Prioritizing replay when future goals are unknown

Yotam Sagiv[a], Thomas Akam[b], Ilana B. Witten[a], Nathaniel D. Daw[a]

[a]*Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey, USA*
[b]*Department of Experimental Psychology, Oxford University, Oxford, UK*

Abstract

Although hippocampal place cells replay nonlocal trajectories, the computational function of these events remains controversial. One hypothesis, formalized in a prominent reinforcement learning account, holds that replay plans routes to current goals. However, recent puzzling data appear to contradict this perspective by showing that replayed destinations lag current goals. These results may support an alternative hypothesis that replay updates route information to build a "cognitive map." Yet no similar theory exists to formalize this view, and it is unclear how such a map is represented or what role replay plays in computing it. We address these gaps by introducing a theory of replay that learns a map of routes to candidate goals, before reward is available or when its location may change. Our work extends the planning account to capture a general map-building function for replay, reconciling it with data, and revealing an unexpected relationship between the seemingly distinct hypotheses.

Keywords: Hippocampal replay, reinforcement learning, geodesic representation

Acknowledgements

This work was supported by U.S. Army Research Office grant W911NF-16-1-0474 and National Institutes of Mental Health grant R01MH121093.

1. Introduction

Much recent attention has been paid to experience replay as a candidate mechanism subserving learning and complex behaviour. In particular, sequential replay of nonlocal trajectories during sharp-wave ripples in the

hippocampal place cell system [1, 2, 3, 4, 5, 6] (and similar events elsewhere [7, 8, 9]) serve as a tantalizing example suggesting some kind of navigation-related computation [10, 5, 11, 12, 13, 14, 15, 16]. However, the precise functional role for these events — what replay is actually computing — remains a central question in the field.

There are, broadly, two schools of thought about this question. One view, the "value hypothesis," suggests that a purpose of replay is to facilitate planning or credit assignment, in the service of directly guiding current or future choices [17, 3, 18, 19, 5, 20, 21, 22, 23, 16, 24, 25, 26]. Under this view, the replay of extended trajectories facilitates connecting candidate actions at some location with their potential rewarding consequences elsewhere in space (e.g., by updating a decision variable such as the value function). A contrasting view, the "map hypothesis," argues instead that replay is concerned with building (or remembering/consolidating) some abstract representation of the environment per se (e.g., a "cognitive map" of its layout), and is not straightforwardly tied to subsequent behaviour or to reward [1, 27, 4, 28, 29, 30, 31, 32]. Both interpretations have been argued to be consistent with data, though their differential predictions in many situations are often not obvious.

A recent theoretical model suggested an approach for improving the empirical testability of these functional ideas. Mattar and Daw [17] formalized a version of the value hypothesis in a reinforcement learning (RL) model, specifying the particular computation (a DYNA-Q [33] value function update) hypothetically accomplished by each individual replay event. This reasoning implies testable claims about how each replay event should affect subsequent choices (by propagating reward information to distal choice-points) [26, 34]. Furthermore, they argued that, given a precise enough hypothesis about the effects of a replay on behavior, it is possible to derive a corresponding formal hypothesis about the *prioritization* of replayed trajectories; that is, if this were indeed the function of replay, then the brain would be expected to favor those trajectories that would maximize expected reward by best improving choices. This idea led to testable predictions (e.g., about the statistics of forward vs. reverse replay in different situations) that are well fit to data in many contexts.

Sharper empirical claims in turn have permitted clearer falsification and refinement. Accordingly, multiple authors using goal-switching tasks (here we focus on work by Gillespie et al. [32] and Carey et al. [35]) have recently reported that replayed trajectories tend to be systematically focused on past

goals rather than current ones, and thus to lag rather than lead animals learning updated choice behavior. The decoupling between the change in behavior and the content of replay has been suggested to disfavor the value hypothesis, which would predict that these quantities should track each other, and instead support the map hypothesis. In the present work, we aim to explain these results, and reconcile them with an updated general account of replay by extending Mattar's approach to encompass the map hypothesis.

Indeed, although the value/map division appears intuitively straightforward, a critical challenge is that the cognitive map hypothesis remains incompletely specified. On one side, the concept of long-term reward prediction from RL theories offers a precise formalization of the value hypothesis, while on the other there has been less formal attention to the map hypothesis, starting with the question of what the "map" is, and therefore what it means for replay to be building it. Outside the replay context, RL models generally operationalize the cognitive map as the local connectivity and barriers (the "one-step" state-action-state adjacency graph) of the environment [36]. However, it seems paradoxical to assert that replay events are involved in building this representation, because replayed trajectories already reflect local connectivity, even immediately after encountering novel barriers [11]. Another suggestion in the "map" camp is that the goal of replay is to form or maintain memories about visited locations [32]; however, it remains unclear what memory content is maintained and also what specific locations are favored for maintenance and why. In short, it remains a central open question what "map" structure is hypothetically being built by trajectory replay, and what the precise computational role of replay is in building it.

