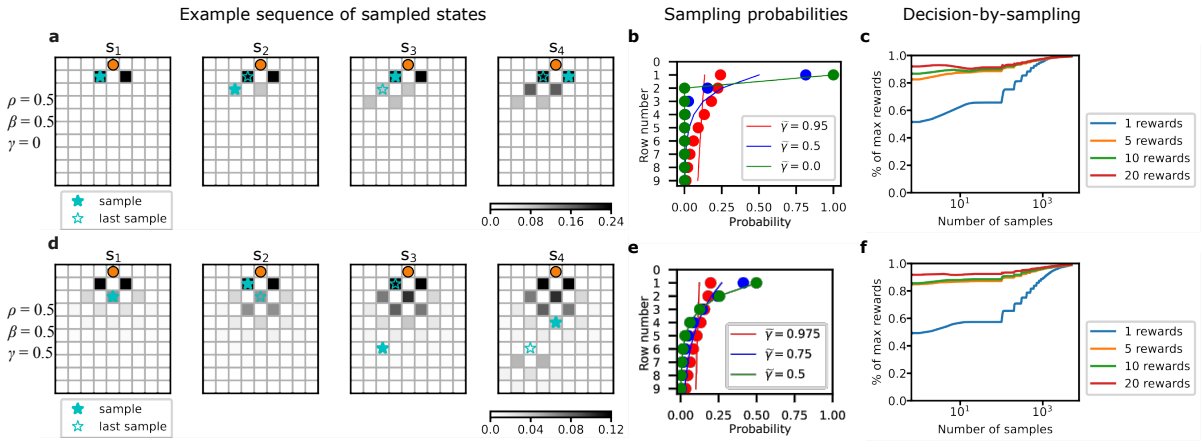

Figure 3.4: **An intermediate regime between i.i.d. sampling and rollouts. (a-c)** Parameters: $\rho = 0.5$, $\beta = 0.5$, $\gamma = 0$. **(a)** An example sequence of memory retrieved when initiating the temporal context as the top-center state (orange circle) of a Plinko board. $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Probability that a sample is drawn from each row of the Plinko board, in this intermediate sampling regime (dots) versus generalized rollout (lines, same as Fig. 3.3b) given the same discount factors. We illustrate these distributions for three values of p_{stop} (0.05, 0.5, and 1), each leading to an effective temporal discount factor $\tilde{\gamma} = 1 - p_{\text{stop}}$. **(c)** We simulate an agent that evaluates two actions (at top-center state and the state immediately adjacent to the right), and then selects the action with the larger estimated value. Rewards are uniformly placed between the second and the seventh row (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 $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d-f)** As in **a-c**, but using parameters: $\rho = 0.5$, $\beta = 0.5$, $\gamma = 0.5$. See the online article for the color version of this figure.

3.1.7 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 temporal difference learning; Gershman et al., 2012). This forms associations corresponding to the sequential contingencies experienced by the agent. If this encoding is followed by TCM retrieval, only states along the experienced trajectory will be retrieved (Fig. 3.5a, "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 trajectories should be encoded in the learned SR. Here, states belonging to either trajectory can be retrieved, with dynamics again depending on the degree of context updating (Fig. 3.5a, "Trial 2"). The learned SR comes to represent a composite of possible trajectories as experiences expand, eventually

converging to the steady-state SR (Fig. 3.5a, 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. e.g. given trajectories ABDE and ACDF, our model predicts that rollouts along ABDF or ACDE are possible. For more general parameter settings, our model predicts state-skipping (if $\gamma > 0$) or backward jumping (if $\beta < 1$). Furthermore, states in the beginning of an experienced trajectory (predictions of the near-future vs. distant-future) are prioritized for retrieval due to discount factor. These differences result from the critical assumption of our model that agents retrieve individual states, rather than trajectories.

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

3.1.8 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 signalled 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 (Yonelinas & Ritchey, 2015). 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 (C. Y. 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 vs. 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 (Fig. 3.2c,f; Fig. 3.3c,g; Fig. 3.4c,f). While performance can be improved by drawing more samples, this longer deliberation can be costly in terms of time and effort.

A potentially more effective way to find the best action might be to bias sampling toward the most relevant states (here, the goal), even if biasing the sampling procedure

might lead to biases of the estimated payoff $q(a)$ (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 under-estimation of true action values. When most samples come from the smaller set of "important" states, samples are less varied, resulting in lower *estimation variance*. Consequently, fewer samples are required to be reasonably precise and fewer retrievals are needed to arbitrate between competing actions. Nevertheless, the eventual decision can be suboptimal, in the sense that the action selected may not be the one associated with most reward. The larger the retrieval preference towards emotionally-salient stimuli, the larger the estimation bias and smaller the variance — thus, a bias-variance trade-off. A similar observation has been previously made in bandit settings (Lieder et al., 2018). Here, we extend this class of Monte Carlo models to sequential tasks, and show that the same observation applies. The main contribution of this section is that TCM-SR allows us to expose how action evaluation in sequential tasks relates to episodic memory, helping to rationalize emotional memory effects.

To illustrate this effect in our Plinko environment, we follow previous modeling work and employ a higher learning rate α_{mod} to encode emotionally salient stimuli into memory according to Eq. 3.4 (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 towards the rewarded states (Fig. 3.5b). Consequently, in the Plinko game, states associated with rewards are sampled more frequently during retrieval (Fig. 3.5d, right). Without

emotional modulation, rewarded states would have been sampled only rarely (Fig. 3.5d, left). The consequences of operationalizing emotional modulation in TCM-SR such that rewarded states are encoded with a larger learning rate are threefold. First, the action value estimates no longer converges to the correct action values. Second, convergence will be faster, resulting in a bias-variance trade-off (Fig. 3.5f, compare with Fig. 3.5e). Third, if the agent selects actions according to this regime, a higher fraction of rewards can be obtained for a given number of samples (Fig. 3.5c), suggesting that biased retrieval can be more favorable, in terms of ultimately guiding choice, than unbiased retrieval.

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 non-zero 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 (Fig. 3.5c, 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.

3.1.9 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 (Fig. 3.1c). From the perspective of mental simulation, this property seems

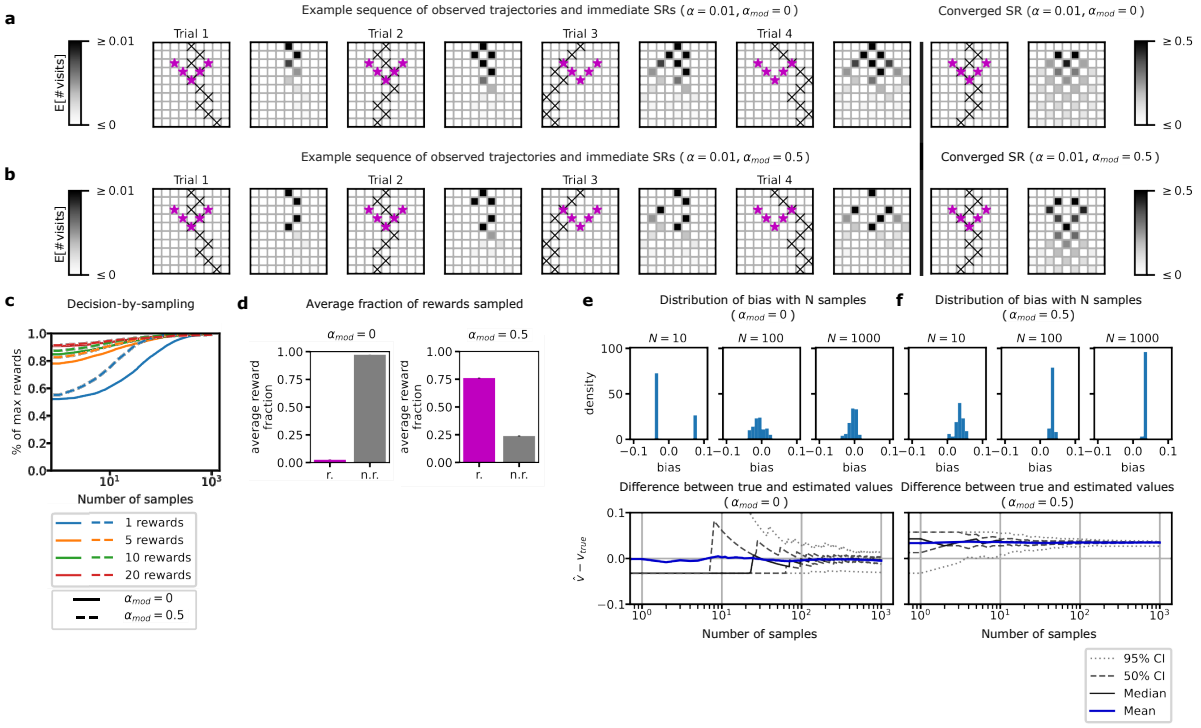

Figure 3.5: **Retrieval with limited experience and with emotional modulation. (a)** Each pair of panels represent a 'trial' where the agent observes the trajectory that follows a single action (left, each visited state denoted in x's, and each rewarded state in $\star$) and the ensuing learned SR after convergence ($\gamma = 0.9$) (right). The impact of accumulated experience is shown by comparing Trials 1, 2, 3, 4, and Trial $\to \infty$, presented in the five pairs of panels going from left to right, all without emotional modulation ($\alpha = 0.01$). SR is updated after each trial using Eq. 3.4. **(b)** The same as (a) but now with emotional modulation ($\alpha = 0.01$ for unrewarded states and $\alpha = 0.5$ for rewarded states). The same rule Eq. 3.4 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 Eq. 3.1, with (dashed lines) and without emotional modulation (solid lines). Rewards are uniformly placed between the second and the seventh row (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 $p_{\text{stop}} = 0.05$ as a function of different numbers of rewards placed on the Plinko board. **(d)** Average fraction of sampled states with (r.) and without a reward (n.r.). Error bars indicate s.e.m. across experiments. Left: no emotional modulation. Right: with emotional modulation. **(e-f)** Bias and variance convergence based on a single observation for $\gamma = 0.9$ without emotional modulation (e) and with modulation (f). Top: mean bias of estimates based on 10, 100, 1000 samples. Bottom: mean discrepancy between the true value and the estimated value as a function of number of samples on a log scale. See the online article for the color version of this figure.

counterintuitive: in our example, it corresponds to rollouts in which the Plinko ball, impossibly, runs uphill. Here we suggest that this type of reversible simulation is actually adaptive for many tasks other than Plinko.

The reason our simulations thus far reproduced only the forward contiguity (Fig. 3.3d,h) is because of one final simplification that have not yet been re-examined. We have assumed that when a memory is retrieved, it directly updates the temporal context with a static, task-independent one-hot representation of the retrieved stimulus ($\mathbf{x}_t$ in Eq. 1.1; Fig. 3.1h). In contrast, the original TCM model explains the two-sided contiguity effect by positing that context update caused by retrieving a stimulus is not static and task-independent; rather, memory retrieval updates the temporal context with a dynamic, task-dependent representation, a representation that changes each time that stimulus is experienced. In particular, TCM assumes that the temporal context is updated by a retrieved *context* associated with a given stimulus, instead of being updated by the stimulus representation $\mathbf{x}_t$ itself. Formally, the temporal context is updated during retrieval according to $\mathbf{c}_i = \rho\mathbf{c}_{i-1} + \beta\mathbf{c}_i^{\text{IN}}$, where $\mathbf{c}_i^{\text{IN}} = \mathbf{M}\mathbf{x}_i$, i.e., $\mathbf{c}_i^{\text{IN}}$ is the column of the SR indexed by the stimulus.

What might be the adaptive purpose of a bidirectional pattern of retrieval? This pattern might appear counterintuitive since an action value is determined by the expectation of *future* rewards. Indeed, in our previous simulations, action values were estimated via strictly forward-looking rollouts, i.e., in terms of future rewards alone. With a bidirectional pattern of retrieval, sampling no longer respects the temporal order of events experienced during encoding. We argue that, in most realistic tasks, the experienced temporal ordering of events is only one of all possible orderings; most state transitions experienced in one order can also be traversed in the reverse order. Although this is never the case in Plinko (since gravity strictly pulls the ball downward), it is often the case in tasks like spatial navigation. In other tasks (like chess), many actions are reversible while some others (e.g. capturing a piece) are not. An agent operating in the low-data regime can leverage this reversibility to infer, after experiencing state A followed by B (A $\rightarrow$ B), that transitioning from B to A (B $\rightarrow$ A) is likely also possible. Similarly, given only a

few experiences in an environment, the agent can infer an exponentially larger number of unexperienced but likely possible trajectories (e.g., extrapolating A $\rightarrow$ B $\rightarrow$ C to not only C $\rightarrow$ B $\rightarrow$ A, but also A $\rightarrow$ B $\rightarrow$ A, C $\rightarrow$ B $\rightarrow$ C, etc), which in turn generalizes action evaluation. Ideally, the relative strength of forward vs. reverse 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 agent can move diagonally in any direction, and it may start from any board position. The agent's goal is to select an adjacent state to move into, after which each subsequent states is selected at random from between the neighbors of the previous state. In this "reversible Plinko", the value of each state is affected by all rewards on the board, with nearby rewards contributing a higher weight to the value. If an agent only experiences top-to-bottom trajectories in the reversible Plinko task, and uses a strictly forward-looking rollout to evaluate actions, the resulting values will correspond to values under the gravity-bound Plinko rules. While they are in line with the agent's experiences, they do not match the true values under the reversible Plinko rules (Fig. 3.6d). A retrieved context aids the agent to go beyond unidirectional experience and correctly estimate the values for the reversible Plinko (Fig. 3.6c). 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.

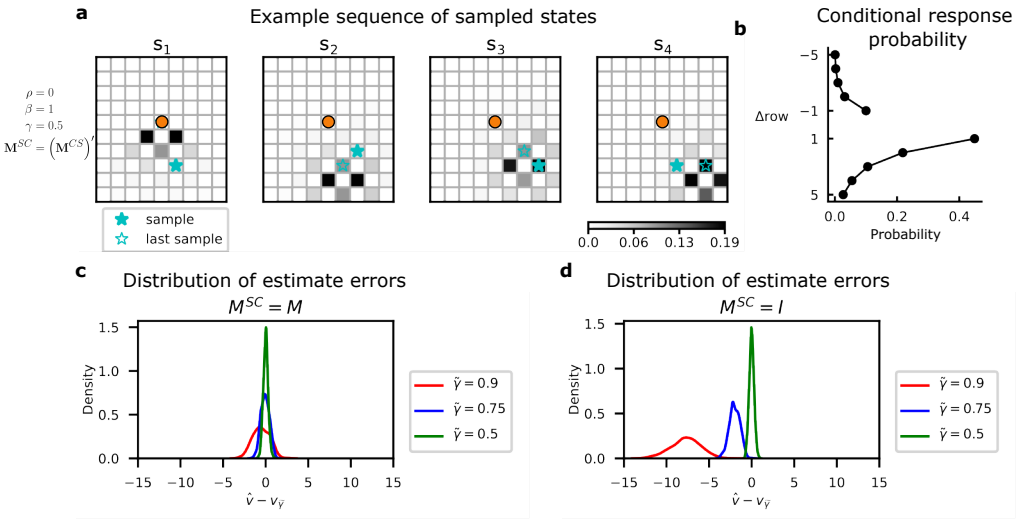

Figure 3.6: **Retrieving a learned context allows backward sampling. (a)** An example sequence of memory retrieved when initiating the temporal context with the state shown as an orange circle, and using $\gamma = 0.5$. $\mathbf{s}_i$ shows the i^{th} stimulus sampled. Greyscale colors indicate the sampling probabilities. **(b)** Contiguity curve implied by the sampled states with respect to their corresponding row number given $\gamma = 0.5$ (zero omitted). Note that both forward and backward sampling are predicted. **(c)** Distribution of estimation error using the SR as the feature-to-context association matrix. Errors are computed as the difference between the sampling-based value estimation and the ground-truth value in a reversible MDP (i.e., a grid world rather than a Plinko game). Rewards are uniformly placed between the second and the seventh row (inclusive). **(d)** As in (c), but using the identity matrix as the feature-to-context association matrix (as in the previous simulations). See the online article for the color version of this figure.

3.2 Discussion

3.2.1 Summary of Findings

In this chapter, 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 Temporal Context Model (TCM) of episodic recall and the normative concept of the Successor Representation (SR), a model of the world that is widely studied in reinforcement learning (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 evaluates then 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; Momennejad et al., 2018; 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.

3.2.2 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 multi-step 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 & Norman, 2017; Bornstein et al., 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 bandit-like evaluation by sampling (Bornstein & Norman, 2017; Bornstein et al., 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. Bandit-like 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 successor representation (SR). Prior work suggests that the SR explains numerous patterns in human behavior (Momennejad et al., 2017; 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 AI where state values are instead computed instantaneously via a dot product $\mathbf{v} = \mathbf{M}_\gamma \mathbf{r}$ (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. While we will argue that this is not an issue, we note that our goal in this chapter (or this dissertation overall) 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 aligns more with iterative sampling than with a parallel dot product formulation. The model we proposed, TCM-SR, is an important step towards 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; Russek et al., 2017). In contrast, TCM-SR can accomplish some degree of replanning by controlling the rollout length, 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, 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 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 the most 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.

Lastly, TCM-SR extends the proposal that people overweight 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) demonstrate the benefit of oversampling extreme events, they do 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 over-predicts the future occurrence of extreme events and biases any of the subsequent learning. TCM-SR also exposes

connections between a 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.

3.2.3 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. Finally, 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 regards 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 $\mathbf{c}_i^{\text{IN}} = \mathbf{x}_i$, 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 is 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 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 here is the simulation of one-shot learning, where we use a

temporal difference 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.

3.2.4 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 and 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 multi-step MDPs; Daw et al., 2005; Keramati et al., 2011; 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 non-reversible environmental dynamics — e.g., playing a card in a poker game, capturing a piece in a chess game, falling under gravity, or consuming a non-replenishable 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* — i.e., 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 (Fig. 3.5) 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, etc) to encourage more or less episodic sampling.

3.2.5 Relation to Existing Models

Temporal Context Model 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, 2019; 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 prior to 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 in 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 towards to the most recent "list" (trajectory), corresponding to the inter-list 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 inter-list effects. Moreover, TCM-SR urges us to re-consider memory intrusion from the decision-making perspective: although recalling words from non-target lists is generally regarded as a "failure" in free recall tasks, "failure" to distinguish experiences between episodes may actually be advantageous for behavior generalization: e.g., 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. Fig. 3.5a specifically illustrates how a handful of experiences may be interpolated to facilitate sampling of trajectories that do 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, 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.

Finally, the artificial task of Plinko involves limited semantic (or, in general, non-temporal) 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, 2019) 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 and Harbison, 2007; J. Miller et al., 2012; Murdock and 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 pre-specified 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 inter-response time during free recall (J. Miller et al., 2012; Murdock & Okada, 1970). 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 3-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 achieve 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 (1) individual states rather than trajectories are retrieved, and (2) retrieved samples may skip over intermediate states. Future work should investigate whether these predictions better describe how humans evaluate actions.

Episodic vs. Model-Based Evaluation

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

in rats causes their behavior to shift from model-based to model-free (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 evidence suggest a close link between episodic and model-based evaluation as a function of experience. Importantly, these considerations point to the importance of future investigation in a memory regime that has not seen much study in list learning: how episodic recall is affected by repeated exposure to lists with overlapping items (Gershman et al., 2012), analogous to the hypothetical transition from individual trajectories to an SR in our model.

Gamma-Models

The drifting temporal context in TCM-SR resembles the learning of a bootstrapped target distribution in the γ-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 γ-model.

Like γ-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 as well as 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 γ-model, showing that the two regimes explored there (sampling from a fixed γ-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., non-geometric) timestep weighting. Most notably, of course, our descriptive model is grounded in a computational model of episodic memory, which makes the parallel with the γ-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 non-neighboring 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've

established in this chapter, 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.

3.2.6 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 multi-step 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 (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 towards semantically-related recall. Finally, 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 chapter.

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 and McClelland, 2012 on REMERGE), and incorporating event boundaries (Clewett et al., 2019), where TCM-SR may be extended to offer normative explanations and 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 suggests 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 the these data. Our model could also leverage corpus statistics to represent items as non-orthogonal 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 interest 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 $\mathbf{c}^{IN}$ 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 $\mathbf{M}^{\text{FC}}$ do not incorporate task-dependent representations (e.g., predecessors, captured by $\mathbf{M}^{\text{FC}}_{\text{exp}}$ 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 $\mathbf{M}^{\text{FC}}$ should be updated to reflect $\mathbf{M}^{\text{FC}}_{\text{exp}}$ at each encoding step (i.e., $\gamma_{\text{FC}} > 0$ throughout encoding), yet Gershman et al. (2012)'s 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 non-zero proportion of $\mathbf{M}^{\text{FC}}_{\text{exp}}$ in $\mathbf{M}^{\text{FC}}$ 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.

Lastly, 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 — i.e., 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.

3.3 Methods

3.3.1 Task Details

We wish to formalize how action values in sequential decision problems can be estimated via episodic memory samples, taking into account several known properties about retrieval dynamics in free-recall. We illustrate this process with a temporally extended game called Plinko (Fig. 3.1a). This game is an analogy to a generic sequential decision task where each action leads to a stochastic sequence of states, and where each state can be reached by potentially multiple actions. We selected the game of Plinko because it allows the visual depiction of the sequential retrieval process in a didactic manner (as rows represent both time and space). The game should therefore not be interpreted literally as choices in a real game of Plinko are unlikely to be guided by episodic memory.

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

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

3.3.2 The Successor Representation in the Temporal Context Model

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

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

which is analogous to the eligibility trace defined in Eq. 2.3 by treating each feature vector as a state in an MDP.

As previously discussed in Section 2.5, Gershman et al. (2012) proved that unique stimuli results in $\mathbf{M}^{\text{CS}}$ being equal to the transposed SR $\mathbf{M}'$ over the stimuli. If we further assume that the equivalence still holds with repeated exposure, the TD-learning rule for

$\mathbf{M}^{\text{CS}}$ should be

$$\mathbf{M}^{\text{CS}}_{t+1} \leftarrow \mathbf{M}^{\text{CS}}_{t} + \alpha \left(\mathbf{x}_{t+1} + \gamma \mathbf{M}^{\text{CS}}_{t} \mathbf{x}_{t+1} - \mathbf{M}^{\text{CS}}_{t} \mathbf{x}_{t}\right) \mathbf{c}'_{t+1}. \tag{3.4}$$

Similar to Eq. 2.4, this learning procedure permits convergence in the limit of time without suffering the problem of unbounded growth in Eq. 2.8.

3.3.3 Value Computation in TCM-SR

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

$$v_{\pi}(S_0) = \mathbf{m}^{\pi}_{S_0,*;\gamma} \mathbf{r} = \sum_{S'} m^{\pi}_{S_0,S';\gamma} r(S'). \tag{3.5}$$

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

$$\widetilde{v}_{\pi}(S_0) = \frac{1}{N(1-\gamma)} \sum_{i=1}^{N} r(S_i). \tag{3.6}$$

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

However, in general, TCM draws are not i.i.d., because a non-zero β would cause

the temporal context to drift towards the most recently experienced stimulus. Subsequent recalls are therefore dependent on preceding memory samples, as manifested by the contiguity effect where subsequent recalls are biased towards successors of the previous sample. In particular, $\mathbf{x}_i$ may be obtained via

$$\mathbf{x}_i \sim \frac{1}{Z}\mathbf{M}^{\mathrm{CS}}\mathbf{c}_i, \tag{3.7}$$

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

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

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

tree structure of these two types. Simulations 1-3 demonstrate the behavior of each of these cases, and we prove the exact state value computation in the next section.

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

Finally, because SR is dependent on the behavioral policy under which it is learned, a large change in the transition structure or reward function may render the previously obtained SR fruitless. For instance, if a behavioral policy poorly represents certain state transitions around the reward location, an agent using its corresponding SR will be inflexible and perform suboptimally in transfer learning (e.g. 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 $\mathbf{M}^{\mathrm{SC}}$ to be the identity matrix - that is, the context associated with a state is exactly its feature vector. Alternatively, $\mathbf{M}^{\mathrm{SC}}$ could encode some backward transitions such as the transpose of $\mathbf{M}^{\mathrm{CS}}$, 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.

3.3.4 Simulation Details

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

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

We set $\rho = 1, \beta = 0$ to simulate the effect of a stationary context, which gave rise to independent draws of memory samples in TCM-SR. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 3.2a-c) and $\gamma = 0.5$ (Fig. 3.2d-f) during encoding, with the latter corresponding to a slower rate of temporal drift (i.e., longer timescale). The stimulus-to-context associative matrix $\mathbf{M}^{\mathrm{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments (games) were conducted for each different discount factor, with 50 trials per experiment and 1000 (independent) samples per trial (i.e., $N = 1000$). At least one reward was placed within the agent's temporal horizon. e.g., given $\gamma = 0$, row 2 contained at least one reward. The sampling distributions over rows (Fig. 3.2b,e) reflect trial averages if starting from the top-center state (marked with a orange circle in Fig. 3.2a,d).

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was computed as the average across samples and trials. Simulations were repeated for games with 1, 5, 10, 20 binary rewards

accessible from either dropping location (Fig. 3.2c,f). The number of rewards were chosen to reflect a spectrum of reward abundance ranging from a single reward to about 50%. The percentage of maximum rewards obtained of a particular game *pmr* was computed as

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

where S_{chosen} is the state selected by the deterministic policy, S^* is the state with highest expected total return, and $v(\cdot) : \mathcal{S} \mapsto \mathbb{R}$ is the state value function. Note an optimal choice implies $pmr = 1$. Fig. 3.2c,f show the average *pmr* across 100 experiments.

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

We set $\rho = 0, \beta = 1$ to simulate the effect of a context fully determined by the most recent retrieval, which gave rise to generalized rollouts in TCM-SR. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 3.3a-d) and $\gamma = 0.5$ (Fig. 3.3e-h) during encoding. For each γ, simulation were repeated using three different probabilities of interruption $p = 0.05, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s for each underlying true γ at retrieval (Fig. 3.3b,f). Thus as long as the ball had not reached the bottom of the Plinko board, at each time step, there was a p probability that the trial will terminate, regardless of the ball's location. Consequently, each trial started from the top-center state (marked with a orange circle in Fig. 3.3a,e) and ended if either the ball hit the bottom of the board or the sampling process terminated due to the non-zero interruption probability. The stimulus-to-context associative matrix $\mathbf{M}^{\text{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments were conducted for each combination of discount factor and interruption probability. The sampling distributions over rows (Fig. 3.3b,f) reflect averages across 1000 trials per experiment if starting from the top-center state. The implied contiguity curves (Fig. 3.3d,h) were computed similarly using the same starting state.