Our new account addresses these questions by providing a new view of "map" replay that extends the key logic from Mattar's model (that trajectory replay propagates local reward information over space to produce long-run value representations) to a setting where the locations of rewards are unknown or dynamic. The analogue of the value function (the target of computation in the Mattar model) in the new setting is a set of long-run routes (effectively, goal-specific value functions) toward different possible goals, similar to a successor representation (SR) [37]. This provides a formal notion of a "cognitive map" that goes beyond local information; a role for trajectory replay in computing it (like DYNA-SR [38, 39]); and a corresponding priority metric quantifying which replays should be most useful. The new prioritization rule generalizes the Mattar one: replay has value to the extent it helps you reach either current goals or potential future ones, weighted ac-

cording to learned beliefs about which goals are more likely. This account strictly generalizes Mattar's and in this way reconciles the planning and map views: value replay arises as a special case of map replay when the agent's behavioural goals are fixed and known.

We show that this theory addresses the Gillespie and Carey results, while maintaining the key insights offered by the original Mattar account when a single, static set of goals predominates. Furthermore, we present predictions made by our model that may be used to validate it using future studies.

2. The model

2.1. The Geodesic Representation

We begin by describing how we operationalize the term "cognitive map". In principle, a cognitive map may be understood as any representation of an agent's task contingencies. In particular, in RL models, a cognitive map (or "internal world model") is traditionally associated with the one-step transition function $T(s_t, a_t, s_{t+1}) = P(s_{t+1} \mid s_t, a_t)$ that captures how local actions a (e.g., directional steps) affect the current state s (e.g., location).

However, given just this local map, plus local goal information (e.g., the one-step reward values $r(s)$ associated with each location), it still takes substantial computation to find the long-run optimal actions, e.g. by computing the long-run aggregate rewards resulting from different candidate actions. Formally, this is the state-action value $Q(s,a) = \mathbb{E}_{s' \sim P(s'|s,a)}\Big[r(s') + \gamma \max_{a'} Q(s', a')\Big]$. Previous theories [17] suggest that a goal of replay is to facilitate computing Q by aggregating reward over replayed trajectories. The resulting value function is goal-specific, in the sense that if the one-step rewards change (e.g., if a rewarding goal moves from one location to another), a new value function must be computed. Thus Q-learning and similar methods are inflexible in the face of changing goals, requiring additional computation.

One way to address this limitation of Q-learning is to represent a map not in terms of local adjacency relationships between neighboring states, but instead shortest path distances from start states to many possible goal states. An alternative way to conceptualize the same approach is to maintain not a single value function, instead a set of value functions for many different reward configurations. One common version of this idea is the successor representation (SR) [37, 40]. Here we introduce a variant of the SR, the Geodesic Representation (GR), inspired by Kaelbling [41], which is based on

the same state-action value function as Q-learning and allows for "off-policy" learning that facilitates transfer to later tasks. The GR aims to learn the shortest paths from each state in the environment to a distinguished subset of possible "goal" states (which may be, in the extreme case, all other states).

In particular, consider an episodic task taking place in an environment with a single terminal state g that delivers unit reward. The state-action value function $Q(s, a)$ for this environment measures, for each state s and action a, their distance from g (i.e. the terminal value 1 discounted by the number of steps optimally to reach it; Fig. 1a). Consequently, the optimal policy in this task can be thought of as a "distance-minimizing" policy that maximizes return by minimizing the number of times the eventual reward is discounted by the temporal discount factor.

Accordingly, we define the GR as a stack of these Q-value tables (Fig. 1b), with each "page" in the stack encoding the state values in a modified version of the underlying environment where the corresponding goal is the only rewarding state, confers a unit reward, and is terminal. As in the earlier example, policies derived from each of these pages facilitate optimal navigation to their associated goal state, as they are return-maximizing (distance-minimizing) in the associated MDP. That is:

$$G(s, a, g) \equiv \mathbb{E}_{\pi_g}\Big[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}_{s_t=g} | s_0 = s, a_0 = a\Big] \tag{1}$$

where g is any state distinguished as a potential goal, $\gamma \in [0, 1)$ is a temporal discount rate, π_g is the optimal policy for reaching g[1], and $\mathbb{I}_\bullet$ is the indicator function that is 1 if $\bullet$ is true and 0 otherwise. This definition simply encodes the intuition from above: $G(s, a, g)$ is the expected (discounted) reward for taking action a in state s, and thereafter following the optimal policy for reaching state g, in an environment where only g is rewarding. Example slices of the GR in an open field environment with a walled area are illustrated in Fig. 1c.

Another way of characterizing the GR is by its Bellman equation:

$$G(s, a, g) = \mathbb{E}_{s' \sim P(s'|s,a)}\Big[\mathbb{I}_{s'=g} + \gamma \mathbb{I}_{s' \neq g} \max_{a'} G(s', a', g)\Big] \tag{2}$$

[1] i.e., is reward-maximizing in the MDP where g is terminal and is the only rewarding state.

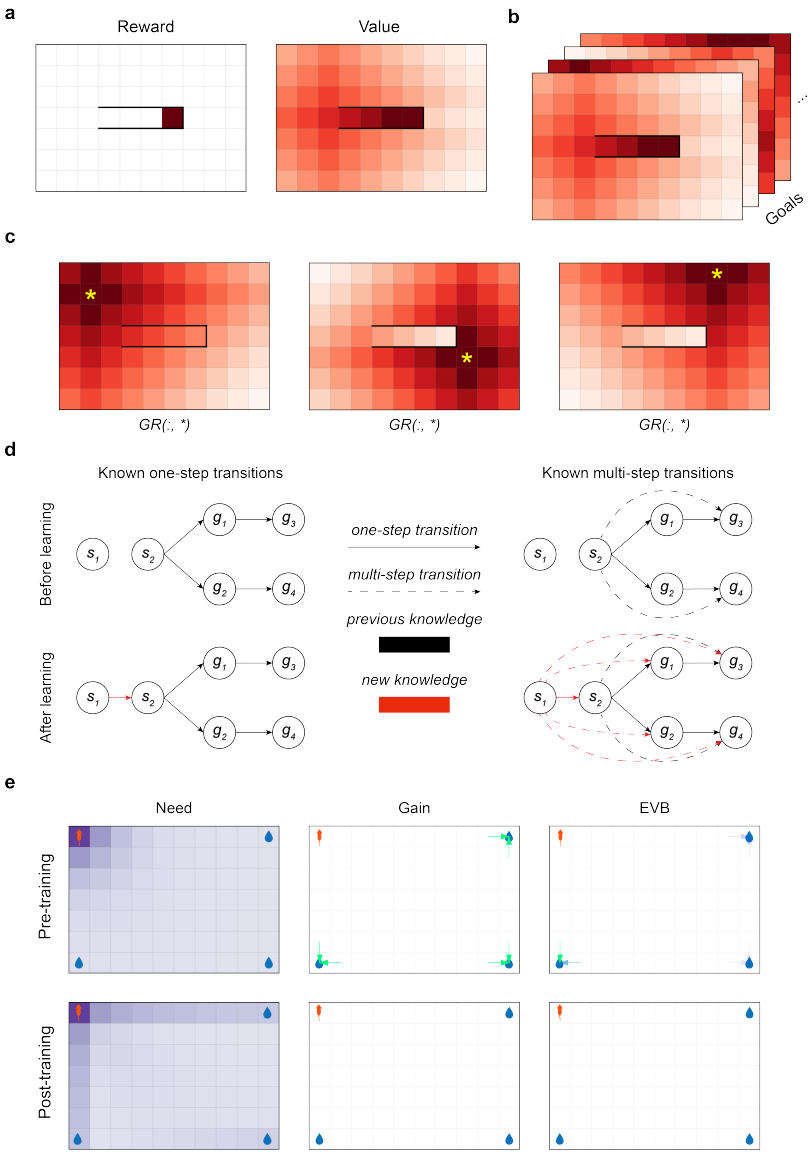

Figure 1: **The Geodesic Representation.** **(a)** Left: an open-field environment with a walled corridor that encloses a single rewarded state. Right: the state-value function induced by the single reward. The reward state has been assigned a value of 1 for clarity. **(b)** The GR is a stack of state-action value functions, one for each goal. (Note that for simplicity, we illustrate the value functions over states rather than over state-action pairs.) **(c)** Illustrations of three different slices of the GR in an open field environment with...

Figure 1: ... a walled enclosure. The candidate goal state corresponding to the slice is indicated with a yellow asterisk. **(d)** An agent learns how to reach a variety of goals via a single learning step from s_1 to s_2. Black arrows: knowledge prior to learning, red arrows: knowledge gained after the transition from s_1 to s_2. Solid lines: one-step transitions, dashed lines: implied longer-horizon connections. Top-left: the agent's knowledge about the environment's one-step transition structure before learning. Top-right: the agent's knowledge about the environment's long-run connections before learning. Bottom-left: the agent's knowledge about one-step structure after learning. Bottom-right: the agent's knowledge about multi-step structure after learning. **(e)** Visualisations of need, gain, and EVB in a simple open field environment where the agent starts in the top left corner, and there are candidate goal locations in the other three corners. Top row: before replay has occurred, bottom row: after the GR has converged. In the gain and EVB plots, arrow colour and opacity indicates the value of the relevant metric (only arrows corresponding to transitions with > 0 gain or EVB are shown).

Intuitively, if s' is the goal state g, then transitioning to it should accrue a reward of 1 and if s' is not, then the current value should be γ times the value at wherever we arrived based on taking the best available action there. This can be also written compactly in vector form:

$$G(s, a, :) = \mathbb{E}_{s' \sim P(s'|s,a)} \left[\mathbb{1}_{s'} + \mathbb{0}_{s'} \odot \gamma \max_{a'} G(s', a', :) \right] \tag{3}$$

where $\odot$ denotes the Hadamard (elementwise) product, $\mathbb{1}_{s'}$ is a one-hot vector at s', $\mathbb{0}_{s'}$ is a vector that is 0 at s' and 1 everywhere else, and the max is taken separately over each goal state. This form of the equation demonstrates that information about distances to multiple goals can be updated through a single learning step (e.g. via a vector of off-policy temporal-difference updates, one for each goal, based on this Bellman equation, as has been proposed for the SR [37, 42]). Consider, for example, the setup in Fig. 1d where an agent knows how to get to goals $g_1, ..., g_4$ from state s_2 but not s_1. If the agent were to undergo the transition $s_1 \rightarrow s_2$, they could learn how to get to all of the goals that s_2 is already connected to in a single learning step.