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed

based on their respective state value estimate, which was obtained by summing samples within each of 5000 trials and averaging across trials. 100 games were simulated and each trial consists of a variable number of correlated samples (at most nine, or $H-1$). The interruption probability is fixed at 0.05. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 3.3c,g). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 3.3c,g show the average *pmr* across 100 experiments.

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

We set $\rho = \beta = 0.5$ to simulate the effect of an intermediate context updating regime in TCM-SR that better explains human behavioral data on free recall tasks. Simulations were repeated using two different discount factors $\gamma = 0$ (Fig. 3.4a-c) and $\gamma = 0.5$ (Fig. 3.4d-f) during encoding. For each γ, simulations were repeated using three different probabilities of interruption $p = 0.05, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s for each underlying true γ at retrieval (Fig. 3.4b,e). Thus as long as the ball had not reached the bottom of the Plinko board, at each time step, there was a p probability that the trial will terminate, regardless of the ball's location. Consequently, each trial started from the top-center state (marked with a orange circle in Fig. 3.4a,d) and ended when the ball hit the bottom of the board or the sampling process terminated due to the non-zero interruption probability. The stimulus-to-context associative matrix $\mathbf{M}^{\mathrm{SC}}$ was equal to the identity matrix $\mathbf{I}_{|\mathcal{S}|}$.

A total of 100 experiments were conducted for each combination of discount factor and interruption probability. The sampling distributions over rows (Fig. 3.4b,e) reflect averages across 100 trials per experiment if starting from the top-center state.

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was obtained by summing samples within each of 5000 trials and averaging across trials. 100 games were simulated and each

trial consists of a variable number of correlated samples. The interruption probability is fixed at 0.05. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 3.4c,f). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 3.4c,f show the average *pmr* across 100 experiments.

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

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

The intermediate and converged SR matrices of the top-center state (4 panels to the left in Fig. 3.5a,b) were learned via TD(λ), where $\lambda = 0.7, \gamma = 0.9$. A ball was dropped four times from the top-center position of a board with predetermined reward locations and reached the bottom following a sequence of transitions, resulting in 4 complete trajectories. An intermediate SR was computed after observation of each complete trajectory. The unmodulated and modulated learning rates were initialized to 0.01 and 0.5 respectively. i.e., $\alpha_0 = 0.01, \alpha_{mod,0} = 0.5$. Both the unmodulated agent (Fig. 3.5a) and the modulated agent (Fig. 3.5b) were trained using the same exponential decay schedule such that the learning rates upon observing trajectory t was defined as

$$\alpha_t = \alpha_0 * e^{-kt}$$

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

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

We used 100 random experiments (games) and drew 1000 samples from the TD-learned SR after one observation (trajectory) in each experiment to compute the average fraction of samples that contained a reward (Fig. 3.5d). The same set of samples (i.e., after observing a single trajectory) were used to compute the bias and variance in the

value estimate of the top-center state, with a random number of binary rewards between 20 (inclusive) and 40 (exclusive) placed on the board (Fig. 3.5e,f).

Given a game, the agent needed to decide where to drop the ball along the top row to maximize expected total return. For clarity, there were two options - either the top-center state or the location directly to its right. A deterministic policy was assumed based on their respective state value estimate, which was computed as the average across 1000 i.i.d. samples and 50 trials. Simulations were repeated for games with 1, 5, 10, 20 binary rewards accessible from either dropping location (Fig. 3.5c). The percentage of maximum rewards obtained follows the same computation as in Simulation 1. Fig. 3.5c shows the average *pmr* across 100 experiments.

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

We chose the generalized rollout regime (i.e., $\rho = 0, \beta = 1$) to illustrate the effect of retrieving a learned context associated with a stimulus as opposed to a task-independent feature representation. The stimulus-to-context associative matrix $\mathbf{M}^{\mathrm{SC}}$ was equal to the SR matrix $\mathbf{M}$. Simulations used $\gamma = 0.5$ during encoding and three different interruption probabilities $p = 0.2, 0.5, 1$, resulting in three different effective discount factors $\tilde{\gamma}$'s at retrieval (Fig. 3.6c,d). Each simulation consisted of 500 experiments and 1000 trials (rollouts) per experiment from the top-center state.

The true state value of the top-center state was computed by assuming full reversibility (i.e., symmetry of conditional transition probabilities), while the estimates are computed similar to Simulation 2 (i.e., as generalized rollouts; Fig. 3.6c,d).

Acknowledgement

This chapter, in full, has been accepted for publication of the material as it may appear in Psychological Review. Zhou, Corey Y.; Talmi, Deborah; Daw, Nathaniel D.; Mattar, Marcelo G., American Psychological Association, 2024. The dissertation author was the primary author of this paper.

Chapter 4

Behavioral Evidence

Temporally extended decision-making through episodic sampling[1]

While the involvement of episodic memory in sequential decision-making has been posited on theoretical grounds, previous empirical work has so far focused only on one-step tasks (Bornstein et al., 2017; Nicholas et al., 2022; Rouhani et al., 2018). They offer a restricted perspective on episodic sampling compared to the largely untapped sequential realm (which we argue is where episodic memory should be most relevant and effective). Yet, if episodic memory also informs decision-making in temporally extended settings as TCM-SR suggests, the retrieval process should leave footprints on our choices, as Chapter 3 predicted.

Formally, we hypothesize that all else equal,

- memories with higher recall have a larger weight on memory-based decisions (1A),
- but such effect may be modulated using temporal contiguity effect (1B),
- and that the choice between options composed of temporally extended events is best predicted by what is recalled (2).

[1]This chapter is based on the following work:

Zhou, C. Y., Talmi, D., Daw, N. D., & Mattar, M. G. (2024). Temporally extended decision-making through episodic sampling. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.

Zhou, C. Y., Talmi, D., Daw, N. D., & Mattar, M. G. (2022). Computing values through episodic sampling. *The 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making*.

These hypotheses bear two important deviations from conventional RL accounts: first, rather than decisions reflecting (incremental averages over) the full trajectory, only samples are used. This is motivated by the observation that even when data is plentiful, people's decisions can still depend on only a handful of individual experiences (Plonsky et al., 2015). In contrast, conventional RL algorithms consider the full trajectory (episode) in evaluating actions and/or state values. Second, the sampling process is psychologically plausible by taking advantage of a shared mechanism with episodic retrieval, unlike previous models that used a stylized memory store (Gershman & Daw, 2017; Lengyel & Dayan, 2007).

To formalize the computational details of this decision-by-sampling account, we leverage TCM (Howard and Kahana, 2002a) – first in an evaluation task (Experiment 1), and then in a decision task (Experiment 2). TCM captures all the memory biases in our hypothesis via learning of appropriate associative matrices and retrieval using an abstract, evolving representation of recently experience items. The learned associations closely corresponds to a sampling distribution, allowing past events to be drawn in a manner consistent with episodic retrieval to compute decision variables (Mattar et al., 2019).

Our current goal is to empirically test the hypotheses above using two novel experimental paradigms. In Experiment 1, we elicit value estimates (an explicit decision variable and a precursor to choices) from subjects and show they manifest biases outlined in hypotheses 1A and 1B. In Experiment 2, subjects engage in sequential decision-making, whose behavior, as we show, is best predicted by a stochastic episodic sampling account compared to a model-free strategy or a stylized memory (veridical storage). Together, these two studies provide novel evidence for a decision-by-sampling mechanism in humans that is subserved by episodic memory.

4.1 Experiment 1

We test whether memory-based evaluation weighs items differently based on their absolute and relative temporal positions. To probe adaptive evaluation, we design a task where the evaluation goal is unknown during encoding.

4.1.1 Methods

Design and Materials

The key manipulation involves ad-hoc categories, which are not well established in memory during encoding (Barsalou, 1983) and thus affect memory performance less. Specifically, during each test trial, subjects studied a list of items and were then told to estimate the total value of a "partial" list, a subset of the encoded items that belonged to a category revealed after the full list had been presented (Fig. 4.1a) For each trial and subject, the specific category was randomly chosen with replacement from a set of possibilities. The possible categories were not exclusively color-coded and include things like "packaged items," "leafy items," or "spherical items". With a total of 13 categories to sample from, 4 of which are colors, this prevents the subject from using any specific encoding strategy (e.g., color coding) to solve the task.

All study lists were constructed such that partial list items were spread across serial positions: the first partial list item appeared either as the first or second item, while the last partial list item appeared as one of the last two items.

66 colored images of common grocery store items were collected from various online sources. For each participant, 45 unique items were randomly selected and grouped into 5 study lists, corresponding to 5 possibly repeating random categories. The first list contained 5 items and was used as the practice trial. The rest contained 10 items each. Each 10-item full list contained a 5-item partial list (each corresponding to a random ad-hoc category), and the 5-item full list corresponded to a 3-item partial list. No instruction ever hinted at the size of the partial lists. All items were shown with an integer value between 1 and 12, which was pre-determined according to the national average price so as to aid value encoding and avoid enhanced encoding due to surprise.

Two pairs of value estimation and free recall tasks followed list studying. The first pair concerned the partial list and the second was about the full list. The value estimation task prompted the subject to estimate the total price of the corresponding list. Subsequently, the recall task asked for the names of the list items. The specific questions were

1. Based on the images, you will first check out the [**category**]. Roughly, what is the **total price** of only the [**category**]?

2. Please write down the **names** of the items you just checked out (i.e. only the [**category**] based on the images).

3. Now, you check out the rest of the items in your cart. Roughly, what is the **total price of all items** in your shopping cart? (including the [category] but also everything else)

4. Please write down the **names** of as many items as possible from your cart (including the [category] but also everything else).

A cued recall test was additionally administered during the last trial, where subjects needed to recall an object from the list based on a written description (not shown in Fig. 4.1a). Unbeknownst to the participants, the answer was always the third (middle) partial list item. The description was not specific enough to infer the answer from common sense, but could be uniquely determined given the presented items. For instance, the prompt may read

> Based on the images, what was the item from your current cart that best fits the following description?
>
> **white item in a cup**

The answer is "yogurt," which is the only item in the entire set of items that fits this description.

Procedure

The experiment consisted of one (1) practice trial and four (4) test trials. Feedback was only provided at the end of the practice trial. Subjects had to pass a short quiz with 100% to ensure good understanding of the experiment procedure before moving onto the test trials. During each test trial, an interval of length uniformly drawn between 800ms and 1200ms was inserted between two item presentations.

For the first three test trials, after a 500ms interval following list presentation, subjects completed the two pairs of value estimation and free recall tasks in the order described in the previous section. Each value estimation task had a time limit of 60 seconds. The two free recall tasks had a time limit of 30 seconds (partial list) and 40 seconds (full list) respectively.

On the last test trial, subjects completed a cued recall test before proceeding with the value estimation and free recall tasks. This test was not timed.

Participants

200 subjects were recruited through Prolific, 66 of which were excluded from analysis due to excessive low effort responses (e.g., zero recall for at least 2/4 of the test trials) or responses that indicate gross overestimation or underestimation of values. Specifically, overestimation is defined as an estimate larger than the maximum possible item value times the number of recalled items, and underestimation is defined as an estimate smaller than the minimum possible item value times the number of recalled items. Neither criterion depends on how accurately individual values were remembered. A total of 134 subjects were included in the subsequent analysis.

4.1.2 Results

Serial Order Effects

Free recall of episodic memory exhibits primacy and recency effects. Here, consistent with the classic findings in the free recall literature, subjects recalled more items more often from either end of the partial list (Fig. 4.1c; position 1 vs. position 2: $t(133) = 3.28, p < 0.001$; 1 vs. 3: $t(133) = 2.97, p = 0.003$; 1 vs. 4: $t(133) = 3.73, p < 0.001$; 5 vs. 2: $t(133) = 3.74, p < 0.001$; 5 vs. 3: $t(133) = 3.46, p < 0.001$; 5 vs. 4: $t(133) = 4.21, p < 0.001$). Thus, these results suggest primacy and recency are also observed in partial list free recall when the retrieval criteria (ad-hoc categories) are unknown during encoding.

Value Estimation

To quantify the effect of serial list position and episodic memory retrieval on value estimation, we fit a mixed-effect regression model of the form on data from the first three

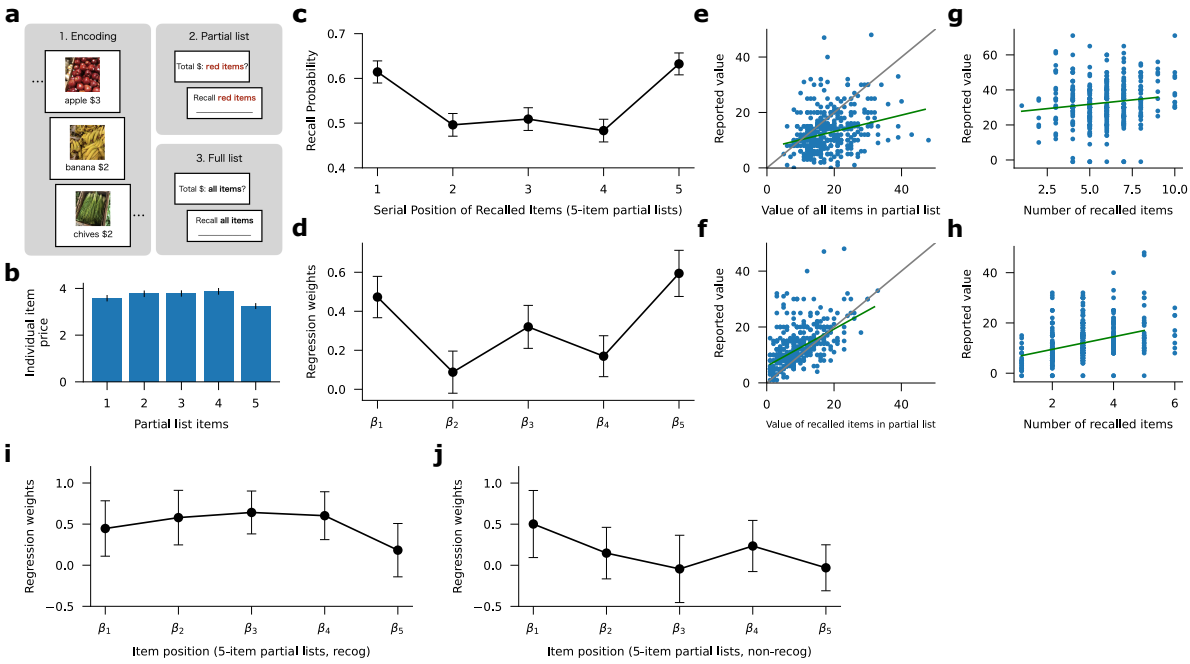

Figure 4.1: **(a)** Example presentation of an item. **(b)** Average price of partial list items as a function of list position. **(c)** Average probability of recall as a function of an item's position in the 5-item partial list. **(d)** Fitted fixed effect regression weights on subjects' reported partial list value. β_i's corresponds to position i in the 5-item partial list. **(e)** Reported partial list value against the true partial list value. **(f)** Reported partial list value against the true value of recalled partial list item(s). Each dot represents one list from one subject. i.e. a subject may contribute to multiple data points. **(g)** Reported full list value against number of recalled items in the full list. **(h)** Reported *partial* list value against number of recalled items in the *partial* list. **(i)** Fitted fixed effect regression weights on reported partial list value by subjects who correctly answered the probing question. β_i's corresponds to position i in the 5-item partial list. **(j)** As in (i) for subjects who answered the probing question incorrectly.

blocks

$$\mathbf{V}_{reported} = \mathbf{V}\boldsymbol{\beta} + \mathbf{Zu} + \boldsymbol{\epsilon}. \tag{4.1}$$

The fixed effects $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m, \beta_n)'$ represent the average effect of different partial-list positions on the overall value estimate, and the random effect $\mathbf{u}$ captures effects specific to individual subjects. $\mathbf{V}_{reported}$ is a vector of reported partial list values with length L equal to the total number of trials. Each row of the design matrix $\mathbf{V}$ contains the individual item prices in a partial list, plus the total number of partial list items recalled. $\mathbf{Z}$ contains subject ids. The inclusion of the number of recalls helps to explain an additional 12.2% fixed effect variance and 5.0% overall model variance. A model comparison also shows that including the number of recalled partial list items significantly improves the model ($\boldsymbol{\chi}^{\mathbf{2}}_{\mathbf{1}} = 57.4, p < 0.001$).

The first and the last items have a significantly higher weight on the value estimate compared to the three middle items (β_{first} vs. β_{middle}: $F(388) = 5.24, p = 0.023$; β_{last} vs. β_4: $F(388) = 10.24, p = 0.001$). Comparison of each pair of partial list positions reveals a similar trend, where the first and last items receive a larger weight in general (Fig. 4.1d; β_1 vs. β_2: $F(388) = 2.51, p = 0.012$; β_1 vs. β_4: $F(388) = 1.92, p = 0.055$; β_5 vs. β_2: $F(388) = 3.18, p = 0.002$; β_5 vs. β_4: $F(388) = 2.64, p = 0.008$), consistent with subjects' serial recall patterns.

The parallel between the recall probability and regression weights in this task agrees with our hypothesis. To further test whether episodic memory recall predicts evaluation, we regress subjects' reported partial list value on two quantities separately using Huber loss: (i) the true partial list value, and (ii) the true total value of only the recalled partial list items. The reported partial list values turn out to strongly correlate with the true value of recalled items (Fig. 4.1f; $\beta_r = 0.640, R^2 = 0.383$), more so than the true partial list value v_t (Fig. 4.1e; $\beta_t = 0.291, R^2 = 0.103$). Note the latter corresponds to an alternative account that people keep track of an aggregated statistic (i.e., independent of the specific items they can remember), akin to a model-free reinforcement learning agent. Regressing the reported values on (i) and (ii) together also show that (ii) was more predictive than (i).

Temporal Contiguity

The cued recall test was inserted in the last trial in an attempt to update the temporal context of the subject before recall and change subsequent recall dynamics. Because the correct answer is always the item in position 3 of the partial list, the temporal contiguity effect would predict an enhanced probability of recalling the items at position 2 and/or 4, while attenuating both the primacy effect and the recency effect (i.e., worse recall accuracy of item 1 and 5 compared to without the cued recall test).

Subjects who correctly recognized the cued item show attenuated primacy and recency effects, as well as a qualitatively higher recall accuracy of the neighboring items (Fig. 4.1i). On average, the middle items have an increased influence on the value estimate compared to the previous trials, as no pairs of the regression weights are significantly different. Those who failed to recognize the cued item do not show any significant primacy or recency effect either (Fig. 4.1i, compared with Fig. 4.1c), likely because the (failed) memory retrieval also changed their internal context to some extent. Thus the results suggest a temporal contiguity effect even in subjects who did not respond correctly to the cue.

4.1.3 Discussion

Results from our study suggests people rely on episodic memory to adaptively compute values in our task, supporting the hypothesis that memory plays an important role in general decision-making. Participants memorized lists of everyday grocery items. They were then asked about the total price of a subset of items – a task meant to mimic the computation of a decision variable – and finally, recalled these items. Regression analysis shows that the first and last items in the subset had greater weights on value estimates, suggesting primacy and recency effects were at play. Furthermore, when an additional question was introduced to shift subjects' temporal context to the middle of the list, previous signs of primacy and recency effects disappeared, suggesting temporal contiguity at play. These results are consistent with hypotheses 1A and 1B, providing novel evidence for a psychologically plausible decision-making mechanism using episodic memory samples.

Two features distinguish our paradigm in comparison to previous recall experiment. First, we introduced a value estimate question in order to investigate the role of episodic memory retrieval dynamics in value judgment. Second, we ensured that encoding strategies could not influence the answer so as to minimize its confounding effect on retrieval dynamics and value computation. By design, our paradigm forced subjects to compute decision variables only after encoding was completed.

An alternative explanation for the results is that subjects employed an average-based strategy, tracking a running average of item values during encoding, rather than encoding individual items and their values. At decision time, they could multiply the average by the number of recalled items. However, this did not appear to be the case, at least on the group level. The number of recalled partial list items was not predictive of reported value estimates (Fig. 4.1h; $\beta_r = 2.50, R^2 = 0.128$), and even less so for the full lists whose length can be easily counted (Fig. 4.1g; $\beta_r = 0.994, R^2 = 0.019$).

Value effects may also complicate episodic memory retrieval dynamics (Stefanidi et al., 2018). For example, subjects might be more likely to recall items of extreme values (Lieder et al., 2018). Nonetheless, the average item value were similar across list positions (Fig. 4.1b), so these additional dynamics induced by value differences were unlikely to be the driving force behind our results.

Our findings thus far provides the initial evidence of an episodic memory footprint on *evaluation*, a precursor to decisions. The next step to follow is to extend the paradigm into a sequential decision making task that recruits episodic encoding and retrieval.

4.2 Experiment 2

Building onto the link between episodic retrieval and evaluation in Experiment 1, we next seek to establish a more direct connection between episodic retrieval and decisions: whether people's choice can be well predicted by what they recall at decision time. Again, to tap into adaptive behavior, decision-relevant information is unknown during encoding. Additionally, we take advantage of the temporal contiguity effect to manipulate subjects' recall rate. By inducing quantitatively different recall patterns *within individual subjects*, we disambiguate several alternative explanations and demonstrate how episodic retrieval

accounts for people's trial-wise choices.

4.2.1 Methods

Design and Materials

The temporal contiguity effect suggests that items encoded close together in time are more likely to be retrieved together, and more so in the same order as presented (Howard & Kahana, 1999). This implies that discontiguously encoded items - that is, items presented at relatively distant temporal steps - would have a worse average recall rate. At the same time, the spatial distance between items should be controlled to minimize the spatial contiguity effect (J. Miller et al., 2013). We hence designed a gridworld which subjects explored in multiple runs. They then decided between two subsets of items that were encoded with different levels of temporal contiguity.

123 black-and-white cartoon battle items were collected from an online resource[2]. For each participant, 108 items were randomly selected to form 9 gridworlds with 12 unique items each, all of which needed to be fully explored (Fig. 4.2a). The items were hidden behind the grey squares. To fully explore a gridworld, they took four zigzag routes composed of gray squares by starting from the top-center location (marked by $\star$) every time and pressing either the left or the right arrow key to move to an (diagonally) adjacent grid (e.g., $\star \to \square \to A \to B \to C \to D$). Each zigzag path therefore contained 4 items. Black squares were inaccessible and the steps were irreversible. The top row did not contain any item and was solely for navigation purposes.

The image of each item was only shown once when the participant first navigated to its location by pressing on a keyboard, but not any future (repeated) visits. All future visits to the location showed a black cross, and the subject could visit a grid twice only if doing so was the only way to access another novel item. For example, to access item F after first taking the path A(item) $\to$ B(item) $\to$ C(item) $\to$ D(item), the participant may take the path A(X) $\to$ E(item) $\to$ C(X) $\to$ F(item). None of the locations on the right half may be accessed thereafter.

Each item had two attributes - one attack value and one defense value, both of

[2]https://game-icons.net/

which were shown along the image upon the first visit. The attack and defense values were integers between 1 and 4 (inclusive). They were assigned manually in accordance with the physical properties of the item. For instance, "death star"[3] has attack 4 and defense 4; "vending machine" has attack 1 and defense 3 (it's hard to throw and has limited damage unless the impact is direct, but pretty decent as a barricade); "death note"[4] has attack 4 and defense 1. The values were determined so that (1) they are intuitive for subjects to encode based on the physical attributes, (2) attack/defense values had roughly equal frequency (≈ 30 items per value), and (3) attack values and defense values did not predict each other.

A decision task appeared after exploration, where the participant chose between two "paths" to maximize either the total attack or total defense points. They were not told which attribute was relevant until this point. Both paths in the decision task were taken from the four zigzag paths the subject actually took and were paired so they did not overlap with one another (i.e., one on the left half, one on the right half). Crucially, exactly one of them was encoded *contiguously*, meaning the subject saw all the items on the path without encountering black crosses and the temporal distance between spatially adjacent items is one (e.g., $A \to B \to C \to D$ in the example above), while the other was encoded *discontiguously*, such that two of the locations were crossed out at exploration and the temporal distance between spatially adjacent items may be three or more (e.g., $A \to E \to C \to F$). We refer to the two types of paths as "full" and "partial" paths respectively. In other words, on the first route through a particular half of the grid, participants saw four images in total (full path). On the second route on the same half, participants only saw two new images with two crossed out squared interleaved in-between (partial path). Each trial thus consisted of two full paths (one on the left half, one on the right half) and two partial paths.

To control for possible primacy and recency effects, for each participant, the paths

[3]The Death Star was the Empire's ultimate weapon in *Star Wars*. It is a giant space station that can destroy an entire planet with superlaser. See https://www.starwars.com/databank/death-star for details.

[4]Roughly speaking, writing down a person's name in the notebook kills them almost immediately, but the notebook itself can't protect the owner from being killed. See https://deathnote.fandom.com/wiki/Rules_of_the_Death_Note for details.

were selected such that the first and/or the last path the participant traversed was only queried at most 2 out of 9 test trials. This means that the majority of the decision tasks (7/9) asked about the middle two runs through the gridworld. The options were not exclusively middle paths so as to avoid attentional biases, since the participant may discover the pattern after a few trials and selectively encode the items.

Two recall tasks then followed the decision, asking the participant to write down names of the items in each path option. The specific questions were

1. Which of the following path would you pick if you want maximum [**attack/defense**] power?

2. What **items** did you collect in the locations highlighted below[5]? Feel free to describe them if you aren't sure about the exact names.

3. What **items** did you collect in the locations highlighted below[6]? Feel free to describe them if you aren't sure about the exact names.

Procedure

The experiment consisted of three (3) practice trials and nine (9) test trials. Participants first familiarized themselves with the keyboard control to navigate through the maze efficiently during practice, and completed a trial similar to the test trials except with a smaller gridworld (6 hidden items). They had to pass a short quiz with 100% correctness to ensure good understanding of the experiment procedure before moving onto the test trials.

During each test trial, participants were given 75 seconds to uncover all hidden items. Each item (along with the two attribute values) was shown no more than once for 4 seconds when the navigation reached its location for the first time. The participant could not move their cursor during the full duration of the item presentation. On subsequent visits to the location, the square simply shows a cross, and the participant can immediately navigate to the next grid without any pause. If a gridworld was not fully explored, the trial was skipped and the participant was warned that the trial had timed out.

[5]Note: orange path

[6]Note: blue path

All decision and recall tasks had a time limit of 60 seconds. The specific attribute queried was randomly selected for each trial, and was not revealed until the decision task. Path highlight colors were random and did not indicate full/partial types. The trial was excluded if no choice was made. All 733 trials had a valid choice response.

Feedback was provided at the end of each trial after the last recall task. Participants could view the full gridworld with items shown, as well as the values of the path options with individual item breakdown.

Participants

100 subjects were recruited through UCSD SONA, 16 of which were excluded from analysis due to excessive low effort responses (zero recall on at least 5/9 of the test trials) or responses that indicate note-taking (perfect recall with exact order as presented in all trials). A total of 84 subjects were included in the subsequent analyses. There were 733 trials with at least one recall.

4.2.2 Results

Temporal Contiguity

The key manipulation of this study is the construction of partial paths: items along the partial paths are experienced in a different order than full paths – specifically, the second item and the fourth item were observed one *temporal* step apart, even though they are two *spatial* steps apart. Since the temporal contiguity effect implies that temporally adjacent items are more likely to be retrieved in succession, we hypothesize that two-step transitions (i.e., relative lag = ± 2, where lag is defined as the number of key presses) in free recall are more probable in the partial path condition.

Indeed, free recall of partial path items shows a clear disruption of the temporal contiguity effect when relative distance between items is defined spatially as opposed to temporally (Fig. 4.2b). While subjects still tend to recall in the forward direction (i.e., recalling items in the order they were observed), the recalled item that immediately follows a previous recall is as likely to be from one (spatial) step ahead (+1) as two (spatial) steps ahead (+2; $t(428) = -0.095$, $p = 0.92$). In contrast, serial recalls of full path items display classic temporal contiguity, such that subsequent recalls are much more likely to

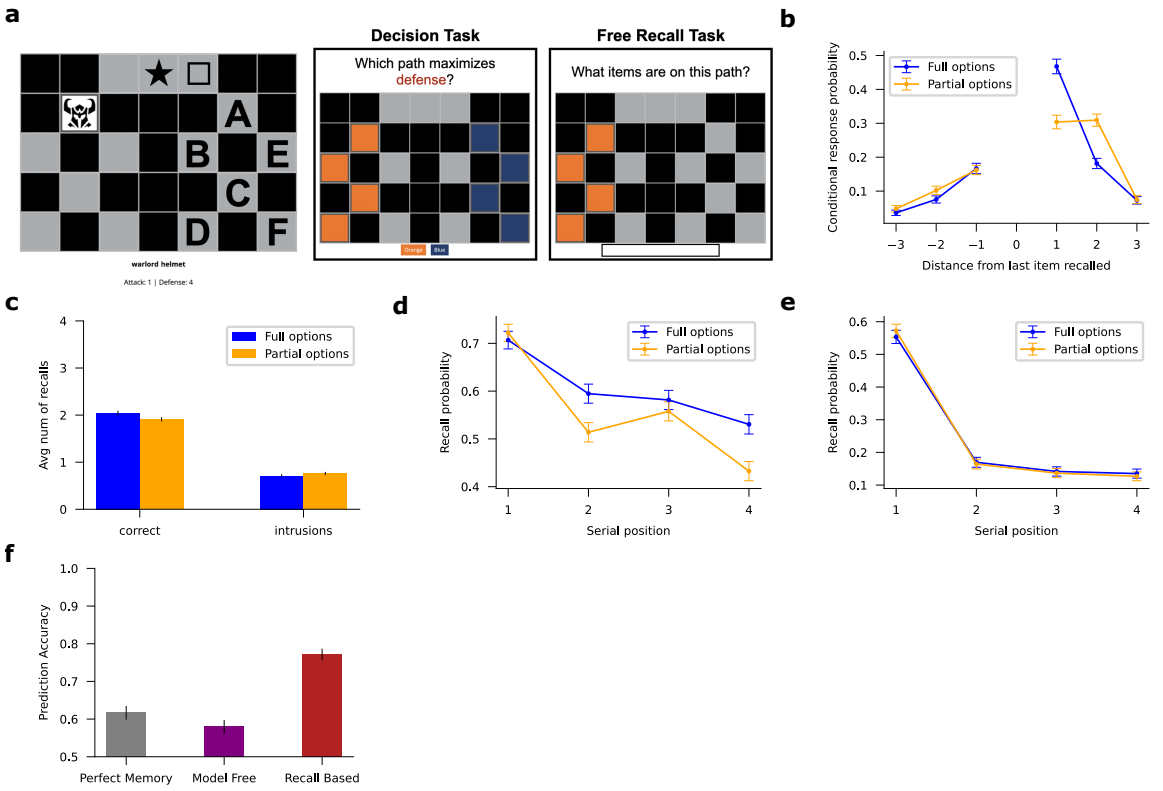

Figure 4.2: **(a)** Example gridworld and tasks during a trial. **(b)** Probability of successively recalled items as a function of their relative spatial distance (number of key presses). For a pair of items on a full path, their spatial distance is equal to their temporal distance. For a pair of items on a partial path, their spatial distance may be longer than their temporal distance. e.g., F is +2 spatial distance away from E, but +1 temporal distance away from E, because C is crossed out when the subject takes the route $A \to E \to C \to F$ after taking $A \to B \to C \to D$. By design, the subject can immediately pass through C without having to wait 4 seconds (presentation time of an item). Results are averaged across trials, and error bars indicate standard errors. **(c)** Trial-wise average number of correctly recalled items and intrusions. **(d)** Trial-wise overall recall probability of an item as a function of its serial position along the queried path. **(e)** Trial-wise probability of the first recalled item as a function of its serial position along the queried path. **(f)** Trial-wise prediction accuracy of the three candidate models (see text for details) with respect to subjects' actual choices.

come from one step ahead.

Serial recall

Temporal contiguity effect indicates that episodic retrieval of items is most likely to ensue in close temporal proximity with respect to the order that items are experienced and encoded. It is less likely for people to subsequently recall an item seen a few temporal steps away than something that directly follows a just recalled item during exploration. Since the four items in each partial path were experienced in two separate runs, we expect more difficulty in retrieving all four items and thus lower recall accuracy.

On average, subjects recalled 1.94 items from a queried path, regardless of the path type (full/partial). Across the 733 trials with at least one recall, an average of 2 items were recalled from each full path (s.e.m. = 0.05), while an average of 1.88 items were recalled from each partial path (s.e.m. = 0.05), which was significantly lower ($t(732) = 2.87$; $p = 0.004$) and is consistent with our hypothesis (Fig. 4.2c). There was no difference in the number of recall intrusions between full and partial paths ($t(732) = -1.32$; $p = 0.19$, suggesting that the reduced recall accuracy is unlikely due to incorrectly recalling items from the full path overlapping with the queried partial path.

A closer look reveals that the poorer recall of the partial path items was primarily driven by worse memory of the second and fourth items (Fig. 4.2d). This may seem counterintuitive, since those were exactly the items that were presented and encoded when participants traversed the partial path (e.g., E and F), while the first and third items were crossed out (e.g., A and C). However, this finding is exactly what we should expect given the disrupted (spatial) contiguity: first, similar to full paths, subjects were most likely to recall the first item in the partial path (Fig. 4.2e). Critically, because item 1 was only contiguously encoded with item 2 on the *full* path, recalling this item would shift their temporal context away from item 2 of the partial path and toward item 2 of the full path. Indeed, if item 1 on a given partial path was recalled earlier, the recall rate of item 2 of the same partial path (mean = 0.346) was lower than that of item 2 of the full path (mean = 0.418), and the effect was significant on the subject level ($t(81) = -2.153, p = 0.034$). In contrast, there was no difference between the recall rate of item 3 if item 1 was recalled

earlier (mean = 0.416 (partial) / 0.415 (full), $t(81) = 0.023, p = 0.98$), since they were encoded contiguously as part of the full path that overlapped with the queried partial path.

Recalling item 3 should also shift the temporal context towards item 4 of the full path more than item 4 of the partial path. We found that if item 3 was recalled earlier, it was much harder to retrieve item 4 of the partial path (mean = 0.297) than item 4 of the full path (mean = 0.382; $t(81) = -2.132, p = 0.036$), despite having the same spatial distance. On the other hand, if the subject managed to recall item 2 of the partial path, the temporal context should evolve towards item 4 instead of item 3, which was not encoded contiguously after item 2. Our data supports this hypothesis as well: while the conditional probability of recalling item 3 on the full path is 0.392, the same probability for the partial path was only 0.295 ($t(81) = 2.20, p = 0.029$). Surprisingly, conditioned on an earlier recall of item 2, the recall rate of item 4 was also lower for the partial path (mean = 0.389 (full) / 0.288 (partial), $t(81) = 2.81, p = 0.0062$). This may be due to the "lost in the middle" serial position effect, since the queried paths were rarely the first or the last path taken by the subject. In particular, TCM predicts that intermediate temporal context (after evolving for more than a few steps) may not be sufficiently informative about the stimulus identity, leading to unsuccessful retrieval of temporally adjacent items encountered in the middle of a trial.

Choice Prediction

We first inspected subjects' performance on the test trials. Despite the perceived task difficulty, they made the optimal choice on 69.09% of the trials. On the subject level, choice accuracy was significantly better than chance ($t(83) = 10.43$, $p < 0.001$), with no obvious prior bias towards either option ($t(83) = 1.35$, $p = 0.180$).

To test the hypothesis that episodic recalls are better predictors of people's adaptive choice, we compare different computational accounts of the decision process to see which one predicts actual trial-wise choices more accurately. Specifically, we consider three candidate models:

1. **Model free (MF)** - the agent accumulates values over the *observed* items for both

attack and defense along each path, but retains no memory about the individual items. This gives perfect value estimation for the full paths. Since individual item values are similarly distributed regardless of the grid location ($F(3,729) = 1.46$; $p = 0.22$), the accumulated value of partial paths over the two observed items is in expectation half of its actual value (four items in total). Thus the agent estimates the partial path value as twice the accumulated number, as it has no information about the individual item values. It chooses greedily based on the two value estimates and has no free parameters.

2. **Perfect memory (PM)** - the agent is assumed to remember perfectly the individual items and their attributes. It always chooses optimally by adding up the individual values and greedily picking the option with the higher value. It has no free parameters. This model is most similar to the stylized view of episodic memory in previous RL models and modeling of one-step decision behavior, where memory is simply treated as veridically storing a handful of experiences. During retrieval, item identities and associated values are recalled exactly (e.g., exact match in a differentiable neural dictionary, or DND) in no particular order or pattern as observed in free recall tasks.

3. **Recall-based (RB)** - the agent draws episodic samples of individual items and adds up the attack/defense values of sampled items as needed. We assume it encodes and retrieves values perfectly. Additionally, it uses a single value as the expected value for any item it fails to recall. Since the total number of items within each option is obvious (unlike Experiment 1), this quantity acts as a reasonable placeholder to make up for anything forgotten. For instance, suppose a subject recalled the four items with probability $[0.7, 0.6, 0.6, 0.5]$. The value of the first item is either sampled with 70% chance (successful recall) or filled with the placeholder value with 30% chance (forget); the value of the second item is either sampled with 60% chance (successful recall) or filled with the placeholder value with 40% chance (forget); and so on. If, say, it recalls the first two with attack value 1 and 4 respectively, with $\mathbb{E}$[value of unrecalled item] = 2, the path value estimate is 1 + 4 + 2 * 2 = 9. The agent has one free parameter $\mathbb{E}$[value of unrecalled item] and also uses a greedy

decision policy as the other two models.

To predict choices, we fit an RB model to each subject – that is, each RB model shares the same subject-level recall probabilities as the corresponding subject (e.g., if the subject recalls the first item of a full path 70% of the time across trials, the model will successfully sample the value of the first full item 70% of the time across all trials completed by the subject; see Fig. 4.2d for group level probabilities), and a subject-specific value filler (i.e., $\mathbb{E}$[value of unrecalled item]) is fitted to maximize choice prediction accuracy. Value fillers are fitted using a grid search with a grid size of 0.1 in the range [0,4]. The fitting procedure included 50 simulations per subject, such that the subject-level value filler $\mathbb{E}$[value of unrecalled item] maximizes the mean prediction accuracy across simulations.

Among the three candidate models, the recall-based model predicts subjects' actual choices most accurately, with a trial-wise average of 75.85% and log likelihood (LL) of -1888.13 (Fig. 4.2f), much better than both the model-free model (57.95%, LL = -3172.796) and the perfect memory account (61.66%, LL = -2650.21). All choice prediction accuracies are evaluated using leave-one-out cross validation. We have tried ϵ-greedy and softmax policies as well, but the additional parameters (ϵ, softmax temperature) of the best-fitting models indicated near-random choice, so we leave them out of the current analysis.

4.2.3 Discussion

Experiment 2 offers additional insight into how episodic retrieval guides choice when value computation is delayed till decision time. In a novel task, participants explored gridworlds to encode game items. They then chose between two subsets of items encoded with different levels of temporal contiguity based on an attribute selected randomly ad hoc. Finally, they performed free recall of the items within each subset. A comparison of people's responses with different decision models reveals that the recall-based account best predicts actual choice, followed by a perfect memory model and a model-free account. This suggests that subjects' choices are most accurately captured by what they recall at decision time (plus a placeholder for items that they *know* they cannot recall). In contrast, the model-free strategy does not retain any item-specific information to be recalled, and explains human decisions the poorest, even though it solves the task.

It is worth noting that the disagreement between the perfect memory account (which always makes optimal decision) and the recall-based model highlights the way memory biases decision – people rely on episodic retrieval to compute action values, so when fewer episodic samples are drawn successfully (e.g., partial paths), their decisions are *predictably* suboptimal. This finding supports the decision-by-sampling hypothesis, suggesting that people's choices are best predicted by their average recall rates at decision time.

The recall-based model operationalizes the hypothesis that "what is remembered is factored into choice." Unlike TCM, which is a mechanistic model of *how* episodic retrieval takes place, RB only concerns about *what* episodic memory is retrieved, and makes a decision based on the recalled values. This decision process is analogous to TCM-SR with no reward discounting (e.g., generalized rollout), which draws episodic samples from each path and adds up their associated rewards as the path value. The main difference is instead of fitting TCM parameters to approximate subjects' actual recall (by minimizing the distance between the simulated recall probability and Fig. 4.2d; see Q. Zhang et al., 2022 for example), the empirical recall probabilities are directly used to generate memory samples.

The focus of our analysis is primarily with respect to the temporal contiguity within a single run through the gridworld (i.e., encoding a single path). Nevertheless, we have glossed over another source of temporal contiguity – encoding across separate paths. Participants could very well perceive each path through the gridworld as a distinct "event." The resultant event boundaries could cause large drifts in the temporal context during encoding (Pu et al., 2022; Rouhani et al., 2019) such that if two paths were traversed further apart in time, the overall recall is less accurate. Therefore, one limitation of the current study design is that there is no way of measuring the effect of within-event and across-event temporal contiguity on decision making. A possible improvement is to expand the gridworld to allow more paths and vary the temporal distance between the queried paths.

Another limitation of our paradigm is that partial paths were always encoded

after full paths. Since half of the items were shared between the full path and the partial path on either side of the gridworld, subjects might have attended to the full path more, thinking it would be more informative than the partial path, which only showed two items. One way to mitigate this order effect is to randomly block the overlapping items during the first move and unblock it for the second move, effectively reversing the order of full - partial paths. For example, when going down the board in Fig. 4.2a, the first path may be $A(\text{X}) \rightarrow B(\text{item}) \rightarrow C(\text{X}) \rightarrow D(\text{item})$, which makes the second path $A(\text{item}) \rightarrow E(\text{item}) \rightarrow C(\text{item}) \rightarrow F(\text{item})$.

4.3 General Discussion

Through a series of experiments, we find that both value estimates (Experiment 1) and adaptive choices (Experiment 2) show footprints of episodic memory biases and can be predicted from episodic recall patterns. Consistent with our hypothesis, memory-based decisions weigh events differently based on their serial position analogous to the primacy and recency effects in episodic memory (1A). Such effect can be modulated by the temporal contiguity effect, causing intermediate items to have qualitatively higher recall than unmodulated (1B). Moreover, people's choices are better predicted by their average recall pattern than aggregated statistics (model-free strategy) or the simplistic view of memory as information storage (perfect memory). The results suggest that an episodic sampling mechanism underlies adaptive decision-making in humans, such that encoded information may be integrated for decisions ad hoc, subject to episodic encoding and retrieval biases.

Two features distinguish our paradigms in comparison to previous recall experiments. First, we extend classic word list learning tasks to investigate the role of episodic retrieval in evaluation and action. Second, we ensure that encoding strategies could not influence the answer so as to minimize its confounding effect. By design, our paradigm forces subjects to retrieve information and compute decision variables only after encoding has been completed.

The rich connection between experiences and choice has attracted interest from various subfields of cognitive science; here, we highlight an under-explored account of

decision by episodic sampling by developing two novel behavioral tasks that suggest the recruitment of episodic memory in evaluation and choice. These findings provide the initial empirical support for computational accounts that combine sample-based decision-making with episodic retrieval models such as TCM, which lays the ground for future efforts into understanding sequential decision making in naturalistic settings.

Acknowledgement

This chapter, in part, is a reprint of the material as it appears in Proceedings of the Annual Meeting of the Cognitive Science Society. Zhou, Corey Y.; Talmi, Deborah; Daw, Nathaniel D.; Mattar, Marcelo G., eScholarship University of California, 2024. The dissertation author was the primary author of this paper.

Chapter 5

Hierarchical TCM-SR

Hierarchical memory mechanisms implement human meta-learning[1]

Thus far, Chapter 3 and Chapter 4 have formalized and empirically tested the hypothesis of decision-by-sampling by recruiting episodic memory mechanisms. However, their assumption bears one critical deviation from our actual living experience – that episodic memory does not encode everything as an uninterrupted streamline. While TCM resets between tasks, our brain creates events. The capacity to effectively represent, organize, and evoke past experiences is core to us as intelligent and adaptive agents navigating through a complex and uncertain world. Yet little is known about the cognitive process that supports such a central ability.

Despite suggestive links among episodic memory, event representation, and cognitive control, few formal models addresses the underlying mechanisms to unify the three. The structure event memory model (SEM; Bezdek et al., 2022; Franklin et al., 2020) leverages Bayesian latent cause inference to model human event cognition with promising results. However, its characterization of episodic memory using recurrent neural networks is highly stylized, with loose connections to the memory literature. SEM also falls short of capturing the hierarchical organization of acquired knowledge. Another recurrent neural network model, the Episodic Generalization and Optimization Framework (EGO; Giallanza et al., 2024), incorporates more psychologically plausible mechanisms and enables hierarchical

[1]This chapter is based on the working paper:
Zhou, C. Y., & Mattar, M. G. Hierarchical memory mechanisms implement human meta-learning.

control via the interaction of different functional modules. While EGO accounts for human performance in a range of decision tasks, it limits the way past experiences are initially represented. The Option Model (Collins & Frank, 2013; Xia & Collins, 2020) instead combines Bayesian inference with the hierarchical reinforcement learning (HRL; Sutton et al., 1999) framework to discover reusable options from past experience. Yet it makes little theoretical connection to any kind of memory and has only been tested on a limited set of empirical studies.

To gain further insight into the cognitive representations and processes that support adaptive composition and generalization of behavior, it is thus important to account for all three components – memory, event representation, and control – in a wide range of decision scenarios that is informed by studied cognitive constraints while limiting additional assumptions. An additional objective is to improve the interpretability of current models. Replication of human behavior is never the sole end of modeling; rather, the greatest value of computational models lies in the formalization of mechanisms that give rise to the human-like behavior in the first place. By explicating the internal process, the model suggests testable hypotheses that can help crack open the blackbox of the human mind. For this reason, this dissertation is motivated to take a primarily symbolic approach to a unified account of EM-for-DM. Specifically, I propose to integrate latent cause inference and hierarchical RL into the basic mechanistic model in Chapter 3. Chapter 5 shows that this framework rooted in episodic mechanisms explains a diverse set of human behavior that underlie and/or is indicative of the ability to generalize past experiences for adaptive use. Remarkably, it also suggests ways continual learning may happen in a purely online and self-supervised regime.

5.1 Model

Episodic memory dynamically interacts with event representation and cognitive control to facilitate adaptive behavior in an ever-changing environment. In the current work, we propose a solution to this three-way interaction through hierarchical episodic encoding and retrieval. We extend a recent model that grounds rational choice behavior in the dynamics of human episodic memory (C. Y. Zhou et al., in press) (Chapter 3)

to additionally infer the latent causes of its observations and represent past experiences accordingly. We hypothesize that organizing knowledge by the inferred latent cause gives rise to what cognitive psychologists call "event models", which may be the basis of (sub)task representation in decision making. Events, therefore, should not be merely seen as perceptual units but should also be studied as functional entities, and we use it interchangeably with "(sub)tasks" in this paper. This extension results in a novel symbolic framework to explain how event cognition and meta-learning arise in humans, and we show its ability to capture human behavior in various tasks previously studied. In doing so, we formalize a unified account of episodic memory, event (task) representation, and cognitive control with respect to efficient knowledge representation and use in humans.

In the current work, we seek an algorithmic model that addresses shortcomings of prior models. In particular, it has the following properties: (1) it allows temporal abstraction over multiple timescales (hierarchical); (2) abstractions on the same level can be combined (compositional); (3) it makes minimal assumptions about how experiences are initially encoded, so the model can scale to real-world complexity; (4) its design is informed by established theories and known cognitive constraints such that it is biologically plausible; (5) its behavior aligns with human behavior across diverse tasks. We outline how properties 1-3 are achieved by the current computational framework below and detail the design motivations (property 4) in Section 5.4. In Section 5.3, we simulate the model on five different behavioral tasks to demonstrate property 5.

For property 1 (hierarchical), we adopt a similar approach as Franklin et al. (2020) and Xia and Collins (2020) by equipping the base model with latent cause inference (Courville et al., 2004; Gershman et al., 2010) so the agent clusters experiences into events and learns the hidden structure of each event. This process explains the timing of episodic memory updates reflected in human behavior (Gershman et al., 2014), place cell remapping (Sanders et al., 2019), consistent neural activation by shared event structures (Baldassano et al., 2016), and reinstatement of event models during recall (Baldassano et al., 2016; Gershman & Niv, 2012). Like SEM, our event representations resulting from latent cause inference enable predictions about upcoming observations and evaluation of

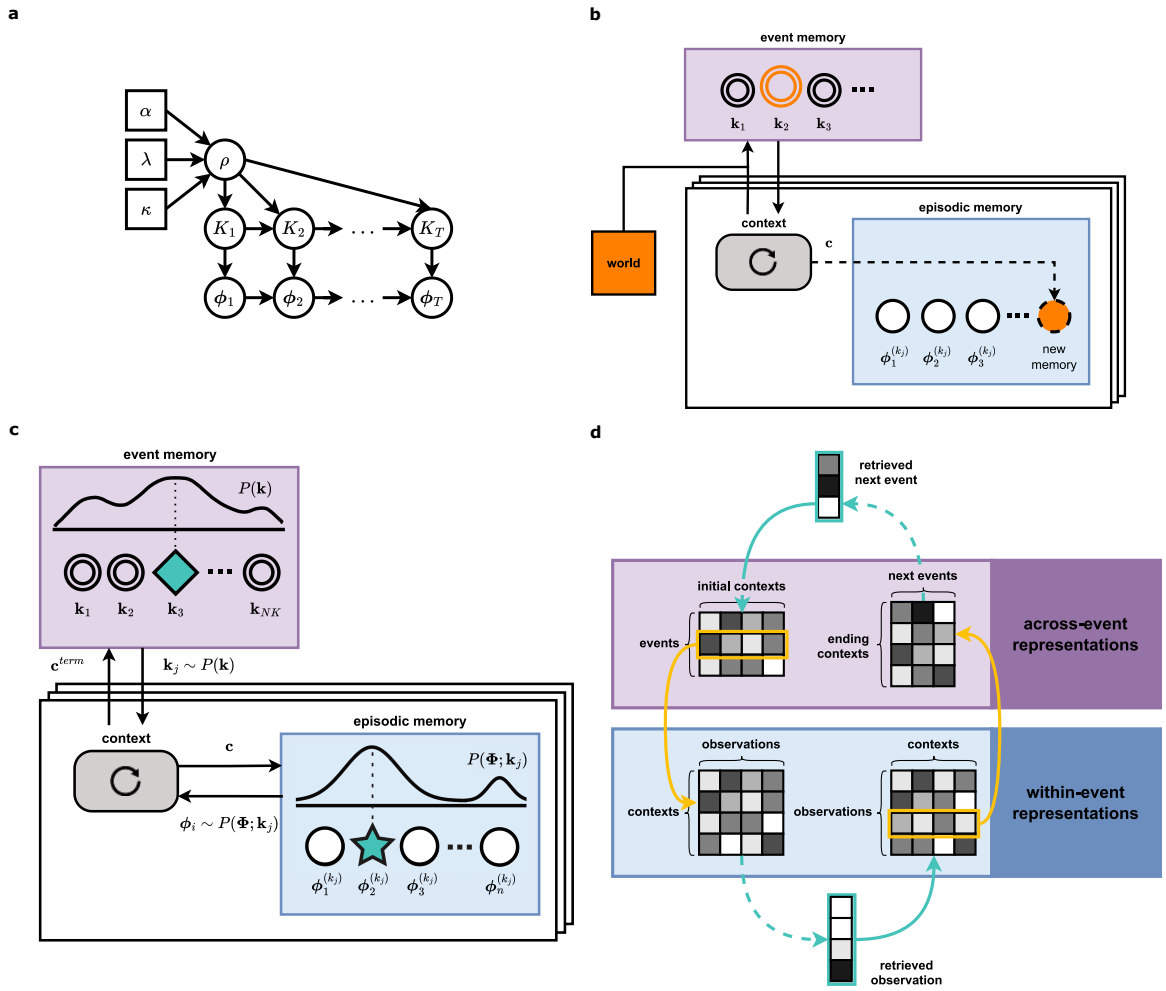

Figure 5.1: **Model schematics.** **(a)** We assume that a generative model underlies observations. Specifically, events dynamics follow a sticky Chinese Restaurant Process (sticky-CRP) parameterized by $(\alpha, \lambda, \kappa)$, such that at each time step t, the observation $\boldsymbol{\phi}_t$ is contingent on the current event K_t and the previous observation $\boldsymbol{\phi}_{t-1}$. A total of T observations are generated. **(b)** Encoding phase of a two-level model. An observation made at time t is first used to infer the currently active event (e.g., $\mathbf{k}_2$). It then updates the temporal context $\mathbf{c}_t$ of the current event. Memory encoding amounts to storing each temporal context present when a stimulus is seen. The first time a stimulus is encountered, a new memory is stored (circle with dashed outline). Each subsequent time the same stimulus is presented, the associated memory is modified (not shown). **(c)** Retrieval phase of a two-level model. The agent retrieves events and stimuli in an interleaved manner: first, it samples an event based on the last temporal context $\mathbf{c}^{term}$ (e.g., $\mathbf{k}_3$) and activates the corresponding event representations from the set of stored event models (lower boxes). It then freely samples one or more observations from $P(\boldsymbol{\Phi}; \mathbf{k}_j)$, the distribution of observations within the specified event $\mathbf{k}_j$. 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. **(d)** Representations learned by the two-level model after encoding. A retrieved event evokes an initial context associated with it (left bolded box), which seeds retrieval of within-event observations (bottom bolded box). Each retrieved observation is incorporated into the evolving temporal context (right bolded box) and affects subsequent retrievals (not shown). At the end of within-event retrieval, the ending context informs what the next event may be. The model samples a successor event (top bolded box) and repeat the retrieval process. Yellow lines indicate temporal contexts and cyan lines indicate knowledge units (observation, event). Solid lines correspond to matrix product operations and dashed lines correspond to sampling. The schematic only shows one retrieved observation from one event for clarity, but the model allows sampling of multiple observations and events.

actions; unlike SEM, however, event representations are further organized based on the temporal relationship between events (Fig. 5.1b,c).