It is worth noting explicitly that the GR object itself is updated strictly based on observing a state-action-successor state tuple (s, a, s'). The off-policy nature of the update, combined with the lack of an environmental reward term, means that the GR is not sensitive to either the exploratory policy generating the updates nor the reward function governing the agent's behavior during learning. This means that once a GR is learned, an agent can adapt to a new goal (e.g., reward moved from one location to another)

simply by switching which "page" $G(:,:,g)$ controls behavior. Such nimble switching is unlike Q-learning (which must relearn a new value function in this case), and importantly implies that replay can have utility (in the sense of increasing future reward by improving future choices) due to "pre-planning." That is, since the GR is robust to changes to goal locations, learning updates made to it in one goal regime remain relevant even when goals change. This means that replay can improve the choices that the agent makes, even later when the goals are different than they were at the time of learning.

Finally, consider the GR's relationship to Q-learning. The GR with a single goal location is exactly equivalent to Q-learning in the case of a single, terminal reward. Thus the new theory generalizes an important case of its predecessor: from one goal to several, mutually exclusive candidate goals, which may be available at different times. Although in the present work we concentrate on this case, the GR can also be used for a more general class of reward functions: those containing multiple, simultaneously available terminal rewards of different magnitudes. (The Q function is then given by the max over per-goal value functions across the corresponding GR "pages," each weighted by their reward magnitudes.)

2.2. Prioritizing replay on the GR

Previous work [17] considered the problem of prioritizing experience replay for Q-value updating. In particular, to extend a theory of replay's function to a theory of replay content, it was proposed that replay of a particular state-action-state event (so as to perform a Q-value update, known as a Bellman backup, for that event) should be prioritized greedily according to its expected utility, i.e. the difference between the agent's expected return after vs. before the update due to the replay. Replay can increase expected return if the update improves future choices. This "expected value of backup" (EVB) can be decomposed as a product of a "need" term and a "gain" term. Briefly, need roughly corresponds to how often the agent expects to be in the updated state (i.e., the state's relevance) and gain roughly corresponds to the magnitude of the change in the updated state's value (i.e., how much additional reward the agent expects to accrue should it be in that state by virtue of the update improving the choice policy there). Under this theory, different patterns of replay then arise due to the balance between need (which generally promotes forward replay) and gain (which generally promotes backward replay) at different locations.

Since the GR aggregates a set of Q functions, it can also be decomposed into the product of need and gain terms. Generalizing the approach from Mattar and Daw [17], we begin by defining an analogue of the state value function for any single goal g in the current setting:

$$H(s,g) \equiv \sum_a \pi_g(a|s) G(s,a,g) \tag{4}$$

where π_g is the policy that tries to reach g as fast as possible. It can be shown that the expected improvement in H after backing up the experience $e_k = (s_k, a_k, s'_k)$ with respect to a particular goal g factorizes (see Methods):

$$\begin{aligned} H_{post}(s,g) - H_{pre}(s,g) &= \text{need}(s_k,g) \times \text{gain}(e_k,g) \\ \text{need}(s_k,g) &= \sum_{i=0}^{\infty} \gamma^i P(s \to s_k, i, \pi_{g,pre}) \\ \text{gain}(e_k,g) &= \sum_a \big(\pi_{g,post}(a|s_k) - \pi_{g,pre}(a|s_k)\big) G_{post}(s_k,a,g) \end{aligned} \tag{5}$$

Here, $\bullet_{pre}$ and $\bullet_{post}$ refer to $\bullet$ before and after the update, respectively (and so, while a_k and s'_k do not explicitly appear above, they affect the equation by affecting the update from $\bullet_{pre}$ to $\bullet_{post}$). $P(s \to s_k, i, \pi_{g,pre})$ is the probability that a trajectory starting in s at time 0 arrives at s_k at time i when following policy $\pi_{g,pre}$. Intuitively, the need term Equation 5 measures how often the agent will reach the state being updated s_k given its current state s and its policy[2]. The gain term quantifies how much additional reward the agent should accumulate due to a change in policy due to the performed update. Roughly, we can understand this equation as saying that the utility of backing up some experience e_k, measured through the expected improvement $H_{post}(s,g) - H_{pre}(s,g)$, is driven by i) how relevant that experience is and ii) by the magnitude of the change induced by the update. See Fig. 1e for visualizations of the need, gain, and EVB terms in a simple environment.

So far, we have essentially followed Mattar's definition of EVB for a single value function. Here, since the GR comprises a set of value functions for multiple goals, and replay of a single experience (via Equation 3) updates all of them, we need to aggregate their value into an overall EVB. We did this simply by taking the expectation (or more generally the expected discounted

[2]In fact, it is precisely the SR evaluated under π_g.

sum) of these per-goal EVBs under a distribution over them encoding the agent's belief about which ones are likely to be relevant in the future. (How this distribution is learned or constructed is itself an interesting question; we make simple assumptions about it in this article since our main point is to expose the effects that goal uncertainty can have on replay.) In particular, an agent may prioritize replay by picking the memory that maximizes the expected improvement from the current state s averaged over all possible goals:

$$e^* = \arg\max_{e_k} \mathbb{E}_{g\sim P(g)}\Big[H_{post}(s,g) - H_{pre}(s,g)\Big] \tag{6}$$

Once an experience is picked for replay, the GR can then be updated for *all* goals towards the target in Equation 3.