For property 2 (compositional), we posit that representations across different levels form a recursive hierarchy (Fig. 5.1d). Indeed, human cortical representations suggest a nested hierarchy across timescales (Hasson et al., 2015) and modalities (Baldassano et al., 2016; Nelson et al., 2017). Information flows primarily from lower to higher levels at event boundaries (Baldassano et al., 2016). Here, we model two levels of representations operationalized explicitly as associative matrices. The upper level captures the relationship between events, while each structure on the lower level learns an event representation. Event representations can be composed together by traversing event boundaries on a higher level and traverse events of different timescales in an interleaved manner.

To achieve property 3 (realistic representation), we model the inference of event dynamics using the temporal associations readily encoded by the temporal context model (TCM; Howard and Kahana, 2002a), a standard model of episodic encoding and retrieval in humans. In TCM, a temporal context evolves to incorporate observations at encoding time, and episodic memory amounts to storing each context-stimulus association, which converges to the SR over one-hot encoded stimuli in the limit (Gershman et al., 2012; C. Y. Zhou et al., in press). During recall, the temporal context guides recall as TCM retrieves stimuli according to their similarity to the context. We extrapolate the analogy between the temporal associations of TCM and SR to successor features (SF; Barreto et al., 2017) by removing the one-hot encoding assumption. One immediate consequence is that the model can learn transitions in the latent feature space using distributed coding with analytically provable properties. Additionally, the inferred temporal dependencies facilitates probabilistic reasoning of incoming observations, which interacts with latent cause inference to organize experiences into events or options, as seen in Franklin et al. (2020), Lu et al. (2023) and Giallanza et al. (2024) except with recurrent neural networks. This allows stronger claims about generalizing this theoretical framework to a much larger set of tasks with arbitrary stimulus representations and improves the implementational feasibility in the brain (Rissman & Wagner, 2012; Z. Zhou et al., 2023).

For the rest of the chapter, we show that this framework rooted in episodic mechanisms explains a diverse set of human behavior that underlie and/or is indicative of the ability to generalize past experiences for adaptive use. It formalizes the interaction of episodic memory, event representation, and cognitive control through means of meta-reinforcement learning and Bayesian inference. Since the model parameters are fully interpretable, we also draw connections between algorithmic elements and cognitive processes to conceptually unify empirical findings across the three areas of study.

5.2 Results

Our current goal is to build a algorithmic account of structure learning and cognitive control where episodic memory plays a functional role. Specifically, we wish to demonstrate how a memory search model with hierarchical organization gives rise to human-like behavior in a variety of tasks that tap into different (and often interacting) aspects of knowledge representation and retrieval. The model therefore performs each task under the same condition as the original human participants. We then analyze model behavior similarly as in the original studies, and show replication of human behavior.

In the following sections, we first provide high-level intuition of the model to unify the three aspects of cognition of interest, and then corroborate its psychological feasibility by showing how it explains various behavioral findings with five experiments.

5.2.1 Model Intuition: Switching Plinko

To illustrate how this computational model ties episodic memory, event representation, and control together, we consider the sequential decision problem "Plinko": a ball is initially placed on a rectangular board with a few rewards and moves through the states (grids) according to a stochastic transition function (Fig. 5.2a). For a regular (unbiased) Plinko game, at each time step, the ball is equally likely to move diagonally down to the left or to the right by one row (i.e., $P(\text{left}) = P(\text{right}) = 0.5$), as long as either transition is possible. After observing a series of Plinko trajectories, the agent, who has to infer the transition probabilities using episodic memory, chooses an action (initial ball placement) to maximize the expected total rewards. However, we further add a twist: the true transition function of the environment may undergo unsignaled changes

(Fig. 5.2b). Specifically, $P(\text{left})$ can change in between trials. While the ball has a 50/50 chance of going left/right when averaged across all trials, given a specific trial, the actual transition probability may also be 80/20 or 20/80. This Switching Plinko task is similar to a nonstationary multi-arm bandit task with three modes, except that each action has a temporally extended consequence as opposed to an immediate reward. At decision time, the agent observes one random trajectory generated by the test game before taking an action.

Without the machinery of event segmentation, the model failed to make adaptive choices based on the observations. It learned a single representation of the game (e.g., a successor representation; C. Y. Zhou et al., in press) over the entire set of observations (Fig. 5.2c), which essentially summarized the mean transition dynamics over time, corresponding to an unbiased Plinko game. While such representation captures the "average" structure, it does not lead to the optimal choice if true environmental dynamics differ significantly. In particular, the model always chose the top-middle location to drop the ball, even if the game was right-biased with $P(\text{right}) = 0.8$, in which case the top-left placement would lead to more rewards in expectation. This is because the model uses the same representation for action evaluation: it performs generalized rollouts by continuously sampling the SR. The Monte Carlo estimate of the action value is the total discounted reward along an arbitrary rollout trajectory. As shown in C. Y. Zhou et al. (in press), this scheme corresponds to TCM retrieval with a drift rate of 1. Thus by assuming the incorrect representation in Fig. 5.2c, the model overestimates the value of the top-center placement.

On the other hand, with hierarchical event knowledge, the model was able to obtain a larger proportion of maximum possible rewards by picking the optimal action more often. It encoded multiple representations based on the observed trajectories using episodic encoding, where each representation captures a different "event", constituting a potential "option" in reinforcement learning terms. While all events were construed in a self-supervised manner as the model didn't know what or how many different games there were, some of them matched the underlying game structure (Fig. 5.2d). At the test

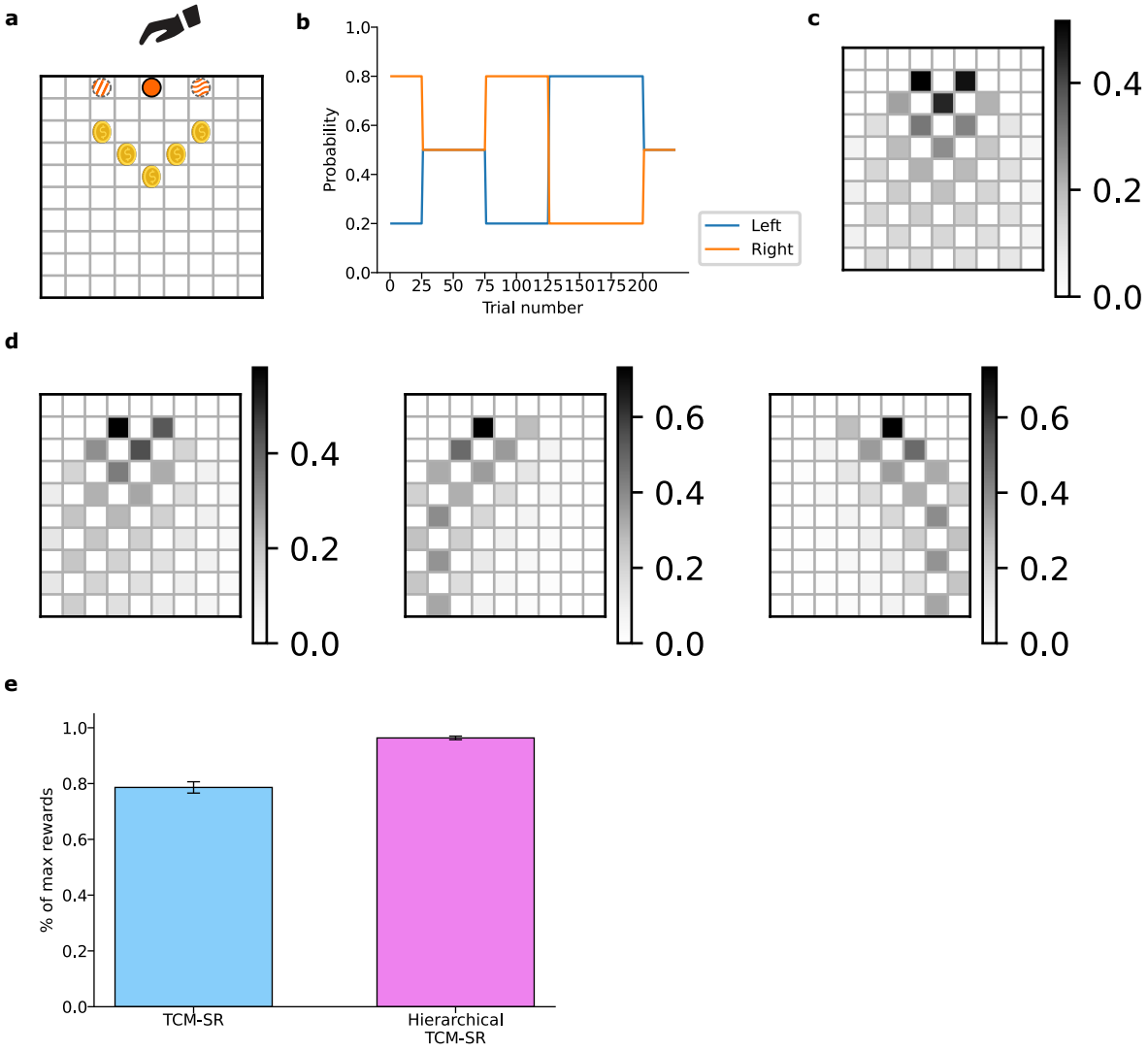

Figure 5.2: **Hierarchical episodic representations predict adaptive choice.** **(a)** Schematic of the Plinko task. The agent chooses between 3 actions, corresponding to 3 different ball placements on the top row (shaded/solid/wave-patterned circles). The dropped ball follows a stochastic trajectory down the board, collecting scattered rewards (gold coins) along the way. The goal is to maximize obtained rewards. **(b)** Sample transition probability during training (225 trials/trajectories). There are three games/modes: (unbiased) $P(\text{left}) = P(\text{right}) = 0.5$; (left-biased) $P(\text{left}) = 0.8, P(\text{right}) = 0.2$; (right-biased) $P(\text{left}) = 0.2, P(\text{right}) = 0.8$. Change points are unsignaled and may occur every 25 trials. **(c)** A model without event structure (TCM-SR) encodes a single SR to represent the Plinko games. Values correspond to the expected number of (discounted) visits to each board location, starting from the location of the solid circle in (a). **(d)** A hierarchical TCM-SR encodes multiple representations (SRs) corresponding to different events, characterized by different transition probabilities. Values correspond to the expected number of (discounted) visits to each board location, starting from the location of the solid circle in (a). Left to right: unbiased, left-biased, right-biased. Different training data and model initialization may result in different learning outcomes, but the three modes were consistently recovered. **(e)** On average, the hierarchical TCM-SR makes more optimal choices on a biased board using the appropriate event representation. In contrast, the model without event knowledge is worse at solving the task.

trial, the appropriate event representation was inferred and evoked to facilitate decision making. This model is able to outperform the other one with a few rollouts when the test game has a biased transition function (e.g., $P(\text{left}) = 0.2, P(\text{right}) = 0.8$; Fig. 5.2e).

This stylized decision task probes the synergic relationship between episodic memory, event representation, and cognitive control in several ways: first, events are represented by episodically encoded, temporally extracted representations, namely successor representations acquired from TCM encoding (Gershman et al., 2012; C. Y. Zhou et al., in press). Because SR separates transition dynamics from rewards and enables control, perceptual events may also act as generalizable task representations. Second, organizing episodic experience based on their latent cause supports adaptive decision making in a changing world. Simply amalgamating all past experience into one flat representation may result in inflexibility, while strategically form options allow efficient control given the task context. Finally, episodic retrieval mechanisms imply a decision-by-sampling approach to evaluate choices with respect to a specific event. We only showed the rollout-based evaluation here for clarity, and we refer the reader to C. Y. Zhou et al. (in press) for a extensive discussion on different retrieval dynamics and action evaluation schemes.

The following sections aim to establish the model's veracity with respect to actual human behavior in tasks that involve episodic memory, event cognition, and/or cognitive control. Franklin et al. (2020) pointed out five core functions that a comprehensive theory of event cognition should have: segmentation, learning, inference, prediction, and memory recall. Based on this proposal, we choose and present experiments in an order with three additional criteria: (1) they are as varied as possible within the current scope in terms of the task objective and experimental manipulation, so we observe a range of behavior from unsupervised event identification to memory scanning, covering the five core functions of event cognition; (2) they involve an assorted set of stimuli, ranging from discrete, hand-crafted word/image lists to continuous, naturalistic movies; (3) they are progressive in terms of the entailed functionalities, such that the model demonstrates competence on the most fundamental puzzles (e.g., event segmentation) before moving on to more complex ones that depend on the fundamental abilities (e.g., memory search).

5.2.2 Event Segmentation on Naturalistic Stimuli

Event boundaries define the start and the end of a sequence perceived as a single event. Prior studies have often operationalized event boundaries as timepoints marked by human participants indicating meaningful units, such as in movies (e.g., Michelmann et al., 2023; Zacks and Tversky, 2001). People readily perform segmentation without instructions or training in these studies, with moderate but significant agreement across participants (Franklin et al., 2020; Michelmann et al., 2023).

Following Franklin et al. (2020), we used the same video stimulus as the last one ("Washing Dishes") in Zacks et al. (2006) and created a set of unstructured scene representations from variational autoencoder[2](VAE; Kingma and Welling, 2014) embeddings as input to the model. This provides an unsupervised way to simulate model behavior on naturalistic stimuli without handcrafting stimulus representations and minimizes baked-in assumptions. The model subsequently inferred the event corresponding to each scene (frame) using maximum a posteriori (MAP) estimates. An event boundary was indicated at time t if the model's predicted event at time $t-1$ was different from its prediction at t. We used point-biserial correlation to quantify the degree of agreement between model boundaries that were discrete and human segmentation results that were averaged across the subjects in Zacks et al. (2006).

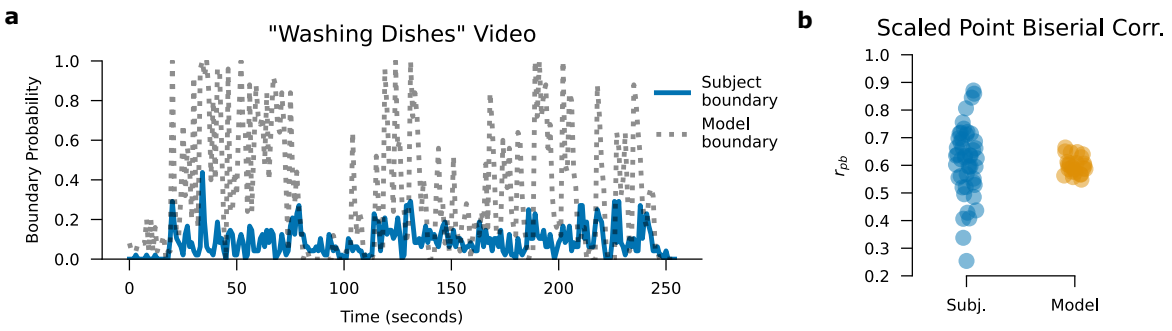

Figure 5.3: **Video segmentation. (a)** Model segmented event boundaries (dotted grey lines), averaged across 25 model instances, compared with human segmentation frequency (solid blue lines). **(b)** Scaled point biserial correlation coefficients for individual human subjects and the model, comparing discrete boundaries to average human segmentation results. Model segmentation falls within the interquantile range (IQR) of human performance.

[2]The specific variational autoencoder implemented can be found at https://github.com/ProjectSEM/VAE-video. The same VAE is used to pre-process the film clip in Section 5.2.6.

The model's event boundaries closely resemble one's indicated by human subjects, with a scaled point biserial correlation of $r_{pb} = 0.600$ (s.e.m. $= 0.006$; Fig. 5.3a). This falls within the interquantile range of particpant responses ($r_{pb} = 0.614$, s.e.m. $= 0.127$; Fig. 5.3b), and is considerably higher than the structured event memory model (Bezdek et al., 2022; highest $r_{pb} = 0.46$; not shown). A permutation test ($N = 1000$) also suggests significant difference between model boundaries and chance ($p < 0.001$). Furthermore, the estimated log-probability of event boundaries highly correlates with the empirical probability of subject-indicated boundaries ($r = 0.23$, s.e.m. $= 0.01$).

Crucially, the only difference between our model and SEM that is relevant to this task is how events are represented in episodic memory: SEM trains a separate RNN for each event, while ours learns the successor features via episodic encoding of TCM (Howard & Kahana, 2002a). The machinery for perceiving events (latent cause inference) as well as the sensory inputs (VAE embeddings) are exactly the same. The better correspondence to human behavior of our model therefore suggests SF as a more accurate account of naturalistic event representations in humans. Next, we examine the consequence of episodic encoding on event cognition in a different setting, where events are formed based on temporal associations instead of predictive uncertainty.

5.2.3 Event Segmentation on Community Structure

Event perception from naturalistic observations likely only constitutes one specific instance of experience partitioning; it has been hypothesized that a more general representational clustering underlies structure learning (Schapiro et al., 2013). Earlier models, such as the Event Segmentation Theory (Reynolds et al., 2007; Zacks et al., 2007), posit that event boundaries are signaled by large prediction errors or surprise. While people's behavior support such a view (Axmacher et al., 2010; Reynolds et al., 2007; Zacks et al., 2011), recent studies have suggested temporal associations as an additional source of representational clustering (Schapiro et al., 2013). In particular, subjects were more likely to cluster observations in a way that reflects the underlying graph structure, despite uniform predictability and controlled local statistics (Fig. 5.4a). Their neural activity also mirrored the hidden structure, which resembled the SR (Stachenfeld et al., 2017).

Different from Stachenfeld et al. (2017), however, in our case, instead of learning a single SR that potentially encompasses multiple events, the model learned an SF for each event it identified in a self-supervised manner.

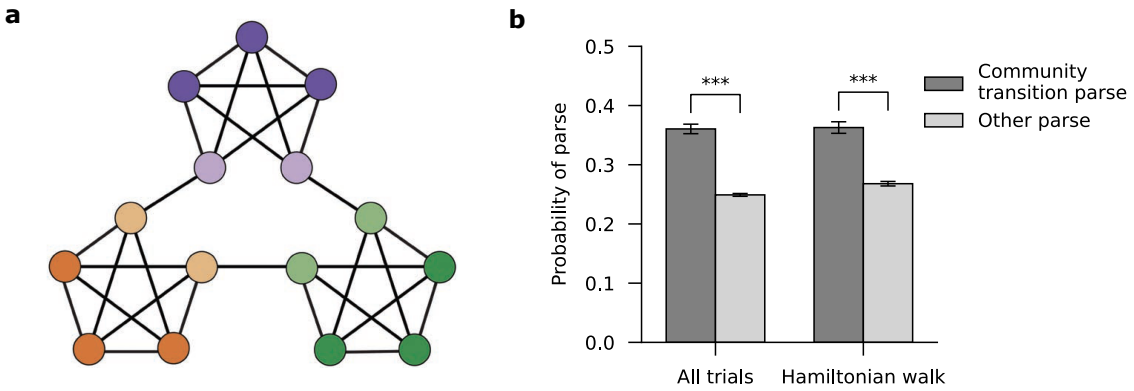

Figure 5.4: **Event boundaries at graph community boundaries. (a)** The graph structure from Schapiro et al. (2013) containing three clusters ("communities"). Sequences of stimuli were generated by drawing from this graph, either through random walks (equal transition probability to each neighbor) or Hamiltonian walks (each node is visited exactly one). **(b)** The proportions of times the model parsed at a community transition (dark grey) or within a community (light grey) out of all opportunities to do so. Results are shown for both all trials (left) and Hamiltonian walks only (right). *** $p < 0.001$. Error bars indicate ± 1 s.e.m. across all experiments.

Our model matched the human behavior on this community structure parsing task. It exhibited a higher probability to parse (i.e., indication of an event boundary) at community transition points, both on Hamiltonian walks ($t(49) = 13.13, p < 0.001$; Fig. 5.4b) where the agent could not use local statistics to infer the generative structure, and over all trials including random walk trials ($t(49) = 9.44, p < 0.001$; Fig. 5.4b). The numerical parse probabilities were also comparable with human data reported in Schapiro et al. (2013).

These findings lend further support to the hypothesis that the SR/SF is a promising candidate to explain human event cognition and hierarchical structure representation. In particular, it not only replicated the pattern of parse, which SEM also achieved, but also replicated the average probability of parse in human subjects. The improvement is again attributed to the event representations alone. Together with the results from simulation 1 (naturalistic stimuli), we have thus shown SF to be more accurate than RNNs in capturing episodic memory's contribution to the formation of the hierarchical organization of the

mind. Having established that the model encodes information in a hierarchical fashion closely resembling humans, we now investigate its consistency with the retrieval outcomes in humans.

5.2.4 Event Representation and Hierarchical Episodic Retrieval

Previous research has suggested behavioral benefits of hierarchically structured representations of temporally extended observations. Organizing experiences by events – either naturally segmented or cued by experimentally manipulated contexts - may facilitate more accurate episodic recall of studied word lists (Hanson & Hirst, 1989; Pettijohn et al., 2016) and sequences of naturalistic observations (Gold et al., 2017; Zacks et al., 2006). For instance, Pettijohn et al. (2016) showed that participants who were exposed to lists of random words in different physical rooms or conceptual contexts had better overall recall accuracy.

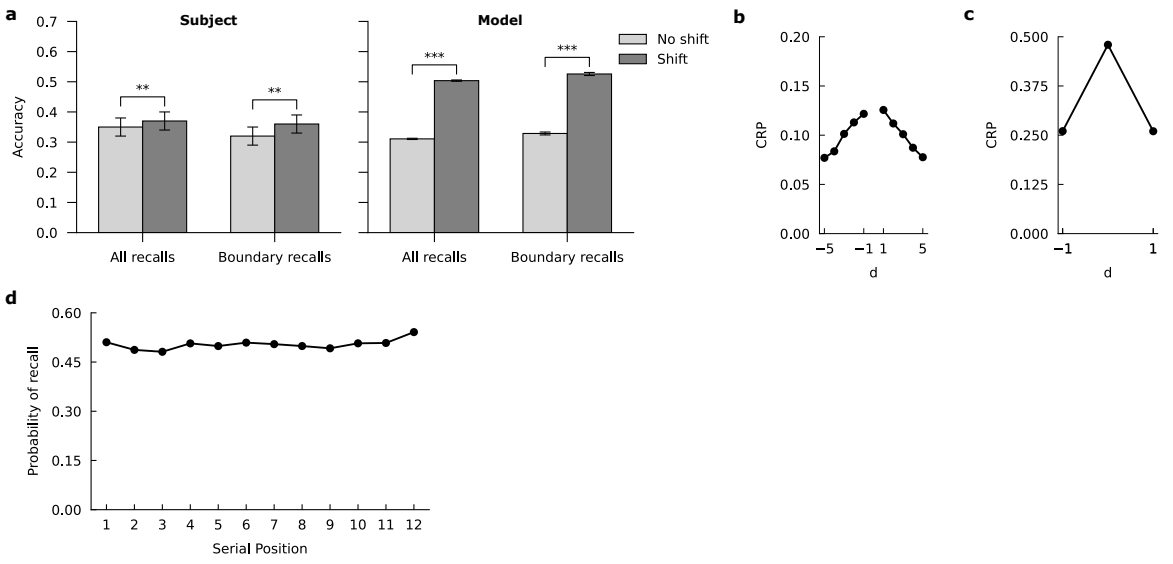

Figure 5.5: **Enhanced free recall with event structure. (a)** Left: human subject recall accuracy as a function of the condition (*shift*/*no-shift*) and the recall type (all recalls, recall of a boundary item). Right: model free recall accuracy. ** $p < 0.01$; *** $p < 0.001$. Error bars indicate ±1 s.e.m. across all subjects/trials. **(b)** Within-event conditional recall probability (CRP) curve produced by the model, averaged across all trials. **(c)** Across-event conditional recall probability (CRP) curve produced by the model, averaged across all trials. 0 indicates recalling from the current event, -1 indicates recalling from the previous event, and 1 indicates recalling from the next event. **(d)** Within-event serial position curve produced by the model, averaged across all trials.

Here, we simulated our model using Experiment 2 from Pettijohn et al. (2016),

where an agent was assigned to one of two conditions. In the *shift* condition, word lists of length 12 were presented in two alternating contexts, corresponding to event 1 and 2. In the *no-shift* condition, all word lists were presented in the same context. Human participants recalled significantly more items in the shift condition than in the no-shift condition (Fig. 5.5a). Consistent with human behavior, the model exhibited a higher recall accuracy in the shift condition for boundary items ($t(398) = 63.43, p < 0.001$) and across all items ($t(398) = 28.05, p < 0.001$; Fig. 5.5a). Prior studies have suggested that events improve memory recollection because they prevent retrospective interference; here, we explicitly modeled this idea with a large update to the temporal context at event boundaries, such that it incorporates little pre-boundary information. Since events are represented by associations between contexts and observations, two different events would then share less overlap in dynamics compared to within each event, where contexts drift in a more moderate manner. This helps to keep event representations separate and essentially parses long sequences of observations into chunks that are easier for recall.

Moreover, the order in which the model recalls matched the pattern observed in a broader range of free recall studies. The probability distribution of the relative position of consecutive recalls *within* each event qualitatively agrees with the temporal contiguity effect observed in humans (Fig. 5.5b). This shouldn't be surprising given our model is based on TCM and a temporal context, which captures the temporal contiguity effect. Importantly, this effect seemed to extend beyond individual lists and apply to inter-list recall patterns as well (Fig. 5.5c), which has also been found in humans (Howard et al., 2008). Moreover, the model qualitatively exhibits a higher probability of recalling items at the end of each list (Fig. 5.5d), similar to the recency effect in humans (Howard & Kahana, 1999). None of the previous models of event cognition have shown these memory effects to our knowledge. Our model is able to capture both the accuracy and order of recall because an evolving temporal context mediates all encoding and retrieval. Specifically, observations within an event update and become associated with the context at their presentation, while the same context shifts at event boundaries to encapsulate the temporal associations across events. Reinstatement of the context at the beginning of each event thus allowed

the agent to reproduce human-like recall behavior over a longer timescale, which has been found neurally (Baldassano et al., 2016). The fact that implementing one mechanism of event perception enables hierarchical retrieval also suggests a close link between event representation and memory organization in humans, which we explore further in the next section.

5.2.5 Event Representation and Memory Organization

Organizing episodic memory by events improves overall recall, but doing so also introduces additional biases. One characteristic impairment is the reduced memory of relative temporal order around an event boundary (DuBrow & Davachi, 2013; Ezzyat & Davachi, 2011; Zacks & Tversky, 2001). Concretely, DuBrow and Davachi (2013) used the frequency of recall transitions in the original serial order as a measure of how well the presentation order is maintained in episodic memory. They found that participants made significantly fewer correct transitions immediately following a boundary compared to right before a boundary or within an event.

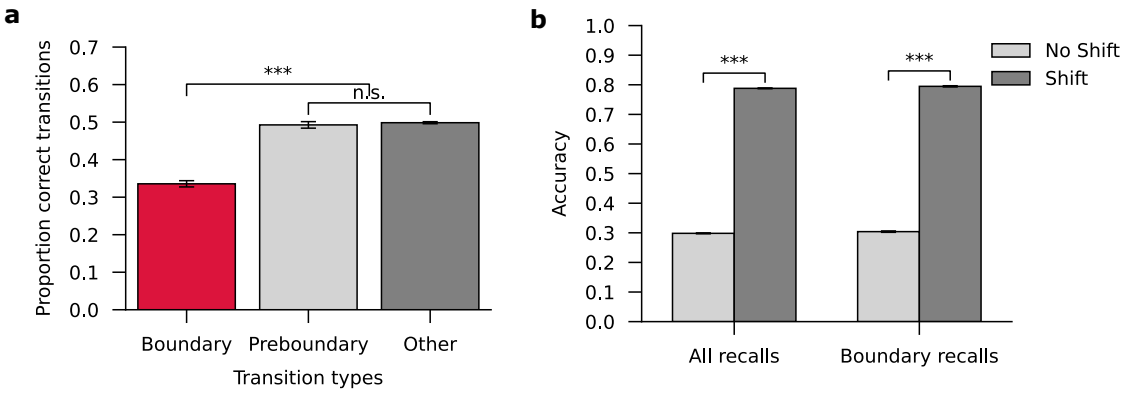

Figure 5.6: **Reduced order memory at event boundaries. (a)** The proportion of correct transitions in the model's consecutive recalls as a function of the pre-transition item and the post-transition item. n.s. $p > 0.05$; *** $p < 0.001$. Error bars indicate ± 1 s.e.m. across all trials. **(b)** Model free recall accuracy as a function of the condition (*shift*/*no-shift*) and the recall type (all recalls, recall of a boundary item). *** $p < 0.001$. Error bars indicate ± 1 s.e.m. across all trials.

We focused on Experiment 1 from DuBrow and Davachi (2013) and simulated the same paradigm by presenting random stimuli in two alternative contexts (the original experiment used images of two categories, faces and objects) followed by a free recall task. Despite the lack of semantic knowledge, the model behaved strikingly similar

to human participants. It too exhibited a compromised order memory after an event boundary (Boundary vs. Pre-boundary: $t(1603) = -13.08, p < 0.001$; Boundary vs. Other: $t(2525) = -19.34, p < 0.001$; Fig. 5.6a), and similarly no difference between pre-boundary and other types of recalls ($t(2573) = -0.34, p = 0.73$; Fig. 5.6a). The transition probabilities were also much closer to the average human subject reported in DuBrow and Davachi (2013) than SEM, which never made correct transitions more than 15% of the time. Furthermore, our model showed the same improvement on recall accuracy with an event structure compared to without any event structures (Overall: $t(3999) = -233.82, p < 0.001$; Boundary: $t(3999) = -156.15, p < 0.001$; Fig. 5.6b), just like the human subjects in DuBrow and Davachi (2013), albeit the average recall rate was higher than in humans (79% vs. 63%).

This lends further evidence to the hypothesis that the degraded order memory is not due to worse associative memory (Heusser et al., 2018), but rather a natural consequence of event representation on memory organization. In particular, while a temporal context slowly drifts during encoding of an ongoing event, at an event boundary, the model's temporal context shifts abruptly, causing it to lose most of the information from the preceding event. The consequence is a double-edged sword: segregating observations into clusters reduces interference and improves overall recall as shown by the previous task, but the points of segregation become discontinuities that are poorly recovered. Our model's ability to faithfully reproduce both behavior thus formalizes the underlying mechanism, which previous work such as SEM and EGO does not explicate.

5.2.6 Memory Search of Temporally Extended Events

In the last study, we examine how cognitive control may be implemented using hierarchical episodic mechanisms. Here, we consider the task of memory scanning after watching a film clip (Michelmann et al., 2023). Specifically, in memory scanning, the agent is cued with one observation and needs to report *where* in the original sequence a different observation occurred. The response time of human subjects on this tasks suggests that memory search happened in a hierarchical manner: they used event boundaries as access points of memory retrieval and searched within each event. When the current

event seemed sufficiently unpromising, they skipped to the next event, again by accessing the event boundary (Michelmann et al., 2023). Our model explicitly implements this search procedure at retrieval time with its recursive hierarchy of representations (Fig. 5.7a, Fig. 5.1d). Thus if it has successfully organized memories by learning an event structure, its search time should mirror the human behavioral patterns.

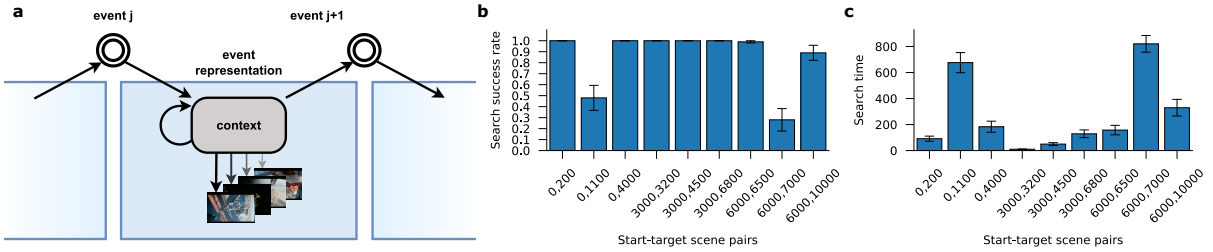

Figure 5.7: **Schematic of memory search performed by the model. (a)**The model accesses an event to evoke the associated initial context and "zooms in". The temporal context evolves as observations are sampled. Sampling terminates when the target observation is found or the skipping threshold is exceeded. If the target is not found, the agent samples the next event using the ending context and continues its search. The dynamics within an event representation can be thought of as the inner loop of an HRL model, while the dynamics across events correspond to the outer loop. **(b)** Average rate of successfully finding the target scene given the start scene. Error bars indicate ±1 s.e.m. across all models and trials. **(c)** Average time the model spent to search for the target scene given the start scene. Error bars indicate ±1 s.e.m. across all models and trials.

Indeed, that is what we found from the simulations. We assumed that each retrieval took a fixed amount of time and used the number of samples as the model's response time RT_{model}. First, we checked whether successful searches took less time as in humans by fitting a simple regression

$$RT_{model} \sim 1 + success.$$

Search is much faster when success rate is higher ($\beta_{success} = -793.82, p < 0.001$; Fig. 5.7b,c).

Next, we address the question whether the model searches in a similar manner as people - that is, it retrieved memory with respect to the encoded event structure and used event boundaries to speed up search when necessary. To operationalize this idea, we fitted the same stepping-stone model as in Michelmann et al. (2023):

$$RT_{model} \sim 1 + nEB + distEBpre + distEBpost,$$

where nEB is the number of event boundaries between the pair of observations in scanning/simulation, $distEBpre$ is the number of frames from the target scene to the previous event boundary, and $distEBpost$ is the number of frames from the target scene to the next event boundary. The number of event boundaries was a significant contributor to overall response time in humans ($\beta = 98.75, p = 0.016$), supporting the hypothesis that memory search uses event boundaries as access points (Michelmann et al., 2023). Our model demonstrated this effect as well ($\beta = 1.91, p < 0.001$). Additionally, the distance to the previous event boundary was significant during memory scanning (human: $\beta = 47.34, p < 0.001$; model: $\beta = 0.98, p < 0.001$), reflecting systematic sequential memory retrieval. In contrast to human data, however, we found the distance to the next event boundary affecting the model's response time (memory scanning: $\beta = -1.79, p < 0.001$). This is likely because the model frequently "skipped" too much in time and subsequently searched in the backward direction. In comparison, humans may have more elaborate control mechanisms or more refined representations to guide their search and avoid jumping back-and-forth (the model has no built-in semantic knowledge).

An alternative account of memory search simply uses the distance between each queried pair of observations to predict search time, which Michelmann et al. (2023) operationalized as the duration model:

$$RT_{model} \sim 1 + dur.$$

Note this account doesn't include any information about event structure and relies on the pairwise distance alone. This would correspond to a model like the current proposal but without any hierarchical representation. Despite penalties in AIC for more parameters, the stepping-stone model (AIC = 9513) explains the variance in model search time better than the duration model (AIC = 9693, ΔAIC = 180).

The stepping-stone search behavior arises from the model's hierarchical organization, where events are associated with their beginning context (event-context associations) *and* the ending context of the previous event (context-event associations). In particular, recalling an event amounts to evoking its beginning context, which acts as the basis

of subsequent within-event retrieval as it evolves according to TCM dynamics. When within-event retrieval terminates, the temporal context provides means for the agent to sample (recall) a next event using the learned context-event associations, thereby traversing up the hierarchy.

5.3 Discussion

In the current work, we proposed a symbolic, process-level model to formalize how the interaction of episodic memory, event cognition, and cognitive control give rise to effective representation and reuse of past experiences. By equipping the basic TCM-SR model (C. Y. Zhou et al., in press) with hierarchical organization, we have created a powerful framework to explain unsupervised structure learning, episodic recall, and hierarchical memory search in humans. Importantly, this framework explains human behavior in various tasks that tax on a diverse set of abilities, including segmentation, statistical learning, structure inference, episodic recall, and goal-directed behavior, thereby reconciling a broad range of experimental findings under a unified account of high-level cognition.

Our proposal departs from the formalism of previous models of structure learning and control, which often rely on recurrent neural networks (e.g., Franklin et al., 2020; Giallanza et al., 2024; Lu et al., 2022), and constitutes a purely symbolic approach. It accounts for similar behavior just as well, if not better, with a higher degree of interpretability and direct connections to various theoretical and empirical results. Specifically, on the theory side, at the core of our model lies a standard theory of episodic encoding and retrieval – TCM (Howard & Kahana, 2002a; Polyn et al., 2009a). The equivalence between TCM encoding and SR learning (Gershman et al., 2012), combined with TCM recall, leads to a surprising sample-based account of decision making, bridging episodic memory and adaptive control (C. Y. Zhou et al., in press). Here, we further establish an equivalence between memory encoding and SF (Bailey & Mattar, 2022; Barreto et al., 2017), and incorporate latent state inference to automatically organize encoded memory based on event structure (Franklin et al., 2020; Gershman et al., 2014) as options/event models. Episodic retrieval maneuvers within this learned hierarchy of representations

to reconstruct the past and predict the future in a model-based manner, similar to the proposal of Konidaris (2016). Cognitive control accomplished this way is thus consistent with earlier theories of category learning through means of auto-associations (McClelland et al., 1995). In terms of empirical evidence, key model features are psychologically and/or biologically plausible: the temporal context account explains a range of episodic recall patterns over various timescales (e.g., Polyn et al., 2009b, 2011), and correlates with activities in the hippocampal complex (Herweg et al., 2018; Sakon & Kahana, 2021). In addition, SR-like representations capture place cell activities (Alvernhe et al., 2011; Ekman et al., 2023), while the hippocampus has also been hypothesized to engage in inferring latent states (Aggleton et al., 2007; Gershman et al., 2015). Just like the brain, within-event and across-event memory are organized in a nested hierarchy (Baldassano et al., 2016; Hasson et al., 2015; Radvansky, 2012), where event boundaries act as ladders between adjacent levels of representations (Michelmann et al., 2023; Zalla et al., 2003). Finally, the representations involved align with the most plausible implementation given what we know about the neurobiology of the hippocampus, namely that they are distributed and can be learned through Hebbian learning with rate decay.

Besides unifying the theoretical and empirical findings, our model also offers a solution to several conceptual taxonomies. First, while episodic control has been proposed as the "third way" of control in addition to model-free and model-based control (Lengyel & Dayan, 2007), here we make a more elaborate argument about how model-free learning and episodic mechanisms account for human behavior that previously require a separate episodic memory module to explain (Giallanza et al., 2024; Lu et al., 2023). Our proposed framework therefore offers a more parsimonious explanation of the cognitive process underlying various behavior that range from event segmentation to control.

Second, while hierarchical reinforcement learning often choose between state and temporal abstraction (Tomov et al., 2018), our model performs both. Specifically, Bayesian inference of the latent cause partitions observations into event clusters. This results in a smaller set of abstract states, as each observation (e.g., a frame in a movie clip) can be naively treated as a state in a large and possibly infinite state space. Additionally, each

partitioned event cluster, as well as the inter-event relation, is represented by a world model that subserves temporal abstraction (Machado et al., 2023). This world model, viz., successor features, not only represents an *understanding* of the environment, but also enables *control* of the environment through episodic sampling, as event representations can function as compositional options. The model thus also suggests one mechanism of action chunking (Sakai et al., 2003). Notably, these different types of abstractions are linked together by an evolving temporal context, which started as part of a descriptive model of episodic memory. Yet we have shown here as well as in C. Y. Zhou et al. (in press) that the descriptive process has a rational basis: the purpose of memory is not simply remembering, but also learning, organizing, and adapting.

In addition to abstractions of different types, we also account for abstractions of different timescales. Humans organize experiences in a nested hierarchy of timescales from coarse to fine (Barreto et al., 2017; Kurby & Zacks, 2008; Murray et al., 2014), such that event boundaries trigger post-hoc encoding (Ben-Yakov & Dudai, 2011; Ben-Yakov et al., 2013) and help reinstate long timescale representations (Baldassano et al., 2016; Zadbood et al., 2017). Relatedly, a multi-scale ensemble has been previously proposed to account for firing patterns in predictive maps (Eichenbaum et al., 1987), and to address the policy dependency of single SRs (Momennejad et al., 2018). What we propose here implements multi-scale successor representation/feature, where the exact timescale is not pre-determined but actively tailored to the experience (e.g., lower discount factor/slower drift in a continuous space). As the result, SFs on the "finer" timescale can inform partial policies within events, while temporal associations on the "coarser" timescale compose them together to form more flexible policies that may behave differently than the originally experienced sequence. Even though we only consider a model with two levels of timescale here, the building blocks of this recursive architecture are clearly stackable. Future work may extend the model hierarchy further and draw more detailed connections to timescales observed in the brain.

The framework integrates many established theories in areas from long-term memory to model-based control; more importantly, it provides the scaffolding for theoretical

extensions in these areas. Our current formulation leans towards the empiricist view of cognition, as little prior knowledge is assumed in most cases and all knowledge comes from a single pass of experiences. However, humans likely possess some form of native representation of the world, such as semantic understanding, that affect episodic encoding and retrieval. Humans also use emotions to guide information processing, which may further improve control efficacy (C. Y. Zhou et al., in press). Many variants of TCM take these factors into account, and our framework allows direct incorporation. In particular, extending the TCM module may explain the effect of semantics with CMR (Polyn et al., 2009a), the effect of repeated learning with CMR2 (Lohnas et al., 2015), the effect of emotional modulation with eCMR (Talmi et al., 2019) and CMR3 (Cohen & Kahana, 2022), or false memory and recall control with TCM-A (Sederberg et al., 2008). Although none of these models are hierarchical, our framework opens up opportunities for them to capture behavior on a longer timescale and over a richer structure. Going beyond memory, the explicit event representations also allows detailed comparison between human learning/planning mechanisms and normative models of knowledge representation, such as policy compression (Lai & Gershman, 2021) and knowledge retention (Kirkpatrick et al., 2016).

Experimental Predictions

Our computational account translates into empirically testable hypotheses, which are critical avenues of future exploration. For instance, our proposal aligns with the hypothesis that episodic and semantic memory are not distinct systems but rather ends of a continuous spectrum. Specifically, episodic memory of past experiences undergoes gradual semantization (e.g., as event representation stabilizes), such that the original temporal context is lost and the relational structure is distilled (Duff et al., 2019; Habermas et al., 2013). As learning advances, our model predicts a shift from relying on episodic memory samples to semantically driven associative responses recapitulated in successor representations/features, much like what has been observed in Nicholas et al. (2022) in one-step tasks. Since TCM-SR provides the means for solving sequential decision tasks via episodic sampling, we hypothesize the same effect also applies to multi-step tasks,

such as the next-state prediction task (Beukers et al., 2024).

By replacing neural networks in prior work with successor features, our proposal also implies that the former may effectively come to implement the latter through training. RNN has been previously shown to compute SR and SF from random walks and natural trajectories (Fang et al., 2023), so it is possible that neural circuits also estimate the successor feature over a long temporal sequence with event structure. Future work may directly compare the event representations of neural network models (e.g., Bezdek et al., 2022; Giallanza et al., 2024; Lu et al., 2023) or hippocampal activities in the brain with corresponding SFs for equivalence. Because world models like SF typically have more straightforward interpretations, viewing the learning outcome of neural networks in light of them could in turn improve our understanding of neuro-symbolic models.

Unlike neural network models, event dynamics captured by SR/SF are all in terms of linear transformations, leading to an unexpected implication that people may naturally approximate nonlinear dynamics using a series of simpler linear functions (i.e., piecewise-linear approximation). Furthermore, this predicts that event boundaries are formed when substantial deviations from a (near-)linear system are observed. While this hypothesis has not been systematically investigated before, early studies showed that people treated curvature extrema as natural segmentation points (Shipley & Maguire, 2008; Singh & Hoffman, 2001). Our model predicts that, more generally, if a series of observations were generated using nonlinear dynamics (e.g., visual stimuli in a latent feature space), participants' segmentation probability should correlate with the underlying curvature. The smoothness of transition dynamics should additionally predict the degree of agreement across participants, where increasingly noisy/jumpy transitions cause more uncertainty in event segmentation. It is worth noting that similar ideas have been applied to reverse engineer nonlinear RNN dynamics using a switching linear dynamical system and fixed points as the basis of linearization (Fox et al., 2009; Smith et al., 2021). This suggests a close relationship between our symbolic approach and previous neural network models, and highlights the theoretical potential our framework offers.

Moreover, our model formalizes the idea that episodic memory serves an adaptive

purpose, where episodically encoded SFs may constitute behavioral options (Barreto et al., 2021). One way to empirically test this hypothesis is through a novel transfer learning task: first, the subject observes a sequence of unannotated actions, which (unbeknownst to the agent) achieves an overall objective $\mathcal{O}$ by accomplishing subgoals $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ in a certain temporal order (permutation) $\mathcal{P}$. The objective may be obtaining a reward in a game and the subgoals may involve manipulation of the game environment. The ostensible task is to segment the observation sequence and/or to make free recall of game elements. Then in a surprise test, they need to solve a different task with a different object $\mathcal{O}'$, which can be achieved by solving a subset of the subgoals $\mathcal{G}' \subseteq \mathcal{G}$ with a temporal order $\mathcal{P}'$. The model predicts that successful solutions entail successful segmentation (i.e., segments match subgoal solutions). Additionally, it predicts that worse order memory at segmentation points correlates with higher success rate when $\mathcal{P}$ and $\mathcal{P}'$ are further apart. The second prediction is a direct result of abrupt changes to the temporal context at event boundaries, which helps to segregate representations and increase compositional flexibility.

The proposed transfer learning task may further be used to uncover subgoal evaluation biases in planning. Specifically, people's evaluation and choice in sequential tasks were predicted by their episodic recall patterns - items with higher recall rates (e.g., the first few observations) were also weighted more in evaluations, and decisions were best explained by what people actually recalled *within standalone events* (C. Y. Zhou et al., 2024). This suggests shared neural infrastructure between episodic memory and decision making as TCM-SR predicted. Here, we have further demonstrated that a hierarchical TCM-SR accounts for the stepping-stone search process *across correlated events*. Together, these two sets of empirical findings suggest that people may evaluate an extended plan by concatenating subplans, each of which exhibit footprints of episodic retrieval - the start is weighted more as events are indexed and traversed using the beginning context, and subgoals are quickly discarded if the front portion seems unpromising.

In summary, we put forth an integrated computational framework to explain how humans efficiently organize and learn from past experiences. We demonstrate its ability

to reproduce and thus explain various human behavior in event cognition, memory recall, and cognitive control, all under a unified mechanistic account. It is our hope that the current work could bridge the existing work in relevant areas to help build a more coherent understanding of high-level cognition.

5.4 Methods

5.4.1 Model

We propose a process-level model that integrates episodic mechanisms, event representation, and model-based control. In a nutshell, the model formalizes the hypothesis that humans segment and organize observations into "events" based on the temporal structure encoded by episodic memory. The resultant event representation is an predictive map that can be used to infer environmental dynamics, modulate memory reconstruction, and inform adaptive behavior.

The model consists of three main elements: temporal context model (TCM; Howard and Kahana, 2002a), successor representation/features (SR/SF; Barreto et al., 2017; Dayan, 1993), and latent cause inference (LCI; Gershman et al., 2014). They respectively function to drive encoding and retrieval of relevant experiences, learn representations from experiences, and structure experiences. The simulations use a two-level model, but we note that the architecture can be further extended.

To start with, we assume that an HMM generates an observation (movie scene, visual stimuli, word) at each time step by drawing from the specific distribution associated with the currently active latent variable (event). The identity of the latent event is determined by a sticky-CRP prior, hidden from the agent. See Section 2.6 for details of event inference using LCI.

5.4.2 Episodic Representation

We model the representation of each event k using the SR or SF learned via the encoding process of TCM (Howard & Kahana, 2002a).

One way SR or SF may be learned by humans is through episodic encoding mediated by a temporal context, as instantiated by TCM (Gershman et al., 2012; C. Y. Zhou et al., in press). TCM is a standard model of memory encoding and retrieval, originally

proposed to explain various patterns in human's free recall data, such as the order in which stimuli are recalled. The centerpiece of TCM is a drifting temporal context $\mathbf{c}$ that evolves according to Eq. 1.1 with $\rho = \rho_{enc}$ and $\beta = \beta_{enc}$. At time t during encoding, the temporal context $\mathbf{c}_t$ is (partially) updated by a new observation $\mathbf{c}^{\text{IN}}$, the degree of which is controlled by the model parameter $\beta_{\text{enc}} \in (0, 1)$. Conversely, the degree in which past contexts are retained is determined by ρ_{enc}, such that $\rho_{\text{enc}} + \beta_{\text{enc}} = 1$. The temporal context is thus a recency-weighted average of past experiences, with decay rate equal to ρ_{enc}.

The drifting temporal context guides the representation of experiences by associating itself with each observation across encoding timesteps. Formally, TCM learns an associative matrix that binds observations to contexts as Eq. 2.8. Gershman et al. (2012) showed that encoding this matrix is equivalent to learning the successor representation, while the temporal context is equivalent to the eligibility trace if stimuli are one-hot encoded and only observed once. Under this condition, the encoded memory according to TCM essentially captures the temporal structure of observations (i.e., which stimuli precede a given stimulus). While strictly applying the Hebbian learning could lead to an unbounded representation, adding a weight decay could prevent the problem while preserving the model's biological plausibility. As the result, the associative matrix converges to the true SR after extensive experience (i.e., $\mathbf{M}^{\text{CS}} \to \mathbf{M}$, with $\rho_{enc} = \gamma\lambda$), giving the same outcome as the temporal difference (TD) learning algorithm.

Furthermore, if the one-hot encoding assumption of $\mathbf{x}$ is relaxed to an arbitrary representation $\boldsymbol{\phi}$, TCM learns the *SF* over the latent feature space instead of SR over the observable state space. See Section 2.5 for details.

With the machinery to infer hidden events and form event-specific episodic representations, we now discuss how events are segmented in our model.

5.4.3 Event Segmentation

The model uses a local maximum a posteriori (MAP) estimate to infer the latent event. As discussed in details in Section 2.6, the MAP event $\hat{K}_t$ maximizes the posterior probability $P(K_t|\mathbf{\Phi}_{1:t})$, or the probability of such event being active at time t given

the historical observations. For computational tractability, we assume that instead of performing inference on the entire observation history, the agent only computes the posterior within a moving window of the most recent observations.

The model assumes a sticky-CRP prior as the generative model, so the prior in Eq. 2.13 is given by Eq. 2.12. The ground truth sCRP parameters are assumed unknown, and they are optimized as hyperparameters of the model. Next, we address the problem of estimating the conditional probability $P(\boldsymbol{\phi}_t|\boldsymbol{\Phi}_{1:t-1}, K_t)$.

5.4.4 Episodic Inference of Event Dynamics

We infer the conditional likelihood of an observation using successor representations, which translates to successor features (SF; Barreto et al., 2017).

If the probability of one-step transitions is known, computing $P(\boldsymbol{\phi}_t|\boldsymbol{\Phi}_{1:t-1}, K_t)$ is trivial by looking up the transition function. When the underlying transition dynamics is not known, which is most likely given a complex, naturalistic setting, the quality of knowledge organization depends on the agent's own inference. Previous models use recurrent neural networks (RNNs) to solve this problem (Franklin et al., 2020; Giallanza et al., 2024; Lu et al., 2022); here, we take a different route. We first note that SR can alternatively be incrementally learned via TD learning. Each TD update follows Eq. 2.4.

As in TCM, $\mathbf{c}_t$ is the temporal context (i.e., eligibility trace) at time t, and $\mathbf{x}_t$ is the one-hot encoded stimulus experienced at time t.