3. Results

3.1. GR replay favors elements of routes shared between multiple goals

Since prioritized GR replay is a generalization of the prioritized replay account by Mattar and Daw [17] it retains all of its notable properties (e.g., exhibiting coherent forward and reverse sequences, spatial initiation biases, etc.). As such, we focus on exploring the novel properties of GR replay, which would not be attributable to Q-value replay for a single goal or reward function.

First, the explicit representation of distinguished goal states in the GR allows for simultaneous planning across multiple candidate goals. To expose this distinction clearly, we first consider a stylized situation. In Fig. 2a, we simulated prioritized replay, using both Q-learning and GR agents, on an asymmetric T-maze task in which the ends of each arm of the T contain rewards (and both are candidate goals), but one arm is shorter than the other. (We assume for the sake of simplicity that the reward states are terminal, that there is a single, known start state, and that the maze is viewed without online exploration, such as through a window as in Ólafsdóttir et al. [21].) In this case, the Q-learning agent (Fig. 2a, middle) replays a path from the closest reward location to the starting location and then stops, whereas the GR agent (Fig. 2a, right) replays paths from both the close and far candidate goal locations to the start, in order of distance. This distinction illustrates the fact that the Q-learning agent's objective is to build a reward-optimal policy – and as such, all it needs to learn how to do is to reach the nearest

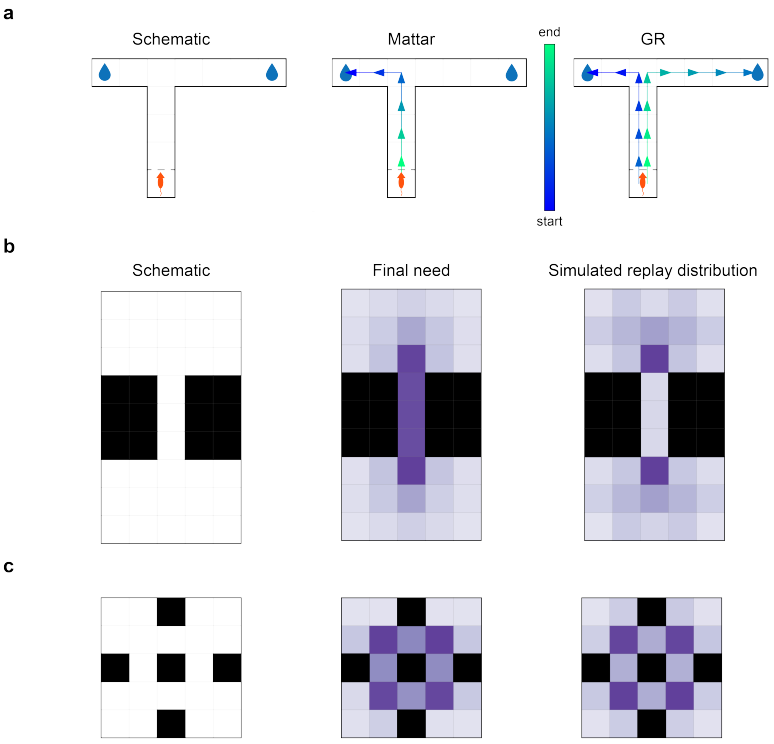

Figure 2: **The GR supports replay to multiple goals, and respects environmental structure.** **(a)** In an asymmetric T-maze task, Q-value replay only learns a path to the nearest goal, whereas GR replay learns paths to both goals. Left: task schematic, middle: Q-value replay, right: GR replay. **(b)** Replay in a bottleneck maze where every state is a candidate goal and also a potential starting location is biased towards topologically important states. Left: environment schematic, middle: asymptotic across-goal mean need after GR convergence in a single simulation, right: mean state replay across $n = 250$ simulated replay sequences. **(c)** As in (b), but in a maze analogue of the community graph. Left: schematic, middle: asymptotic need, right: mean state replay across $n = 250$ simulations.

reward – whereas the GR agent's objective is to learn the structure of the environment.

We next consider a less constrained setting, in which all locations in an environment are both candidate goals and potential starting locations, so that the target GR is an all-to-all map of shortest paths. One hallmark of GR-based replay in this case is that, because priority is averaged over goals, replays are particularly favored through locations that are shared between optimal routes to many different goals. For example, replay should be focused

around bottleneck states — states through which many optimal routes must pass in order to connect different starting states and goal states. To exemplify this prediction, we simulated GR replay while learning two environments with bottleneck states: a chamber with two large rooms, connected by a narrow corridor (Fig. 2b), and a four-room environment (based on Schapiro et al.'s [43, 44] community graph) in which each room can only be entered or exited by passing through a single location (Fig. 2c).

Recall that replay priority in the model is the product of need and gain terms. A preference for bottleneck states arises algebraically from the need term, as can be seen in the middle figures, which plot the need term under the optimal GR (i.e., after learning has converged). In both graphs, the bottleneck states have the highest need since they are required for all paths that cross between the rooms. This preference arises formally because need in the GR model corresponds to a variant of graph-theoretic betweenness centrality (BC, or the fraction of shortest paths in which a node participates; GR need is the same but counts participation for each step discounted by its distance from the goal). Supp Fig. A.6 shows that BC closely corresponds to need.