Assuming $\mathbf{M}$ is invertible, Eq. 2.2 then allows $\mathbf{T}$ to be expressed in terms of $\mathbf{M}$ as

$$\mathbf{T} = (\mathbf{M}^{-1} + \gamma\mathbf{I})^{-1}. \tag{5.1}$$

Consequently, the unique one-step transitions can be inferred from a TD-learned SR. This operation requires the SR to be always non-singular throughout the learning process, which is theoretically unlikely but practically achievable. Specifically, SR is learned using small batches of data as opposed to being updated upon every new observation (much alike batched gradient descent versus stochastic gradient descent, except there is no gradient or back-propagation in our case). This kind of "lazy update" in practice reduces large oscillations in the intermediate SR, which also helps to stabilize the event representation

and improve generalization across observations.

The same procedure is applicable to arbitrary stimulus representations through SF. Let $\boldsymbol{\phi}_t \in \mathbb{R}^d$ denote the feature representation of the stimulus experienced at time t, and $\boldsymbol{\Phi}$ the feature matrix where each row corresponds to the feature representation of a state. The successor features $\boldsymbol{\Psi}$ is

$$\boldsymbol{\Psi} = (\mathbf{I} - \gamma \mathbf{P})^{-1} \mathbf{P}.$$

$\mathbf{P}$ can be expressed in terms of $\boldsymbol{\Psi}$ as

$$\mathbf{P} = (\boldsymbol{\Psi}^{-1} + \gamma \mathbf{I})^{-1}. \tag{5.2}$$

Given the current stimulus $\boldsymbol{\phi}_t$, the next observation can be estimated as a Gaussian random vector centered at $\boldsymbol{\phi}_t' \mathbf{P}$ (the variance is inferred from experience). Because $\boldsymbol{\phi}$ may encode complex features of an observation (state), such as its latent representation, $\mathbf{P}$ defined using successor features corresponds to a generalized transition function over the *feature* states; if the feature states are equivalent to the state identity, i.e., $\boldsymbol{\Phi} = \mathbf{I}$, $\mathbf{P}$ is reduced to the pairwise state transitions $\mathbf{T}$ (Eq. 5.1). Importantly, the hidden dynamics of each event is assumed to be linear, since the predicted next observation is a linear transformation on the most recent observation $\hat{\boldsymbol{\phi}}_t = \boldsymbol{\phi}_{t-1}' \mathbf{P}$.

5.4.5 Organization of Episodic Representations

If an event boundary is identified at time t (note: t indexes the entire observation sequence), the current (inferred) event becomes associated with two abstract representations: the "triggering" temporal context of the event $\mathbf{c}^{\text{trig}}$, and the end temporal context of the *previous* event $\mathbf{c}^{\text{term}}$, represented by matrices $\mathbf{M}^{\text{KC}}$ and $\mathbf{M}^{\text{CK}}$ respectively. Specifically, if the model segments event k and k',

$$\mathbf{c}^{\text{trig}(k')} = (1 - \eta)\mathbf{c}_{t-1}^{(k)} + \eta \mathbf{c}_t^{(k')}, \tag{5.3}$$

where η is the context drift rate at event boundaries and is likely larger than TCM's β (Pu et al., 2022) to signal more abrupt shifts in context (Dubrow et al., 2017). The two components – the terminal context of the preceding event $\mathbf{c}^{\text{term}} = \mathbf{c}_{t-1}^{(k)}$ and the temporal context of the current event $\mathbf{c}_t^{(k')}$ – together account for a set of behavioral findings. First, the association between each event and its current context enables finer-grained search within an event by activating a specific event representation, which we hypothesize underlies the "stepping stone" phenomenon in memory scanning and simulation (Michelmann et al., 2023). On the other hand, previous studies have found strong but not complete context drift at boundaries driven by prediction error (Pu et al., 2022; Rouhani et al., 2019), which may explain reduced order memory across event boundaries (DuBrow & Davachi, 2013; Dubrow & Davachi, 2016), and is consistent with the hypothesis that pre-boundary information is activated to encode boundaries (Clewett et al., 2019).

Moreover, binding each event to the temporal context just before the event boundary complements the effect of $\mathbf{c}^{\text{trig}}$. During encoding, $\mathbf{c}_k^{\text{term}}$ of a particular event k becomes associated with possibly multiple successor events, the strength of which depends on the discount rate and frequency of a specific successor event – much like learning the item-to-context associations in basic TCM. At the end of memory retrieval within the event, $\mathbf{M}^{\text{CK}}$ specifies the distribution over successor *events* given the context, which the model samples and continues its retrieval thereafter. Thus these associations allow coarser-grained search by stepping out of the current event. Neuroimaging data has revealed extensive episodic encoding at the end of events (Ben-Yakov & Dudai, 2011; Ben-Yakov et al., 2013), particularly from lower- to higher-level structures (Baldassano et al., 2016), while reinstatement of contexts belonging to the event that just terminated has been posited to consolidate long-term memory (Sols et al., 2017) and optimize event understanding under the constraint of limited cognitive resources (Lu et al., 2022).

Intuitively, event-context associations $\mathbf{M}^{\text{KC}}$ are analogous to the item-to-context associations $\mathbf{M}^{\text{SC}}$ in TCM, while context-event associations $\mathbf{M}^{\text{CK}}$ are analogous to the context-to-item associations $\mathbf{M}^{\text{CS}}$, except that both represents a larger timescale and are more temporally abstracted. By chaining $\mathbf{M}^{\text{KC}}$, $\mathbf{M}^{\text{CS}}$, $\mathbf{M}^{\text{SC}}$, and $\mathbf{M}^{\text{CK}}$ together, the agent

can flexibly navigate its acquired event knowledge by alternating between within-event and across-event representations. In the next section, we further discuss how this nested hierarchy of representations accounts for hierarchical control.

5.4.6 Model-Based Control by Episodic Retrieval

Within each event/option, our proposed model performs model-based evaluation similar to TCM-SR (C. Y. Zhou et al., in press). In particular, TCM-SR hypothesizes that episodic retrieval in humans implements a Monte Carlo sampling mechanism to inform decisions on sequential tasks. At decision time, TCM retrieval employs an evolving temporal context just like in Eq. 1.1 with $\rho = \rho_{rec} \in [0, 1], \beta = \beta_{rec} = 1 - \rho_{rec}$, while recursively drawing one-hot memory samples according to Eq. 3.7.

Notably, because learning $\mathbf{M}^{\mathrm{CS}}$ is equivalent to learning an SR over the experience, while SR is reward agnostic and temporally abstracted, the retrieved samples can be flexibly combined to compute unbiased estimates of state and/or state-action values based on decision-relevant reward information over multiple timesteps (C. Y. Zhou et al., in press):

$$V_\gamma(\mathbf{s}) = \begin{cases} \frac{\beta_{\mathrm{rec}}}{\beta_{\mathrm{enc}}} \mathbb{E}\left[\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{r}\right] & \text{if } \beta_{\mathrm{rec}} > 0 \\ \frac{1}{N\beta_{\mathrm{enc}}} \mathbb{E}\left[\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{r}\right] & \text{if } \beta_{\mathrm{rec}} = 0. \end{cases} \tag{5.4}$$

Here, N denotes the total number of samples and $\mathbf{r}$ is the reward function over states (observations). More precisely, when $\beta_{\mathrm{rec}} = 0$, TCM-SR gives unbiased value estimates with a temporal discount equal to ρ_{enc}, which specifies the model's encoding timescale, or no discount if $\beta_{\mathrm{rec}} > 0$. Value estimates of different state or state-action pairs subsequently inform choices of action.

In fact, TCM-SR comprises a spectrum of decision-by-sampling algorithms corresponding to different retrieval dynamics, including classic i.i.d. sampling ($\rho_{\mathrm{rec}} = 1, \beta_{\mathrm{rec}} = 0$), generalized rollouts ($\rho_{\mathrm{rec}} = 0, \beta_{\mathrm{rec}} = 1$), and intermediate regimes that are similar to the vine sampling method (Schulman et al., 2015). TCM-SR also formalizes the effect of limited experiences and emotional modulated memory on evaluation, which we do not discuss further here.

As the first major extension, we have introduced arbitrary representations into TCM-SR. Consequently, TCM learns a successor *feature* representation by encoding real-valued latent features of experience as opposed to the identity of each experience. Retrieval still follows the form of Eq. 3.7, namely

$$\boldsymbol{\phi}_i \sim \frac{1}{Z}(D_{\cos}(\mathbf{M}^{\mathrm{CS}}, \mathbf{c}_i) + \zeta), \tag{5.5}$$

where

$$\mathbf{c}_i = \rho_{\mathrm{rec}}\mathbf{c}_{i-1} + \beta_{\mathrm{rec}}\mathbf{M}^{\mathrm{SC}}\boldsymbol{\phi}_{i-1} \tag{5.6}$$

and

$$\zeta = \min(D_{\cos}(\mathbf{M}^{\mathrm{CS}}, \mathbf{c}_i))$$

This generalized sampling applies (1) cosine distance $D_{\cos} : \mathbb{R}^{d\times d} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ to compute pairwise distance between each row in $\mathbf{M}^{\mathrm{CS}}$ and the current temporal context $\mathbf{c}_i$, and (2) an offset to ensure non-negative sampling probabilities for even stimuli with a negative distance to the context (e.g., pointing at opposite directions in the semantic space). Note if $\mathbf{M}^{\mathrm{CS}}$ is the SR, pairwise cosine distances are equal to the product $\mathbf{M}^{\mathrm{CS}}\mathbf{c}_i$, which by definition has only non-negative elements. In that case, Eq. 5.5 is reduced to Eq. 3.7, so this is a strict generalization. Moreover, we define the sampling distribution over cosine distances for two reasons: first, cosine distance is widely used to predict similarity judgment (e.g., Deyne et al., 2018; Michelmann et al., 2023; see Richie and Bhatia, 2020 for a review); second, it is insensitive to the magnitude of vectors. The latter may appear disadvantageous, but is in fact desirable in our case. During encoding, observations with large magnitudes will disproportionally update event representations, which may distort the inferred event dynamics and cause excessive segmentation. These observations may also be sampled more often (e.g., if similarity is defined by Euclidean distance) and bias retrieval in unwanted ways.

To evaluate a state $\mathbf{s}$ with respect to an event k with arbitrary representations, given the reward function is a linear function of the latent features, i.e., $\mathbf{r}^{(k)} = \boldsymbol{\phi}_i' \mathbf{w}^{(k)}$, we have

$$V_\gamma^{(k)}(\mathbf{s}) = \begin{cases} \frac{\beta_{\text{rec}}}{\beta_{\text{enc}}} \mathbb{E}\left[\sum_{i=1}^N \boldsymbol{\phi}_i' \mathbf{w}^{(k)}\right] & \text{if } \beta_{\text{rec}} > 0 \\ \frac{1}{N\beta_{\text{enc}}} \mathbb{E}\left[\sum_{i=1}^N \boldsymbol{\phi}_i' \mathbf{w}^{(k)}\right] & \text{if } \beta_{\text{rec}} = 0. \end{cases} \tag{5.7}$$

Extending TCM-SR to a hierarchical architecture also enriches the original hypothesis of episodic control. In C. Y. Zhou et al. (in press), although the task involves sequential dependencies, only a single decision needs to be made. TCM-SR essentially solves a local policy improvement problem, but does not address decision making when choices depend on each other, as most sequential decision tasks entail. We explore one possible solution here, which is to segment the vast state space of a sequential task into options ("events") and greedily compose them together to solve tasks with similar sub-problems. Specifically, assume event representations $\mathbf{M}^{\text{CS}(k)} = \boldsymbol{\Psi}_k^{(k)}$ have been learned for events $k \in \{1, 2, \ldots, NK\}$ such that they partition the entire observation sequence, and the model is cued with an initial observation $\mathbf{s} = \boldsymbol{\phi}_0$ that belongs to event $\mathbf{k}_j$ ($j = 0$). The model first retrieves a context associated with $\mathbf{k}_j$ by

$$\mathbf{c}_0^{(k_j)} = \mathbf{M}^{\text{KC}} \mathbf{k}_j, \tag{5.8}$$

This operation provides the first input context to Eq. 5.5. Retrieval then proceeds as in TCM-SR, where the model draws N samples while evolving its temporal context according to Eq. 5.6. At the end of retrieval, the model may sample a successor event according to

$$\mathbf{k}_{j+1} \sim \frac{1}{Z_k}(D_{\cos}(\mathbf{M}^{\text{CK}}, \mathbf{c}_N^{(k_j)}) + \zeta_k), \tag{5.9}$$

, with

$$\zeta_k = \min(D_{\cos}(\mathbf{M}^{\text{CK}}, \mathbf{c}_N^{(k_j)}))$$

and Z_k is the normalization constant. Alternatively, if the distribution of successor events

is different from the observations but nonetheless available, $\mathbf{k}_{j+1}$ can be obtained from it as well. In either case, the sampled event is provided as the input to Eq. 5.8. This whole procedure repeats until a target experience is recalled (e.g., in memory scanning), or the model runs out of time.

5.4.7 Simulation Details

Simulation 0: Switching Plinko

Switching Plinko is a temporally extended game based on the task of Plinko (C. Y. Zhou et al., in press; see Section 3.3.1) to formalize action evaluation via episodic memory samples in sequential decision problems. In a trial of Plinko, the agent chooses a place on the top row of the game board to drop a ball. At each time step t, the ball falls diagonally either to the left or to the right by one row with equal probability unless the ball is at a wall, in which case it falls to the opposite side with probability of 1 to stay within the board boundary. A trajectory is defined as the sequence of ball locations starting from the top row and ending at the bottom of the board, with length equal to the number of rows on the board. Rewards are scattered across the board and obtained if hit by the falling ball. The agent's objective is to maximize the total reward.

Plinko is formally a Markov Reward Process (MRP) problem, which differs from a Markov Decision Process (MDP) as the transition dynamics are not under control of the agent (equivalent to using a fixed policy in an MDP). To align with existing decision making literature, we hereby formalize the task in the language of MDP: a Plinko game is defined by a 5-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ denotes the set of states, $\mathcal{A}$ denotes the set of actions corresponding to each possible ball placement on the top row, $\mathcal{P} : \mathcal{S} \mapsto \mathcal{S}$ is the Markov transition function that defines the probability distribution $\mathcal{P}(s'|s)$ of transitioning from state s to state s'. $\mathcal{R} : \mathcal{S} \mapsto \mathbb{R}$ is the reward function $\mathcal{R}(s)$ specifying the reward magnitude received upon visiting state s, and $\gamma \in [0, 1)$ is the discount factor that controls the temporal horizon of computations by reducing the importance of rewards in distant future. By construction, $\mathcal{P}(s'|s)$ is equal to 0.5 if s is not next to a wall and s' is diagonally adjacent to s below, 1 if s is next to a wall and s' is diagonally adjacent to s below, and 0 otherwise. The goal of the agent is to choose the

action that maximizes the cumulative discounted return $G = \sum_{t=1}^{\infty} \gamma^t \mathcal{R}(S_t)$, where S_t is a random variable denoting the state at time t. We additionally assume that the ball enters an unrewarded absorbing state once it reaches the bottom of the board.

To select an action among $\mathcal{A}$, the agent estimates the action value $q(a)$ for each candidate $a \in \mathcal{A}$. Specifically, since each a is deterministically related to a unique state s_a, $q(a)$ can be defined in terms of states as $q(a) = v(s_a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t \mathcal{R}(S_t) \mid S_1 = s_a]$. $q(a)$ fulfills the function of a decision variable to inform subsequent choice. Various reinforcement learning methods can be applied to estimate $q(a)$. In particular, if $\mathcal{P}$ and $\mathcal{R}$ are known, a model-based agent can perform rollout to generate a plausible trajectory $(S_1, R_1, S_2, R_2, S_3, R_3, \ldots, S_T, R_T)$ such that $S_1 = s_a$, $S_{i+1} \sim \mathcal{P}(\dot{|}S_i)$ for $i > 1$, and $R_i = \mathcal{R}(S_i)$ (Tesauro & Galperin, 1996). The total discounted reward along a rollout trajectory is a Monte Carlo estimate of the action value, i.e., $q(a) = \mathbb{E}[\sum_{t=1}^{\infty} \gamma^i R_i]$. This approach is shown to be compatible with TCM, by continuously drawing episodic memory samples from the context-item association matrix in TCM with $\rho_{rec} = 0, \beta_{\text{rec}} = 1$ (C. Y. Zhou et al., in press). If γ is large, the model could skip multiple steps at a time (i.e., generalized rollout) and compute action values over an extended temporal horizon, and we adopt this decision-by-sampling scheme in the current simulation.

In Switching Plinko, the transition function is additionally a function of the trial. Denote the trial number as m. Then $\mathcal{P}(s'|s, m)$ is the transition probability from state s to state s' *given the trial context* - that is, the transition function is time-varying, as well as the resultant MDP. Note that the transition probabilities never change during a trial, but may change between two trials. We simulated a Switching Plinko task with three modes or sets of possible transition probabilities: (1) unbiased (U): $\mathcal{P}(s'|s, m) = \mathcal{P}_U(s'|s) = 0.5$ if s' is diagonally adjacent to s below; (2) left-biased (L): $\mathcal{P}(s'|s, m) = \mathcal{P}_L(s'|s) = 0.8$ if s' is diagonally down to the left of s and $\mathcal{P}(s'|s, m) = 0.2$ if s' is diagonally down to the right of s; (3) right-biased (R): $\mathcal{P}(s'|s, m) = \mathcal{P}_R(s'|s) = 0.2$ if s' is diagonally down to the left of s and $\mathcal{P}(s'|s, m) = 0.8$ if s' is diagonally down to the right of s. The other four components – $\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma$ – remain time-invariant throughout trials. The exact $\mathcal{P}$ is random but remain fixed within a block of 25 trials. For instance, Fig. 5.2b shows a

sequence of observations consisting of 25 right-biased trials/games, 50 unbiased games, and so on. Importantly, the change points are unsignaled, and the agent has no knowledge about either the possible transition dynamics or the cardinality of this set.

Observations consisted of 225 random trajectories by dropping a ball at the top-middle board location, each consisting of 10 ordered states as the board contained 10 rows and 9 columns. Observations were provided as input to the model in one pass. Five rewards of size 10 were placed as shown in Fig. 5.2a. During test, one observation from a biased board was provided, and model needed to pick among the three actions shown in Fig. 5.2a by assuming the test game had the same transition dynamics as the last observed trajectory. The hierarchical model was parameterized as follows:

Table 5.1: Model parameters for Simulation 0

Description	**Parameter**	**Value**
error threshold	pe_{thres}	2.0
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(=\gamma)$	0.9
within-event context drift rate at retrieval	β_{rec}	1
context drift rate at event boundaries	η	0.9
eligibility trace	λ	1
new event spawn rate	α_{sCRP}	10
history weight	κ_{sCRP}	1
stickiness	λ_{sCRP}	.1

Event dynamics were updated after each observation. The learning rate of event representations exponentially decays as a function of time:

$$\alpha_{t+1} = \alpha_0 \times \exp(-rt/\text{step}),$$

where decay rate $r = 0.01$ and step size step $= 10$.

The baseline model without event structure does not have any of the event-related parameters but otherwise shares the same parameterization (e.g., $\alpha_0, \beta_{\text{enc}}, \beta_{\text{rec}}, \lambda$) and

learning schedule. Each action value $q(a)$ is computed as the average across 10000 generalized rollouts for both models.

Simulation 1: Event Segmentation on Naturalistic Stimuli

The material consisted of the full "Washing Dishes" video clip from Zacks et al. (2006). Observations were generated by a variational autoencoder such that each frame was embedded in a 100-dimensional space with real values. Observations were passed through the model once. Each model instance was parameterized as follows:

Table 5.2: Model parameters for Simulation 1

Description	**Parameter**	**Value**
error threshold	pe_{thres}	0.8
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(=\gamma)$	0.4
context drift rate at event boundaries	η	0.95
eligibility trace	λ	1
new event spawn rate	α_{sCRP}	0.1
history weight	κ_{sCRP}	1
stickiness	λ_{sCRP}	100

We trained 25 model instances on the data. Event dynamics were inferred in batches of size 30 (i.e., every 30 frames). The learning rate of event representations exponentially decayed as a function of time with decay rate $r = 0.01$ and step size step $= 10$.

Model and subject boundaries were grouped into 1-second bins, as was reported in Zacks et al. (2006). The scaled point biserial correlation was used to quantify the agreement between discrete model boundaries and averaged human marked boundaries, because it accounts for the number of boundaries inferred by the model (Bezdek et al., 2022; Lu et al., 2023). To compute the scaled version, we first computed the regular (unscaled) point biserial correlation as

$$r_{pb} = \frac{M_1 - M_0}{s}\sqrt{\frac{n_1 n_0}{n^2}},$$

where M_1 is the average human segmentation frequency at a model-marked boundary, M_0 is the average human segmentation frequency when a model did not indicate an event boundary, n_1 is the number of model boundaries, and $n_0 = n - n_1$ is the number of time points without any boundary according to the model. Lastly, s is the standard deviation of human segmentation frequency across the entire length of the video clip. The scaled point biserial correlation is obtained by scaling r_{pb} using the maximum and minimum possible r_{pb} given the total number of model-marked boundaries.

Simulation 2: Event Segmentation on Community Structure

The stimuli ($N = 15$) were simulated as random feature vectors of size 10. Specifically, each vector was sampled as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Following the experiment procedure of the original study (Experiment 2 of Schapiro et al., 2013), the model was trained on a sequence of 1400 observations for one pass and tested on a separate set of 600. Training data was generated from a random walk on the community structure (Fig. 5.4a) starting from a random node. Testing data was generated by interleaving random walks of length 30 and Hamiltonian walks of length 15. The construction of observation sequences followed the exact setup as Schapiro et al. (2013). Each model instance was parameterized as follows:

Table 5.3: Model parameters for Simulation 2

Description	**Parameter**	**Value**
error threshold	pe_{thres}	0.4
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(= \gamma)$	0.6
context drift rate at event boundaries	η	1.0
eligibility trace	λ	1
new event spawn rate	α_{sCRP}	1
history weight	κ_{sCRP}	1
stickiness	λ_{sCRP}	1

We trained 50 model instances on the data. Event dynamics were inferred at every encoding time step. All representations were frozen after the training stage so no further

learning was possible during the test phase. The learning rate of event representations exponentially decayed as a function of time with decay rate $r = 0.01$ and step size step $= 1$.

Simulation 3: Event Representation and Hierarchical Episodic Retrieval

The stimuli (288 distinct word tokens) were simulated as random feature vectors of size 10. Specifically, each vector was sampled as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As in the original study (Experiment 2 of Pettijohn et al., 2016), each experiment consisted of 12 trials with 2 lists of 12 words each. The model was trained on each trial (24 random vectors) for one pass and then performed free recall for that trial. An additional sequence was provided as input to the model annotating the event that the current stimulus belongs to. Depending on the condition, the model either took the event annotation into account (shift condition), or ignored it (no-shift condition). Each model instance was parameterized as follows:

Table 5.4: Model parameters for Simulation 3

Description	**Parameter**	**Value**
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(= \gamma)$	0.3
within-event context drift rate at retrieval	β_{rec}	0.77
context drift rate at event boundaries	η	0.99
eligibility trace	λ	1

Considering the task mainly involved stochasticity in retrieval but not encoding, as the model (as well as participants in the original study) was provided with the ground truth event label, we trained one model instance for each combination of trial and condition and performed 200 free recall experiments with each model. Free recall was initiated with a random stimulus from the trial, simulating the first recall (note we did not impose any sequential order effect of human free recall on the model, which usually includes primacy and recency effects). Event dynamics were computed at the end of each list (i.e., twice in a single trial). The learning rate of event representations exponentially decayed as a function of time with decay rate $r = 0.01$ and step size step $= 1$.

Simulation 4: Event Representation and Memory Organization

Stimuli ($N = 1000$) were simulated as random feature vectors of size 10. Specifically, each vector was sampled as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Consistent with the original study (Experiment 1 of DuBrow and Davachi, 2013), each experiment trial consisted of 5 lists with 5 distinct stimuli each. There were two types of events (face and object in the original study), and the event switched after each list presentation. The model was trained on each trial (24 random vectors) for one pass and then performed free recall for that trial. An additional sequence was provided as input to the model annotating the event that the current stimulus belongs to. Depending on the condition, the model either took the event annotation into account (shift condition), or ignored it (no-shift condition). Each model instance was parameterized as follows:

Table 5.5: Model parameters for Simulation 4

Description	**Parameter**	**Value**
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(= \gamma)$	0.3
within-event context drift rate at retrieval	β_{rec}	0.77
context drift rate at event boundaries	η	0.9
eligibility trace	λ	1

Similar to Simulation 3, the task mainly involved stochasticity in retrieval but not encoding, as the model (as well as participants in the original study) was provided with the ground truth event label. We therefore trained one model instance for each combination of trial and condition and performed 100 free recall experiments with each model. Free recall was initiated with a random stimulus from the trial, simulating the first recall (note we did not impose any sequential order effect of human free recall on the model, which usually includes primacy and recency effects). Event dynamics were computed at the end of each list (i.e., five times in a single trial). The learning rate of event representations exponentially decayed as a function of time with decay rate $r = 0.01$ and step size step $= 1$.

Simulation 5: Memory Search of Temporally Extended Events

The material consisted of the first clip from the movie *Gravity* used by Michelmann et al. (2023). Observations were generated by a variational autoencoder such that each frame was embedded in a 100-dimensional space with real values. We further limited the L2 norm of the embeddings since some of them were more than ten times larger than the median and caused large perturbations in the inferred event representation. Specifically, the top 5% outlier were scaled down to the 95-percentile L2 norm. The observations were passed through the model once. Each model instance was parameterized as follows:

Table 5.6: Model parameters for Simulation 5

Description	**Parameter**	**Value**
error threshold	pe_{thres}	2.0
initial learning rate	α_0	0.01
within-event context drift rate at encoding	$\beta_{\text{enc}}(=\gamma)$	0.4
within-event context drift rate at retrieval	β	0.77
context drift rate at event boundaries	η	0.9
eligibility trace	λ	1
new event spawn rate	α_{sCRP}	0.1
history weight	κ_{sCRP}	1
stickiness	λ_{sCRP}	10

We trained 5 model instances on the data. Event dynamics were inferred in batches of size 30 (i.e., every 30 frames). The learning rate of event representations exponentially decayed as a function of time with decay rate $r = 0.01$ and step size step $= 30$.

The memory scanning task consisted of nine pairs of start and target scenes, indexed by their frame number: (0, 200), (0, 1100), (0, 4000), (3000, 3200), (3000, 4500), (3000, 6800), (6000, 6500), (6000, 7000), (6000, 10000). These pairs were picked so that they span different points in the movie (a total of 10796 frames) with varying duration and number of events between the each pair. Each model performed 20 memory search trials for each start-target scene pair, resulting in a total of 900 trials (5 models $\times$ 9 pairs $\times$ 20 trials)

Similar to Michelmann et al. (2023), the model adopted a skipping threshold to decide when to terminate search in the current event. Specifically, it kept track of the total cosine distance between each recalled scene and the target scene as a measure of dissimilarity. If the accumulated dissimilarity exceeded the skipping threshold, search within the current event was terminated, and the model sampled a new event to continue memory scanning. A skipping threshold of 60 was used and was reset at the beginning of every new event.

The model successfully "found" the target scene $\mathbf{x}^*$ and terminated the search altogether if it recalled an observation $\tilde{\mathbf{x}}$ such that

$$d_{\cos}(\tilde{\mathbf{x}}, \mathbf{x}^*) \leq 0.5.$$

The match criterion was determined by an exploratory analysis on the distribution of cosine distances between an arbitrary scene in the movie and its neighbors (either preceding or following) versus random scenes from the film (Fig. 5.8).

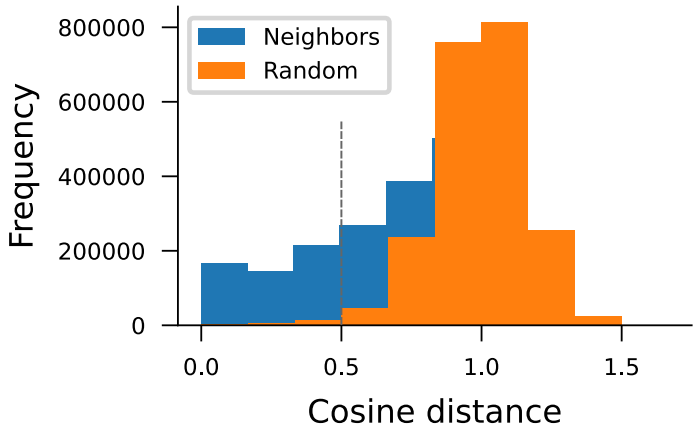

Figure 5.8: **Distribution of cosine distances between pairs of observations in the movie used in Simulation 5.** Cosine distances were computed between each embedded frame and its predecessor/successor frames (blue, 100 in each direction, corresponding to roughly 4.2 seconds before and after the frame), as well as 200 random scenes (orange). The grey dashed line indicates the cutoff value of 0.5 to minimize false positives in memory search.

Acknowledgement

This chapter, in full, is currently being prepared for submission for publication of the material. Zhou, Corey Y.; Mattar, Marcelo G. The dissertation author was the primary author of this material.

Chapter 6

Conclusion

This dissertation sets out to address the guiding question "**What is memory for?**" By positing the theoretical grounds and exploring empirical evidence for sequential decisions from episodic sampling, we attempt to break open the blackbox of cognitive processes behind adaptive behavior in humans and postulate the purpose of memory beyond mere remembering. This work is motivated by the need for mechanistic explanations that are psychologically plausible from both memory and decision making perspectives.

The three chapters build on each other: Chapter 3 lays the basic computational framework that grounds model-based evaluation in episodic recall. It extends a phenomenological model of episodic memory, TCM, into a family of sample-based algorithms to solve sequential decision tasks. The novel TCM-SR framework suggests that seemingly arbitrary features of episodic memory – including serial order effects, the contiguity effect, and emotional modulation – serve an adaptive purpose. Moreover, it makes several empirical predictions about how episodic retrieval dynamics bias memory-informed sequential decisions.

Chapter 4 experimentally tests the the TCM-SR theory using two novel paradigms that involve temporal dependency over multiple time steps, unlike previous one-step bandit task studies. The findings show that participants' behavior aligns with an episodic sampling account, such as the one predicted by TCM-SR in Chapter 3. In particular, we found decision biases analogous to memory biases observed in free recall tasks, such as primacy, recency, and temporal contiguity effects.

Finally, Chapter 5 extends the TCM-SR framework in Chapter 3 to capture adaptive behavior beyond simple experiences across a few episodes. The hierarchical model introduced in this chapter hypothesizes that structure knowledge is represented by layers

of successor features from unsupervised episodic encoding. During evaluation/decision-making, the TCM dynamics guides sampling within and across events, empowering flexible and generalizable control. Critically, this chapter unifies empirical literature across episodic memory, event cognition, and adaptive control. The symbolic approach adopted by both Chapter 3 and Chapter 5 further opens up the opportunity to connect with existing models (especially TCM-based ones) and improves the interpretability of individual components.

6.1 Limitations & Future Directions

6.1.1 Memory

The work presented here lays the groundwork for future scientific inquiry at the intersection of memory and adaptive control. As the first formal attempt to bridge the psychologically-informed function of memory with the mechanism of sequential decision-making, it recontextualizes various curious properties of episodic memory. In particular, Chapter 3 explores the role of episodic retrieval dynamics in temporally extended model-based evaluation, supported by preliminary evidence in Chapter 4. Chapter 5 further examines the interactions between event structures and episodic memory, highlighting the shared basis of continual learning and phenomena such as event segmentation patterns, enhanced within-event memory, reduced memory at event boundaries, and the "stepping stone" memory search pattern.

Nonetheless, by no means does this dissertation reflect the full scope of current memory literature. First, it does not discuss how the magnitude of rewards affects memory, despite evidence that only moderate rewards enhance episodic encoding. Emotional modulation of episodic encoding, modeled in Section 3.1.8, only captures the main effect of emotion but leaves out many nuances. For instance, positive and negative reinforcers have different effects on memory encoding and retrieval (Bowen et al., 2017; Madan et al., 2019, 2020; Williams et al., 2022), and they are sensitive to the specific context (Madan et al., 2020; Schmidt et al., 2011). While some studies show an enhanced associative memory (Madan et al., 2020; Rimmele et al., 2012), a few others show a disruptive effect (Bisby et al., 2016), and sometimes emotion does not affect associative memory at all (Sharot & Yonelinas, 2008). Given that the field of memory research has not reached a

consensus, it would be premature for TCM-SR to take a definitive stance.

Fortunately, grounding the core theory in a model like TCM allows for future expansions to incorporate new findings As we have discussed in Section 3.2.5, TCM-SR as well as its hierarchical extension in Chapter 5 can be readily extended to capture more nuanced effects of reward, episodic replay, and semantic information. The TCM components in the current theory are consistent across all models derived from TCM, even in certain neural network architectures trained to perform free recall (Salvatore & Zhang, 2024).

Another overlooked aspect of the current work is the distinction between pre-experimental and experimental contexts. The retrieved context models, starting from CMR, all assume the encoded stimulus-specific context as a combination of both contexts (Polyn et al., 2009a). Formally, at each time step during encoding, $\mathbf{M}^{\mathrm{SC}}$ is updated according to

$$\mathbf{M}^{\mathrm{SC}} = (1 - \gamma_{\mathrm{FC}})\mathbf{M}^{\mathrm{SC}}_{\mathrm{pre}} + \gamma_{\mathrm{FC}}\Delta\mathbf{M}^{\mathrm{SC}}_{\mathrm{exp}}$$

Setting $\gamma_{\mathrm{FC}} = 0$ therefore results in $\mathbf{M}^{\mathrm{SC}} = \mathbf{M}^{\mathrm{SC}}_{\mathrm{pre}} = \mathbf{I}$ throughout encoding and retrieval. This is what we assumed for all simulations in Chapter 3 except the last one. Since $\mathbf{c}^{\mathrm{IN}}_i = \mathbf{M}^{\mathrm{SC}}\mathbf{x}_i$ and $\mathbf{M}^{\mathrm{SC}} = \mathbf{I}$, the simplified expression $\mathbf{c}^{\mathrm{IN}}_i = \mathbf{x}_i$ was used in most of the text, suggesting no update to $\mathbf{M}^{\mathrm{SC}}$ during encoding. However, the last simulation where learned associations affect retrieval implies that $\mathbf{M}^{\mathrm{SC}}$ should reflect the task-dependent representation $\mathbf{M}^{\mathrm{SC}}_{\mathrm{exp}}$ at each encoding step (i.e., $\gamma_{\mathrm{FC}} > 0$ throughout encoding), not just at retrieval. Yet Gershman et al. (2012)'s result only applies if $\mathbf{M}^{\mathrm{SC}} = \mathbf{I}$ (i.e., $\gamma_{\mathrm{FC}} = 0$ throughout encoding).

To examine the consequence of updating $\mathbf{M}^{\mathrm{SC}}$ with the task-dependent $\mathbf{M}^{\mathrm{SC}}_{\mathrm{exp}}$ to some extent (instead of caching it until retrieval), we simulated encoding when $\gamma_{\mathrm{FC}} = 1$ and the degree of temporal drift is moderate ($\beta_{\mathrm{enc}} \approx 0.6$). Intermediate representations during the early training (encoding) phase show an interesting pattern, where distant futures are more likely to be retrieved than immediate successors Fig. 6.1. This is most likely due to the incorporation of *predecessors* from $\mathbf{M}^{\mathrm{SC}}_{\mathrm{exp}}$ encoded during previous trials.

As the agent gains more experience, $\mathbf{M}^{\mathrm{CF}}$ eventually captures both the successors and predecessors of a stimulus Fig. 6.1. In the case of repeated exposure (dropping the ball from the same board location over and over), the predecessors happen to correspond to the positions near the end of each trajectory. This approach may lead to underrepresentation of intermediate states, suggesting a need to investigate the computational properties and theoretical guarantees of such learned representations, with potential ties to existing RL algorithms.

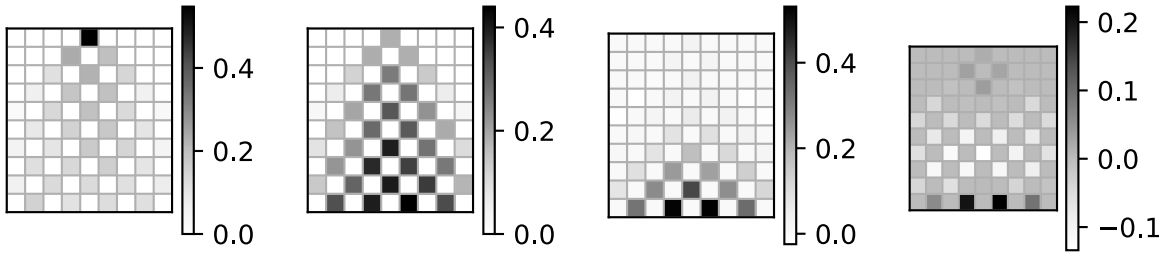

Figure 6.1: **Learned representation ($\mathbf{M}^{\mathbf{CS}}$) from updated stimulus-context association $\mathbf{M}^{\mathbf{SC}}$.** Left to right: encoded $\mathbf{M}^{\mathrm{SC}}$ of the top center state after 1%, 5%, 10%, and 100% (converged) training.

6.1.2 Control

A major simplification of this dissertation is made regarding the concept of control. In conventional RL as well as in reality, control entails choosing an action at every step. However, in Chapter 3, model-based "control" only invovles one initial action, with no further control over subsequent episodes. This approach simplifies sequential decision-making and leaves multi-step control problems for future exploration. As one potential solution, Chapter 5 suggests that a single sequence of experiences may be segmented based on latent factors and organized into decision-relevant (sub)task representations, such that the agent can apply a greedy divide-and-conquer decision strategy by activating the multiple subtask representations during decision time. Hierarchical TCM-SR thus posits that effective model-based control in humans is contingent on the statistical regularities within each (sub)task, as well as discernable dissimilarities across tasks. While the model's predictions (e.g., Section 5.2.1) should be empirically tested, prior work on episodic memory and structure learning supports this approach (Beukers et al., 2024; Giallanza et al., 2024; Lu et al., 2023). Moving beyond one-step problems, future work is also needed

to explore new paradigms with nontrivial temporal dependencies and varying degrees of within- and/or across-event similarities in order to more directly examine the effect of hierarchical episodic mechanisms on adaptive behavior.

One common issue with SR and SF is policy dependency, which our models (especially TCM-SR) are not immune to. SR and SF are learned based on the agent's own experience, so they only reflect the behavioral policy then; if environmental dynamics deviate significantly when control is required, performance may suffer. Therefore, just like any other RL algorithms, there is an inherent exploitation-exploration tradeoff associated with the use of SR. Nonetheless, the goal of this dissertation is to reverse engineer adaptive decision-making using episodic memory, not to propose a novel algorithm to balance flexibility and speed. Despite limitations, TCM-SR aligns with behavioral data, showing that human choices are consistent with the SR (Momennejad & Howard, 2018), that the hippocampus is involved in model-based and SR behavior (Stachenfeld et al., 2017), and that model-based computations become more accurate the more time is devoted to a decision (Keramati et al., 2011).

With the caveats above, the TCM-SR framework and its hierarchical extension offers substantial advantages over standard SR-based models. It enables independent control over the timescale of individual representations (the discount factor cached within the SR/SF) and the overall decision horizon, by chaining prediction steps together in MB-like rollouts of tunable depth to provide more flexibility in temporal abstraction. This feature has been acknowledged as important in previous SR formulations (Momennejad et al., 2018) but often lacked detail in previous models. The hierarchical TCM-SR further mitigates policy dependency by caching multiple SRs/SFs for different event contexts (using event segmentation in conjunction with temporal abstraction) and storing higher-level relationships. This approach breaks decision problems down into more manageable and practical units, such that (sub)policies can be flexibly composed together like options (Sutton et al., 1999). This enables specialized solutions tailored to subtask dynamics (e.g., by adjusting the agent's sensitivity to prediction errors and the degree of generalization) rather than a blanket policy. It also counteracts the inflexibility of SR with modularity –

for instance, it allows partial evaluation and improvement to adjust for relatively local changes without the need of global re-learning, which could explain the adaptiveness observed in humans.

One limitation of the current approach to disentangle prediction timescales is its relative crudeness in disentangling prediction timescales. In the current work, we choose to use p_{stop} as opposed to more elaborate mechanisms such as a diffusion process, which prior models including TCM-A adopted to model specific aspects of episodic recall, such as inter-response delays (Murdock & Okada, 1970), individual self-motivation (Dougherty & Harbison, 2007), and recall instructions (Osth et al., 2021). While the use of a stopping probability allowed for formal proof of the value estimator's unbiasedness, future work could incorporate more detailed process-level mechanisms, such as those in TCM-A, or explore other retrieval strategies to enhance the model's precision in decision-making contexts (Badre et al., 2014; Naim et al., 2020).

6.1.3 Behavior

The rich mechanistic details of TCM-SR suggests numerous avenues for empirical research, with Chapter 4 serving as a starting point towards a systematic verification of the theory. Specifically, Section 4.1 found biases in value-based evaluation to be analogous to serial position biases in episodic retrieval, while Section 4.2 showed that manipulated recall patterns better explain people's choice than habit-like MF strategies or episodic memory as a veridical storage of experience. Still, a gap remains between these two experiments: it is unclear how the serial position of an item affects its weight in the final *decision*. Although our results illustrated the effect of serial position on *evaluation* and trial-wise prediction accuracy, no strong statistical differences were found by regressing decisions on item values as a function of serial position (e.g., how Fig. 4.1c and Fig. 4.1d mirror each other). We suspect that the paradigm used in Section 4.2 may have been too difficult, as subjects' choices were highly noisy. Future work should iterate the gridworld design for a more precise characterization of episodic memory effects in sequential decision making.

The next step could be fitting TCM-SR to individual subjects. Unlike neural

network models, TCM-SR and its elements are interpretable. For instance, one could relate the temporal drift rate to altered performance due to discontiguous encoding. TCM-SR predicts that higher drift rate leads to worse associations of discontiguous paths (since the agent is less likely to recall in the backward direction), which can be directly tested by model fitting on the subject level. Another possibility is to simulate TCM-SR on the behavioral tasks and observe the consequences of different parameter settings. The simulated results can then be compared against human performance, including those who have memory deficits, to identify behavioral aspects that are captured by the framework.

While this dissertation focuses exclusively on behavior due to TCM being a phenomenological model of episodic memory, recent research suggests involvement of neural substrates including CA1 and the dentate gyrus (DG) in maintaining and reinstating temporal contexts (Dimsdale-Zucker et al., 2022; Kragel et al., 2020; Moscovitch et al., 2016; Sakon & Kahana, 2021). Understanding the implementational details of TCM (or Retrieval Context Models in general) is crucial for theories like TCM-SR, as the interaction between memory and complex decision-making spans beyond the process level. TCM-SR theorizes the integration of episodic encoding and retrieval dynamics into adaptive behavior mechanisms. Likewise, the hippocampal formation may be extensively recruited in decision making, planning, and cognitive control in ways we have overlooked. One prominent model of the hippocampus, the Tolman-Eichenbaum Machine (TEM; Whittington et al., 2019), bears a few similarities with TCM. For instance, both TCM and TEM posits that the brain maintains an evolving representation of the "where" information (temporal context in TCM, abstract location $\mathbf{g}$ in TEM), and that episodic memory consists of associative conjunctions of "where" and "what" ($\mathbf{M}^{\mathrm{CS}}$ in TCM, $\mathbf{p}$ in TEM). Linking TCM to a neural circuit model like TEM can further unify existing EM-for-DM frameworks, offering richer explanation of behavior and brain functions.

6.2 Final Remark: Building Bridges

I have always adored works that bring seemingly irrelevant or incompatible things together. It does not really matter what the "things" are – they can be TCM and SR, as in Gershman et al. (2012), or Bayesian inference and causal learning, as in Tenenbaum

et al. (2011), or biblical tales and nuclear science, as in Dali's *Crucifixion*. To me, the most thrilling experience is discovering the slightest change in perspective that transforms the entire picture, much like anamorphic installations.

This general preference of connections over specific domains has largely shaped my taste in science, and, ultimately, my own research. My original plan before entering college was to study film cinematography or screenwriting, but O soon found the cognitive theory of visual scene understanding more interesting. While that plan eventually gave way to the allure of causal reasoning, it somehow crept back in a slightly different form as a problem of event cognition in Chapter 5. Other than the three projects included in this dissertation, my first-ever publication essentially recast the framework of Marr's three levels of analysis for the field of interpretable AI[1]. My final project in grad school linked CMR, a descendant of TCM, to the attention mechanism of Transformer-based large language models[2]. The topics and methodologies are quite eclectic (to the point I've been asked several times "What is your research *actually* about?"), and I am still excited to discover yet another connection between classic theories and modern tools to illustrate fundamental principles underlying intelligence.

I don't believe in any grand unifying theory, but I do believe in interconnectedness. By pure coincidence, the theme of my undergraduate orientation was "Building Bridges," which now seems like fitting foreshadowing. We don't dwell on bridges; in fact, almost everyone uses them as a means to an end, to get to another place where they spend much more time doing hopefully more interesting things than they would on one isolated piece of land. That's how I hope my work fits into the never-ending quest to understand the human mind: to initiate exchanges, motivate ententes, ground expansions, but never cease explorations.

---

[1]Zhou, Y., & Danks, D. (2020). Different "Intelligibility" for Different Folks. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

[2]Ji-An, L., Zhou, C. Y., Benna, M. K., & Mattar, M. G. (2024). Linking In-context Learning in Transformers to Human Episodic Memory. *arXiv, abs/2405.14992*.

Bibliography

Aggleton, J. P., Sanderson, D. J., & Pearce, J. M. (2007). Structural learning and the hippocampus. *Hippocampus*, *17*(9), 723–34. https://doi.org/10.1002/hipo.20323

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/https://doi.org/10.1037/xge0001034

Alvernhe, A., Save, E., & Poucet, B. (2011). Local remapping of place cell firing in the Tolman detour task. *European Journal of Neuroscience*, *33*(9), 1696–705. https://doi.org/10.1111/j.1460-9568.2011.07653.x

Anderson, J. R., & Fincham, J. M. (2014). Extending problem-solving procedures through reflection. *Cognitive Psychology*, *74*, 1–34.

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*(2), 82–90.

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, *63*, 173–209.

Axmacher, N., Cohen, M. X., Fell, J., Haupt, S., Dümpelmann, M., Elger, C. E., Schlaepfer, T. E., Lenartz, D., Sturm, V., & Ranganath, C. (2010). Intracranial EEG Correlates of Expectancy and Memory Formation in the Human Hippocampus and Nucleus Accumbens. *Neuron*, *65*, 541–549.

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), 1928–48. https://doi.org/10.1162/jocn_a_00596

Bailey, D., & Mattar, M. G. (2022). Predecessor Features. *The 5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making*.

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 (T. Kahnt, M. J. Frank, L. K. Fellows, & S. Gluth, Eds.). *eLife*, *8*, e46080. https://doi.org/10.7554/eLife.46080

Baldassano, C. A., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2016). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron*, *95*, 709–721.e5.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*, 1382–1407.

Baldwin, D. A., & Kosie, J. E. (2020). How Does the Mind Render Streaming Experience as Events? *Topics in Cognitive Science*, *13*, 79–105. https://doi.org/10.1111/tops.12502

Barreto, A., Borsa, D., Hou, S., Comanici, G., Aygün, E., Hamel, P., Toyama, D., Hunt, J. J., Mourad, S., Silver, D., & Precup, D. (2021). The Option Keyboard: Combining Skills in Reinforcement Learning. *Neural Information Processing Systems*.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., Silver, D., & Hasselt, H. V. (2017). Successor Features for Transfer in Reinforcement Learning. *Neural Information Processing Systems*.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bellman, R. (1957). A Markovian Decision Process. *Indiana University Mathematics Journal*, *6*, 679–684.

Bengio, Y., Bengio, S., & Cloutier, J. (1991). Learning a synaptic learning rule. *International Joint Conference on Neural Networks*, *2*, 969–974. https://doi.org/10.1109/IJCNN.1991.155621

Ben-Yakov, A., & Dudai, Y. (2011). Constructing Realistic Engrams: Poststimulus Activity of Hippocampus and Dorsal Striatum Predicts Subsequent Episodic Memory. *The Journal of Neuroscience*, *31*, 9032–9042.

Ben-Yakov, A., Eshel, N., & Dudai, Y. (2013). Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. *Journal of Experimental Psychology: General*, *142*(4), 1255–63.

Beukers, A. O., Collin, S. H. P., Kempner, R. P., Franklin, N. T., Gershman, S. J., & Norman, K. A. (2024). Blocked training facilitates learning of multiple schemas. *Communications Psychology*, *2*(28). https://doi.org/10.1038/s44271-024-00079-4

Bezdek, M. A., Nguyen, T. T., Gershman, S. J., Bobick, A. F., Braver, T. S., & Zacks, J. M. (2022). Modeling human activity comprehension at human scale: Prediction, segmentation, and categorization.

Bisby, J. A., Horner, A. J., Hørlyck, L. D., & Burgess, N. (2016). Opposing effects of negative emotion on amygdalar and hippocampal memory for items and associations. *Social Cognitive and Affective Neuroscience*, *11*, 981–990.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J. W., Wierstra, D., & Hassabis, D. (2016). Model-Free Episodic Control. *ArXiv*, *abs/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), e1003387. 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*, 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*. https://doi.org/10.1038/nn.4573

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & memory*, *11 5*, 485–94.

Bowen, H., Kark, S., & Kensinger, E. (2017). NEVER forget: Negative emotional valence enhances recapitulation. *Psychonomic Bulletin & Review*, *25*, 870–891. https://doi.org/10.3758/s13423-017-1313-9

Braun, E. K., Elliott, W. G., & Shohamy, D. (2018). Retroactive and graded prioritization of memory by reward. *Nature Communications*, *9*, 4886. https://doi.org/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), e1005003. https://doi.org/10.1371/journal.pcbi.1005003

Chang, C. Y., Esber, G. R., Marrero-Garcia, Y., Yau, H.-J., Bonci, A., & Schoenbaum, G. (2015). Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nature Neuroscience*, *19*, 111–116.

Chang, C. Y., Gardner, M. P. H., Tillio, M. G. D., & Schoenbaum, G. (2017). Optogenetic Blockade of Dopamine Transients Prevents Learning Induced by Changes in Reward Features. *Current Biology*, *27*, 3480–3486.e3.

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/https://doi.org/10.1002/hipo.23074

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

Cohen, R. T., & Kahana, M. J. (2022). A memory-based theory of emotional disorders. *Psychological Review*, *129*(4), 742–776. https://doi.org/https://doi.org/10.1037/rev0000334

Collins, A. G., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229.

Coulom, R. (2006). Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. *Computers and Games*.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2004). Similarity and Discrimination in Classical Conditioning: A Latent Variable Account. *Neural Information Processing Systems*.

Cushman, F. A., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences*, *112*, 13817–13822.

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), 20130478.

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

Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Deyne, S. D., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987–1006.

Dimsdale-Zucker, H., Montchal, M., Reagh, Z., Wang, S.-F., Libby, L., & Ranganath, C. (2022). Representations of Complex Contexts: A Role for Hippocampus. *Journal of Cognitive Neuroscience*, *35*, 1–21. https://doi.org/10.1162/jocn_a_01919

Dimsdale-Zucker, H., Ritchey, M., Ekstrom, A., Yonelinas, A., & Ranganath, C. (2018). CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications*, *9*, 294. https://doi.org/10.1038/s41467-017-02752-1

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.

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), 1108–17.

DuBrow, S., & Davachi, L. (2013). The Influence of Context Boundaries on Memory for the Sequential Order of Events. *Journal of Experimental Psychology: General*, *142*(4), 1277–1286. https://doi.org/10.1037/a0034024

Dubrow, S., & Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory*, *134*, 107–114.

Dubrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, *17*, 141–146.

Duff, M. C., Covington, N. V., Hilverman, C., & Cohen, N. J. (2019). Semantic Memory and the Hippocampus: Revisiting, Reaffirming, and Extending the Reach of Their Critical Relationship. *Frontiers in Human Neuroscience*, *13*, 471.

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

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College Press. https://doi.org/https://doi.org/10.1037/10011-000

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

Eichenbaum, H., Kuperstein, M., Fagan, A. M., Nagode, J. C., Cohen, N., Keefe, J. G., Ranck, J. B., & Winson, J. (1987). Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task. *Journal of Neuroscience*, *7*(3), 716–32. https://doi.org/10.1523/JNEUROSCI.07-03-00716.1987

Ekman, M., Kusch, S., & de Lange, F. P. (2023). Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *eLife*, *12*, e78904.

Ezzyat, Y., & Davachi, L. (2011). What Constitutes an Episode in Episodic Memory? *Psychological Science*, *22*, 243–252.

Fang, C., Aronov, D., Abbott, L. F., & Mackevicius, E. L. (2023). Neural learning rules for generating flexible predictions and computing the successor representation. *eLife*, *12*, e80680. https://doi.org/10.7554/eLife.80680

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. *International Conference on Machine Learning*.

Flesch, T., Nagy, D. G., Saxe, A., & Summerfield, C. (2023). Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLOS Computational Biology*, *19*(1), 1–32. https://doi.org/10.1371/journal.pcbi.1010808

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2009). A Sticky HDP-HMM With Application to Speaker Diarization. *The Annals of Applied Statistics*, *5*, 1020–1056.