The analysis so far neglects the contribution of gain, and of the step-by-step progression of learning. These reflect the partly opposing contribution of one additional key feature of the model: the ability of a single replay event to drive learning about many different paths at once. Accordingly, the full simulated replay distributions (right plots) reflect the asymptotic need, but with an interesting elaboration. Namely, the internal states of the corridor (and similarly, the door states in the community graph) are replayed relatively less when compared to the BC values of those states. This is because (to the extent the agent first learns to come and go from the exit state to all other states in a room), all paths between the rooms can be bridged by a replay through the bottleneck. Stated differently, since a single GR update facilitates learning across many goals simultaneously, it is in principle adaptive to learn as much as possible about paths to goals within a given room, transfer that knowledge to the mouth of the bottleneck, and then carry it through to the next room in a single replay sequence (indeed, the "globally optimal" learning sequence for the bottleneck chamber should clearly only visit the interior states of the bottleneck precisely twice: once to update their policies for reaching left-room goals from right-room states, and once more for the reverse). Thus the model tends particularly to favor the endpoints of bottlenecks, relative to the middles.

3.2. GR replay accounts for previous-goal bias in maze navigation tasks

Recent studies [32, 35] examining replay in mazes with dynamically changing goals have presented a critical challenge to the "value view." Specifically, replay in these contexts tends to "lag" choice behavior in adapting to new goals, and thus displays a bias *away* from the current behavioral goal. This pattern appears incompatible with models in which replay directly drives behavioral adjustment; for instance, if replay modifies (for example) Q-values, and these Q-values dictate choice behavior, then replay should, if anything, lead changes in choice behavior. In this section, we show how this decoupling of replay from behavior is naturally explained in our GR model, due to the way it separates learning to reach candidate goals from learning what the current and likely future goals are.

In one study by Gillespie et al. [32], rats dynamically foraged for reward in an eight-arm maze, in which a single arm stably dispensed reward for a block of trials, but this target moved after the rewarded arm was sampled a fixed number of times (Fig. 3a). In each block, rats thus had to first identify by trial and error which arm was rewarding (the "search" phase) and then repeatedly go to it once it was found (the "repeat" phase). The value view is challenged by two key findings about the content of replay during the repeat phase (i.e., once the rats have discovered the new target and reliably visit it):

1. Overall, replayed locations featured the goal arm from the *previous* block more often than any other arm.
2. Replay of the *current* goal arm increased gradually throughout the repeat phase (i.e., over repeated sampling of that arm).

To capture these effects, we simulated the Gillespie task using a GR agent. The agent separately learned a GR (a set of routes to each goal), a representation of the current goal (a value estimate for each arm, used with a softmax to select goals for the GR agent to visit), and finally a representation of the overall distribution of goals (to prioritize replay for maintaining the GR).

The key insight explaining Gillespie et al.'s findings is the distinction between these two representations of the goals, which serve different purposes: while choice behavior needs to nimbly track the current goal (i.e., it must track the *within-block* reward function), replay must be prioritized in part to learn routes to locations where future rewards are likely to be found (i.e., it must be guided by the *across-block* distribution of goals). Accordingly, in the model simulations we capture these two timescales using

Rescorla-Wagner learning with different learning rates (higher to track the current goal; lower to capture the distribution of goals across blocks). The former achieves quick behavioral switching while the slower rate focuses replay on previously rewarded arms (reflecting where reward density, viewed across blocks, has recently been most common), thus only gradually turning its focus to the current goal. We have also assumed that the agent's GR was subject to a small amount of decay (i.e., forgetting) on every time-step (see Methods). This is a standard assumption in learning models, typically justified by the possibility of contingency change [45, 46], and has the effect of ensuring that learning continues in ongoing fashion rather than stopping at asymptote. As such, one can interpret the role of replay after the initial structure learning as maintaining the learned representation in the face of forgetting or environmental change.

Accordingly, prioritized GR replay from our agent qualitatively matched the patterns observed in Gillespie et al. [32]. Overall, within a block, GR replay displayed a bias for the previous-block goal arm; in contrast, a Q-learning agent using the prioritized replay scheme from Mattar and Daw [17] preferred the current goal arm (Fig. 3b). Furthermore, and also consistent with the data, replay of the current goal arm increased over the course of the block for the GR agent while it decreased for the Q-learning agent (Fig. 3c).

The same considerations explain similarly challenging results from Carey et al. [35]. Here, rats repeatedly traversed a T-maze where one arm provided food reward and the other provided water reward (Fig. 4a). Each day, the rats were alternately deprived of either food or water. Echoing the goal-switching result [32], even though choice behavior favored the motivationally relevant reward, replay recorded during the task was largely biased towards the behaviourally non-preferred arm (Fig. 4b).

We again simulated the Carey task using a GR agent equipped with a fast-learning behavioural module and a goal distribution created by slow Rescorla-Wagner learning. The effects of alternating food and water deprivation were realized by having asymmetric reward values for the two arms that switched between sessions [47]. Replay simulated from this model recapitulated the mismatched behaviour-replay pattern observed in the data (Fig. 4b), as before, because GR updates are guided by across-block reward experiences. In contrast, replay simulated from a Q-learning agent displayed a matched preference for the relevant reward in both behavior and replay (Fig. 4b).