Franklin, N., Norman, K., Ranganath, C., Zacks, J., & Gershman, S. (2020). Structured Event Memory: A Neuro-Symbolic Model of Event Cognition. *Psychological Review*, *127*(3), 327–361. https://doi.org/10.1037/rev0000177

Fukai, T., Asabuki, T., & Haga, T. (2021). Neural mechanisms for learning hierarchical structures of information. *Current Opinion in Neurobiology*, *70*, 145–153.

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*, e17086. https://doi.org/10.7554/eLife.17086

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*, 101–128.

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior*, *43*, 243–250.

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M. H., & Niv, Y. (2013). Gradual extinction prevents the return of fear: Implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, *7*.

Gershman, S. J., Moore, C. D., Todd, M. T., Norman, K. A., & Sederberg, P. B. (2012). The Successor Representation and Temporal Context. *Neural Computation*, *24*, 1553–1568.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*, 255–268.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical Computations Underlying the Dynamics of Memory Updating. *PLoS Computational Biology*, *10*(11), e1003939.

Giallanza, T., Campbell, D., & Cohen, J. D. (2024). Toward the Emergence of Intelligent Control: Episodic Generalization and Optimization. https://doi.org/10.1162/opmi_a_00143

Gold, D. A., Zacks, J. M., & Flores, S. (2017). Effects of cues to event segmentation on subsequent memory. *Cognitive Research: Principles and Implications*, *2*, 1.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–9.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–95.

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

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*, 767–773.

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

Gutbrod, K., 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/https://doi.org/10.1016/j.neuropsychologia.2006.01.014

Habermas, T., Diel, V., & Welzer, H. (2013). Lifespan trends of autobiographical remembering: Episodicity and search for meaning. *Consciousness and Cognition*, *22*, 1061–1073.

Hanson, C., & Hirst, W. (1989). On the representation of events: A study of orientation, recall, and recognition. *Journal of Experimental Psychology: General*, *118*(2), 136–47.

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, *19*, 304–313.

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23.

Healey, M. K., Kahana, M. J., Lohnas, L. J., Ramayya, A. G., Kuhn, J. R., Hower, K., & Healey, K. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575–96.

Healey, M. K., & Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory & Cognition*, *47*, 719–737.

Herweg, N. A., Sharan, A. D., Sperling, M. R., Brandt, A., Schulze-Bonhage, A., & Kahana, M. J. (2018). Reactivated Spatial Context Guides Episodic Recall. *The Journal of Neuroscience*, *40*, 2119–2128.

Heusser, A. C., Ezzyat, Y., Shiff, I., & Davachi, L. (2018). Perceptual Boundaries Cause Mnemonic Trade-Offs Between Local Boundary Processing and Across-Trial Associative Binding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1075–1090.

Horwath, E. A., Rouhani, N., DuBrow, S., & Murty, V. P. (2023). Value restructures the organization of free recall. *Cognition*, *231*, 105315. https://doi.org/https://doi.org/10.1016/j.cognition.2022.105315

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review*, *112*(1), 75–116.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–41.

Howard, M. W., & Kahana, M. J. (2002a). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299. https://doi.org/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*, 85–98.

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/https://doi.org/10.1111/j.1756-8765.2010.01112.x

Howard, M. W., Youker, T., & Venkatadass, V. (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

Iran-Nejad, A., & Winsler, A. (2000). Bartlett's Schema Theory and Modern Accounts of Learning and Remembering. *Journal of Mind and Behavior*.

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* (pp. 1724–1735, Vol. 33). Curran Associates, Inc.

Johnson-Laird, P. N. (1983). *Mental Models : Towards a Cognitive Science of Language*. Harvard University Press.

Kahana, M. J., Howard, M. W., & Polyn, S. (2008). Associative Retrieval Processes in Episodic Memory. *Psychology*, *3*.

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

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive science*, *34*(7), 1185–243.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*, 10687–10692.

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

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*, 3521–3526.

Klein, S. B. (2013). The temporal orientation of memory: It's time for a change of direction. *Journal of Applied Research in Memory and Cognition*, *2*, 222–234. https://doi.org/10.1016/j.jarmac.2013.08.001

Klein, S. B. (2016). Autonoetic consciousness: Reconsidering the role of episodic memory in future-oriented self-projection. *Quarterly Journal of Experimental Psychology*, *69*(2), 381–401.

Konidaris, G. D. (2016). Constructing Abstraction Hierarchies Using a Skill-Symbol Loop. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1648–1654.

Kragel, J. E., Ezzyat, Y., Lega, B. C., Sperling, M. R., Worrell, G. A., Gross, R. E., Jobst, B. C., Sheth, S. A., Zaghloul, K. A., Stein, J. M., & Kahana, M. J. (2020).

Distinct cortical systems reinstate the content and context of episodic memories. *Nature Communications*, *12*.

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.

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/https://doi.org/10.1037/a0028681

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*, 72–79.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. *Psychology of Learning and Motivation*, *74*, 195–232. https://doi.org/10.1016/bs.plm.2021.02.004

Laming, D. R. J. (2010). Serial position curves in free recall. *Psychological Review*, *117*(1), 93–133.

Lehnert, L., & Littman, M. L. (2019). Successor Features Combine Elements of Model-Free and Model-based Reinforcement Learning. *Journal of Machine Learning Research*, *21*, 196:1–196:53.

Lehnert, L., Tellex, S., & Littman, M. L. (2017). Advantages and Limitations of using Successor Features for Transfer in Reinforcement Learning. *Lifelong Learning: A Reinforcement Learning Approach workshop @ICML*.

Lengyel, M., & Dayan, P. (2007). Hippocampal Contributions to Control: The Third Way. *Neural Information Processing Systems*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of Extreme Events in Decision Making Reflects Rational Use of Cognitive Resources. *Psychological Review*, *125*, 1–32.

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

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

Lu, Q., Hasson, U., & Norman, K. (2022). A neural network model of when to retrieve and encode episodic memories. *eLife*, *11*, e74445. https://doi.org/10.7554/eLife.74445

Lu, Q., Hummos, A., & Norman, K. A. (2024). Episodic memory supports the acquisition of structured task representations. *bioRxiv*. https://doi.org/10.1101/2024.05.06.592749

Lu, Q., Nguyen, T., Zhang, Q., Hasson, U., Griffiths, T. L., Zacks, J., Gershman, S., & Norman, K. A. (2023). Reconciling Shared versus Context-Specific Information in a Neural Network Model of Latent Causes.

Machado, M. C., Barreto, A., & Precup, D. (2023). Temporal Abstraction in Reinforcement Learning with the Successor Representation. *Journal of Machine Learning Research*, *24*, 1–69.

Madan, C. R., Knight, A., Kensinger, E., & Mickley Steinmetz, K. (2020). Affect enhances object-background associations: Evidence from behaviour and mathematical modelling. *Cognition and Emotion*, *34*, 1–10. https://doi.org/10.1080/02699931.2019.1710110

Madan, C. R., Scott, S. M. E., & Kensinger, E. A. (2019). Positive emotion enhances association-memory. *Emotion*, *19*(4), 733–740.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt; Co., Inc.

Marshall, P. H., & Werder, P. R. (1972). The effects of the elimination of rehearsal on primacy and recency. *Journal of Verbal Learning and Verbal Behavior*, *11*(5), 649–653. https://doi.org/https://doi.org/10.1016/S0022-5371(72)80049-5

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/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*.

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

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

Mattar, M. G., Talmi, D., & Daw, N. D. (2019). Memory mechanisms predict sampling biases in sequential decision tasks. *The 4th Multi-disciplinary Conference on Reinforcement Learning and Decision Making*.

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.

McGovern, A., & Barto, A. G. (2001). Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density. *International Conference on Machine Learning*.

Michelmann, S., Hasson, U., & Norman, K. A. (2023). Evidence That Event Boundaries Are Access Points for Memory Retrieval. *Psychological Science*, *34*, 326–344.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

Miller, J., Lazarus, E., Polyn, S., & Kahana, M. (2013). Spatial Clustering During Memory Search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 773–781. https://doi.org/10.1037/a0029684

Miller, J., Weidemann, C., & Kahana, M. (2012). Recall termination in free recall. *Memory & cognition*, *40*(4), 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.

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

Momennejad, I., & Howard, M. W. (2018). Predicting the Future with Multi-scale Successor Representations. *bioRxiv*. https://doi.org/10.1101/449470

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*, e32548. https://doi.org/10.7554/eLife.32548

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic Memory and Beyond: The Hippocampus and Neocortex in Transformation. *Annual Review of Psychology*, *67*, 105–134.

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C. L., McAndrews, M. P., Levine, B., Black, S. E., Winocur, G., & Nadel, L. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: A unified account based on multiple trace theory. *Journal of Anatomy*, *207*(1), 35–66. https://doi.org/10.1111/j.1469-7580.2005.00421.x

Murdock, B. B. J. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482.

Murdock, B. B. J. (1972). Short-Term Memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (pp. 67–127, Vol. 5). Academic Press. https://doi.org/https://doi.org/10.1016/S0079-7421(08)60440-5

Murdock, B. B. J., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology*, *86*, 263–267.

Murray, J., Bernacchia, A., Freedman, D., Romo, R., Wallis, J., Cai, X., Padoa Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X.-J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, *17*, 1661–3. https://doi.org/10.1038/nn.3862

Musslick, S., Bizyaeva, A. S., Agaron, S., Leonard, N. E., & Cohen, J. D. (2019). Stability-Flexibility Dilemma in Cognitive Control: A Dynamical System Perspective. *Annual Meeting of the Cognitive Science Society*.

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757–775.

Naim, M., Katkov, M., Romani, S., & Tsodyks, M. (2020). Fundamental Law of Memory Recall. *Physical Review Letters*, *124*, 018101. https://doi.org/10.1103/PhysRevLett.124.018101

Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L. D., Koopman, H. J., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*, E3669–E3678.

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

Nielson, D., Smith, T., Sreekumar, V., Dennis, S., & Sederberg, P. (2015). Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences*, *112*(35), 11078–83. https://doi.org/10.1073/pnas.1507104112

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646.

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.

Osth, A., Reed, A., & Farrell, S. (2021). How do recall requirements affect decision-making in free recall initiation? A linear ballistic accumulator approach. *Memory & Cognition*, *49*, 968–983. https://doi.org/10.3758/s13421-020-01117-2

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.

Pashevich, A., Schmid, C., & Sun, C. (2021). Episodic Transformer for Vision-and-Language Navigation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15922–15932.

Patterson, K. E., Meltzer, R. H., & Mandler, G. (1971). Inter-Response Times in categorized free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 417–426.

Pettijohn, K., Thompson, A., Tamplin, A., Krawietz, S., & Radvansky, G. (2016). Event boundaries and memory improvement. *Cognition*, *148*, 136–144. https://doi.org/10.1016/j.cognition.2015.12.013

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

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

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

Polyn, S. M., Erlikhman, G., & Kahana, M. J. (2011). Semantic cuing and the scale insensitivity of recency and contiguity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(3), 766–75.

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.

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/https://doi.org/10.1016/j.neuropsychologia.2009.02

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., & Blundell, C. (2017). Neural episodic control. *International Conference on Machine Learning*, 2827–2836.

Pu, Y., Kong, X.-Z., Ranganath, C., & Melloni, L. (2022). Event boundaries shape temporal organization of memory by resetting temporal context. *Nature Communications*, *13*, 622. https://doi.org/10.1038/s41467-022-28216-9

Radvansky, G. A. (2012). Across the Event Horizon. *Current Directions in Psychological Science*, *21*, 269–272.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140.

Ratcliff, R., & McKoon, G. (1981). Does Activation Really Spread? *Psychological Review*, *88*(5), 454–462. https://doi.org/10.1037/0033-295x.88.5.454

Redish, A. D., Jensen, S., Johnson, A., Kurth-Nelson, Z., Larson, E. B., Gewirtz, J. C., Niv, Y., Murray, B., Jackson, J. C., Thomas, M. J., Smith, D., Nader, K., Gordon, J., Quirk, G., Paulus, M. P., Floresco, S. B., Schoenbaum, G., Steve, & Redish, D. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A Computational Model of Event Segmentation From Perceptual Prediction. *Cognitive Science*, *31*(4), 613–43.

Richie, R., Aka, A., & Bhatia, S. (2022). Free association in a neural network. *Psychological Review*, *130*(5), 1360–1382.

Richie, R., & Bhatia, S. (2020). Similarity Judgment Within and Across Categories: A Comprehensive Model Comparison. *Cognitive Science*, *45*(8), e13030.

Rimmele, U., Davachi, L., & Phelps, E. A. (2012). Memory for time and place contributes to enhanced confidence in memories for emotional events. *Emotion*, *12*(4), 834–46.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: Insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101–28.

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. *International Conference on Machine Learning*, 4354–4363.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings*

of the National Academy of Sciences of the United States of America, *102*(20), 7338–43.

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

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2019). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269.

Rumelhart, D. E. (1980). Schemata: The Building Blocks of Cognition (R. J. Spiro, B. C. Bruce, & W. E. Brewer, Eds.). *Theoretical Issues in Reading Comprehension*, 33–58. https://doi.org/10.4324/9781315107493-4

Rundus, D. (1971). Analysis of Rehearsal Processes in Free Recall. *Journal of Experimental Psychology*, *89*, 63–77. https://doi.org/10.1037/h0031185

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

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*, 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

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*, 229–242.

Sakon, J. J., & Kahana, M. J. (2021). Hippocampal ripples signal contextually mediated episodic recall. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(40), e2201657119. https://doi.org/10.1073/pnas.2201657119

Salvatore, N., & Zhang, Q. (2024). Parallels between neural machine translation and human memory search: A cognitive modeling approach. *Annual Meeting of the Cognitive Science Society*.

Sanders, H., Wilson, M. A., & Gershman, S. J. (2019). Hippocampal remapping as hidden state inference. *eLife*, *9*, e51140. https://doi.org/10.7554/eLife.51140

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, *129*, 241–255.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 773–786.

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.

Schapiro, A., Rogers, T., Cordova, N., Turk-Browne, N., & Botvinick, M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486–492. https://doi.org/10.1038/nn.3331

Schmidt, K., Patnaik, P., & Kensinger, E. A. (2011). Emotion's influence on memory for spatial and temporal context. *Cognition and Emotion*, *25*, 229–243.

Schulman, J., Levine, S., Moritz, P., Jordan, M., & Abbeel, P. (2015). Trust Region Policy Optimization. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 1889–1897, Vol. 37). PMLR.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*, 1593–1599.

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.

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*, 927–939.

Sharot, T., & Yonelinas, A. P. (2008). Differential time-dependent effects of emotion on recollective experience and memory for contextual information. *Cognition*, *106*, 538–547.

Sharpe, M. J., Batchelor, H. M., Mueller, L. E., Chang, C. Y., Maes, E. J. P., Niv, Y., & Schoenbaum, G. (2020). Dopamine transients do not act as model-free prediction errors during associative learning. *Nature Communications*, *11*.

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., & Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*, 735–742.

Shipley, T. F., & Maguire, M. J. (2008). Geometric Information for Event Segmentation. In *Understanding Events: From Perception to Action* (pp. 415–435). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195188370.003.0018

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Singh, M., & Hoffman, D. D. (2001). Part-Based Representations of Visual Shape and Implications for Visual Cognition. *Advances in Psychology*, *130*, 401–459.

Smith, J., Linderman, S. W., & Sussillo, D. (2021). Reverse engineering recurrent neural networks with Jacobian switching linear dynamical systems. *Neural Information Processing Systems*.

Sols, I., Dubrow, S., Davachi, L., & Fuentemilla, L. (2017). Event Boundaries Trigger Rapid Memory Reinstatement of the Prior Events to Promote Their Representation in Long-Term Memory. *Current Biology*, *27*, 3499–3504.e4.

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

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

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

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1), 181–211. https://doi.org/https://doi.org/10.1016/S0004-3702(99)00052-1

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

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

Tan, L., & Ward, G. (2000). A Recency-Based Account of the Primacy Effect in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1589–625. https://doi.org/10.1037//0278-7393.26.6.1589

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*, 1279–1285.

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* (pp. 1068–1074, Vol. 9). MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2018). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, *16*.

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

Tversky, B., Zacks, J., & Hard, B. (2008). The Structure of Experience. Oxford University Press.

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Li, Y., & 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.

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.

Wang, F., & Diana, R. A. (2017). Temporal context in human fMRI. *Current Opinion in Behavioral Sciences*, *17*, 57–64.

Wang, L., & Dunson, D. B. (2011). Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *20*, 196–216.

Wen, T., & Egner, T. (2022). Retrieval context determines whether event boundaries impair or enhance temporal order memory. *Cognition*, *225*, 105145. https://doi.org/10.1016/j.cognition.2022.105145

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2019). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, *183*, 1249–1263.e23.

Williams, S. E., Ford, J. H., & Kensinger, E. A. (2022). The power of negative and positive episodic memories. *Cognitive, Affective & Behavioral Neuroscience*, *22*, 869–903.

Wimmer, G. E., & Shohamy, D. (2011). The striatum and beyond: Contributions of the hippocampus to decision making. In *Decision making, affect, and learning: Attention and performance xxiii*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199600434.003.0013

Xia, L., & Collins, A. G. E. (2020). Temporal and state abstractions for efficient learning, transfer and composition in humans. *bioRxiv*.

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

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction Error Associated with the Perceptual Segmentation of Naturalistic Events. *Journal of Cognitive Neuroscience*, *23*, 4057–4066.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–93.

Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the alzheimer type. *Psychology and Aging*, *21*(3), 466–82.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3–21.

Zadbood, A., Chen, J., Leong, Y., Norman, K., & Hasson, U. (2017). How We Transmit Memories to Other Brains: Constructing Shared Neural Representations Via Communication. *Cerebral Cortex*, *27*, 4988–5000. https://doi.org/10.1093/cercor/bhx202

Zalla, T., Pradat-Diehl, P., & Sirigu, A. (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia*, *41*, 1619–1627.

Zhang, Q., Griffiths, T. L., & Norman, K. A. (2022). Optimal policies for free recall. *Psychological Review*.

Zhang, X., Liu, L., Long, G., Jiang, J., & Liu, S. (2020). Episodic memory governs choices: An RNN-based reinforcement learning model for decision-making task. *Neural Networks*, *134*, 1–10.

Zhao, W. J., Richie, R., & Bhatia, S. (2021). Process and content in decisions from memory. *Psychological Review*, *129*(1), 73–106. https://doi.org/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. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*, 1682–1688.

Zhou, C. Y., Talmi, D., Daw, N. D., & Mattar, M. G. (2024). Temporally extended decision-making through episodic sampling. *Annual Meeting of the Cognitive Science Society*.

Zhou, C. Y., Talmi, D., Daw, N. D., & Mattar, M. G. (in press). Episodic retrieval for model-based evaluation in sequential decision tasks.

Zhou, Z., Singh, D., Tandoc, M., & Schapiro, A. (2023). Building Integrated Representations Through Interleaved Learning. *Journal of Experimental Psychology: General*, *152*, 2666–2684. https://doi.org/10.1037/xge0001415

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–85.

Chapter 7

Supplementary Materials: TCM-SR

7.1 Proofs

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

In addition, derivations and proofs in this section assume all feature vectors are one-hot coded, and that the starting context is the same as the feature vector associated with the starting state. i.e. $\mathbf{c}_0 = \mathbf{x}_0$. We use $\mathbf{x}(s_n)$ to indicate the location of one at s_n in feature vector $\mathbf{x}$. For clarity, the policy π and discount factor γ during the encoding phase are implicit in the following proofs. e.g. using $\mathbf{M}$ as a shorthand for $\mathbf{M}_\gamma^\pi$.

Independent samples from memory yield unbiased value estimates

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

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

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

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

$$\begin{aligned}
\mathbb{P}(S_1) &= \frac{1}{1/(1-\gamma)} \left(\mathbf{M}^{\text{CS}} \left(\rho\mathbf{c}_0 + \beta\mathbf{M}^{\text{SC}}\mathbf{x}_0\right)\right)' \mathbf{x}_1 && \text{(Eq. 2.10)} \\
&= (1-\gamma)\left(\mathbf{M}^{\text{CS}}\mathbf{x}_0\right)'\mathbf{x}_1 \\
&= (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_1
\end{aligned}$$

Now consider arbitrary time step $i > 1$. By Eq. 1.1, $\mathbf{c}_i = \mathbf{c}_{i-1} = \cdots = \mathbf{c}_0 = \mathbf{x}_0$. Thus $\mathbb{P}(S_i) = (1-\gamma)\mathbf{x}_0'\mathbf{M}\mathbf{x}_i$. □

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

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

Proof. Denote the feature representation of state $s_k \in \mathcal{S}$ as $\mathbf{x}(s_k)$. Consider the expected

reward of the i-th sample:

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

By linearity of expectation,

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

Rearranging the terms, we have

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

□

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

The contiguity effect suggests value estimation via rollouts

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

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

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

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

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

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

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

□

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

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

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

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

By linearity of expectation,

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

Rearranging the terms, we have

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

□

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

Proposition 4.1. *Given $\rho = 0, \beta = 1, p_{stop} \in [0, 1]$, and arbitrary encoding γ, the effective discount factor $\tilde{\gamma}$of the estimated value satisfies $\tilde{\gamma} = \gamma p_{stop} - p_{stop} + 1$.*

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

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

By linearity of expectation,

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

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

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

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

In summary, in a generalized rollout sampling regime, an action can be evaluated in an unbiased manner by adding up rewards retrieved from episodically sampling the

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

Data from free recall experiments suggests an intermediate regime

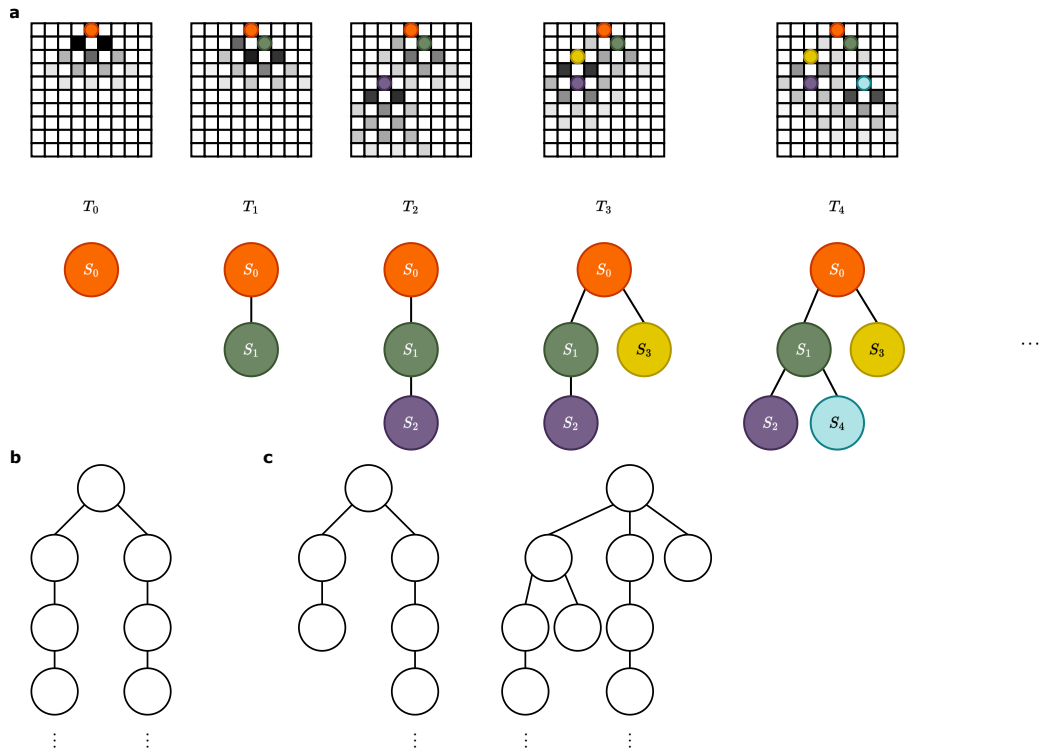

Figure 7.1: **Visualization of the intermediate sampling regime. (a)** A possible sequence of samples obtained using the intermediate sampling regime ($\rho, \beta > 0$). The tree starts with a single root node in orange representing the state-action to be evaluated. The other colored circles on the 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 SR-defined distribution defined at the parent (at a higher level of the tree). Greyscale 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 "stub", or short branches counting from the root node. See the online article for the color version of this figure.

Observe that the sequentially obtained samples can be conceptualized as a random tree with root at S_0 (Fig. 7.1a). 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 Eq. 1.1. In general, for any $i > j \geq 0$ we have

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

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

Lemma 5. *Assume $\rho + \beta = 1$, $\rho, \beta > 0$. As $N \to \infty$, T_N is expected to be a tree with $1/(1-\rho)$ degrees at the root and linear graphs thereafter.*

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

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

An illustration of such a tree structure in expectation is shown in Fig. 7.1b.

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

Corollary 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 $\rho + \beta = 1$, $\rho, \beta > 0$,*

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

Proof. Here we provide a sketch of the formal proof: note that the extreme cases where one of ρ, β is 1 can be realized as a random recursive tree describe above. Specifically, as $N \to \infty$, $\rho = 1$ corresponds to a tree with height 1 and infinitely many branches at the root; $\rho = 0$ corresponds to a path graph with infinite height. Importantly, given such a tree, we know $v_{\gamma=1}(S_0)$ may be computed as the expected total return of an arbitrary path from the root node to a leaf node in a random tree T_∞. i.e., sum along paths and average across paths from Theorem 2 and 4 respectively. The result then directly follows from Lemma 5 noting $1 - \rho = \beta$. □

Lemma 7. *Given $\rho + \beta = 1$, $\rho, \beta > 0$, $N < \infty$, there is a non-zero probability that the shortest path from the root node to a leaf has length 2.*

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

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

A few examples of such possible tree structures are shown in Fig. 7.1c. □

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 resulted 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 downwards in the case of relatively small number of samples, but like previous cases, given a sufficiently large number of samples, the estimate approaches the true value. 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 non-salient stimuli to learn $\mathbf{M}^{\mathrm{CS}}$. For clarity, a state s either contains nothing (i.e. $\mathcal{R}(s)=0$) or a small reward ($\mathcal{R}(s)=1$). All else being equal, the resultant, emotionally modulated 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

$\overline{\mathbf{M}}' \neq \mathbf{M}'$. To compute an estimation of expectation, it needs *importance sampling* to translate distributions of $\mathbf{M}$ to $\overline{\mathbf{M}}$.

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

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

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

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

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

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

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

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