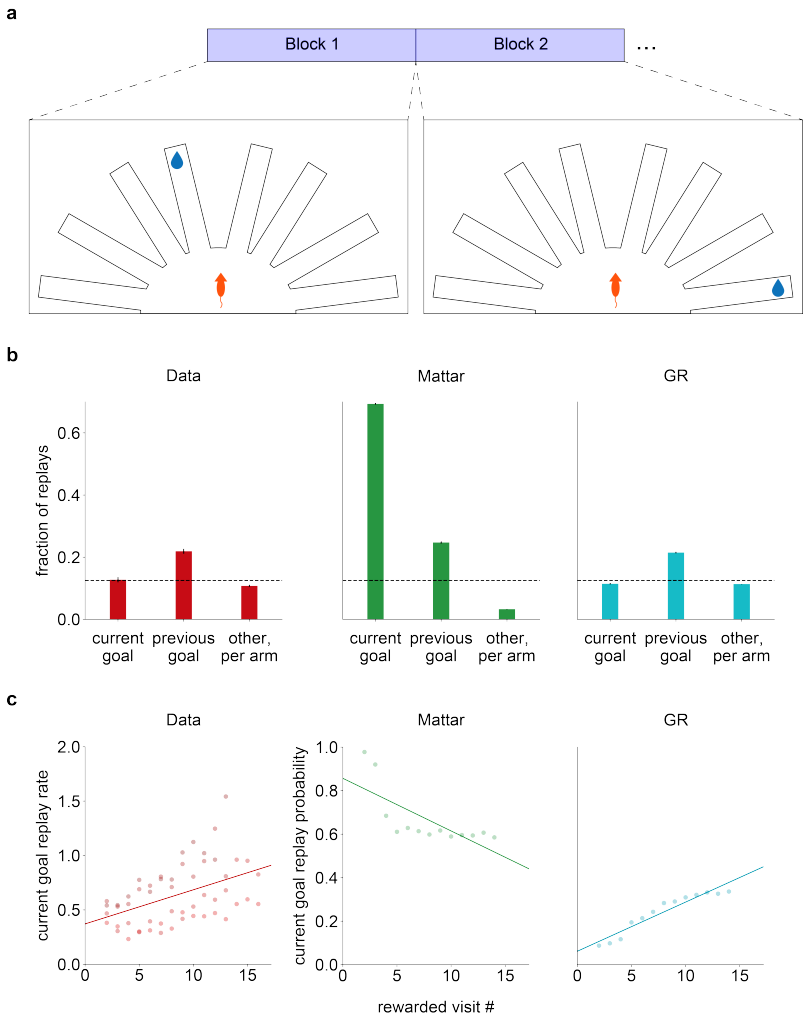

Figure 3: **GR replay captures replay-behavior lag in Gillespie et al. [32] (a)** Task schematic. **(b)** Fraction of replays including either the current goal arm, the previous goal arm, or any of the other six arms (normalized per arm). Left: replotted data from Gillespie et al., averaged across rats, middle: Q-value replay, right: GR replay. **(c)** Rate of current goal replay within a block as a function of rewarded visit number. Left: replotted data from Gillespie et al., averaged across rats, middle: Q-value replay, right: GR replay.

3.3. GR replay trades off current goals against future goals by occupancy

We have so far emphasized that the current model extends the previous value view to, additionally, favor replay of routes to candidate goals to the extent these may be expected in the future. This role of expectancy in

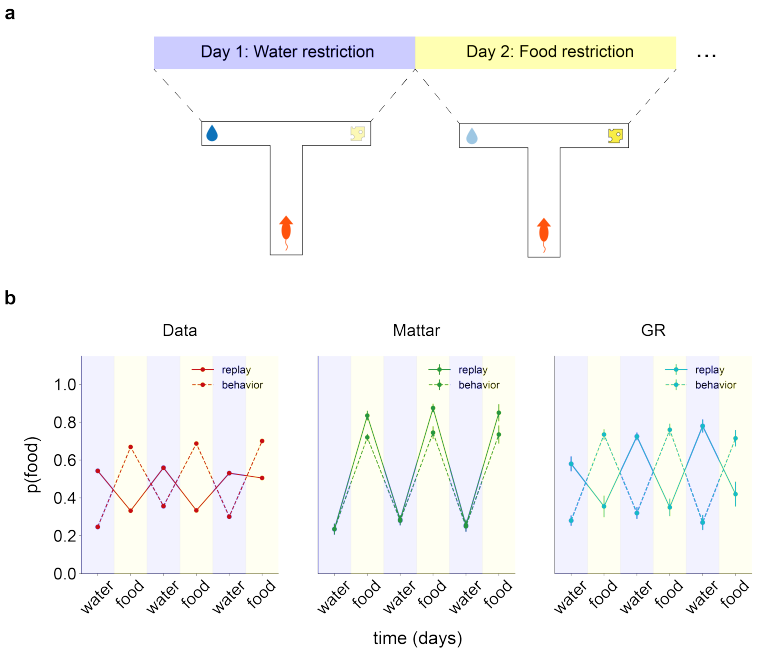

Figure 4: **GR replay captures replay-behavior lag in Carey et al. [35] (a)** Task schematic. **(b)** Probability of seeking the food reward ("behavior") and of replaying the food arm ("replay") as a function of session number/deprived substance. Left: replotted data from Carey et al., middle: Q-value replay/behavior, right: GR replay/behavior.

weighting the utilities of replays with respect to different goals also implies one of the key predictions of this model to be tested in future empirical work: that replay of current vs. other possible goals in an environment should be sensitive to their switching statistics.

To illustrate this, we simulated a dynamic maze navigation task with a detour re-planning manipulation (Fig. 5a). Here, an agent learned to navigate a maze to get to one of two candidate goal states. On any trial only one of the goals was active; given the active goal on trial $t-1$, the active goal on trial t was determined by the dynamic switching process in Fig. 5b. In this maze structure, the shortest path from the start state to either goal state passes through the middle bottleneck state. Consequently, an agent executing this task should learn to always route through that bottleneck regardless of the current goal. However, if on some trial it were to find that the middle bottleneck is inaccessible (for example, if the path to it was blocked by a wall), it would need to re-plan using replay in order to build new policies for

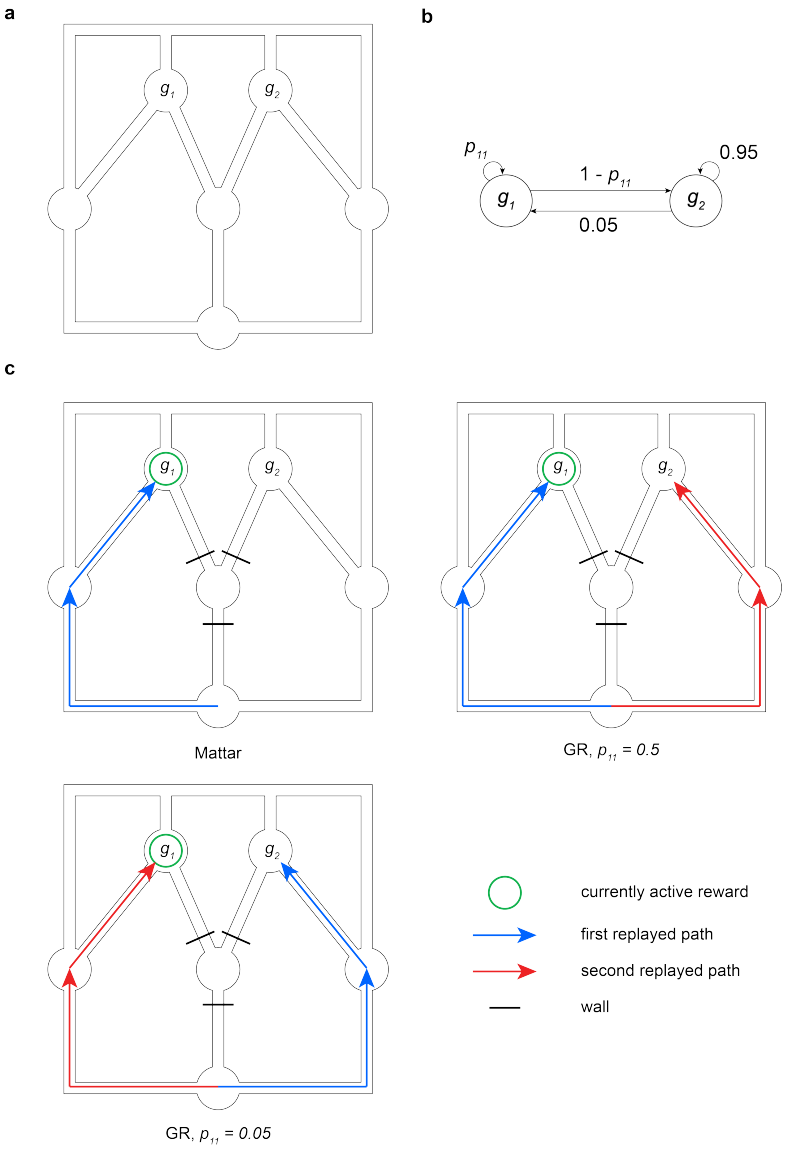

Figure 5: **GR replay trades off future and current goals.** **(a)** Maze schematic. **(b)** Goal dynamics process. p_{11} indicates the probability of goal 1 being active on trial t if it is active on trial $t-1$. **(c)** First and, if relevant, second replays for different models and parameter settings when an agent discovers that the middle bottleneck is inaccessible on a trial where g_1 is active (denoted by the green circle). Top-left: Q-value replay, top-right: GR replay with high p_{11}, bottom: GR replay with low p_{11}.

reaching each goal. Furthermore, if goal 1 is the currently active goal on this trial, then the p_{11} parameter controls the extent to which an agent should prioritize navigating to the *current* goal vs. optimizing for *future goals*. This is because it controls for how long the current goal will likely persist, and conversely how imminent will be the need to visit the alternative.

We simulated this setup using both a Q-learning agent and a GR agent performing prioritized replay. Unsurprisingly, the Q-learning agent only plans how to reach the current goal, regardless of the setting of p_{11} (Fig. 5c, top-left). In contrast, the GR agent is sensitive to the environment's statistics. If $p_{11} = 0.5$ (high), it first replays a path to g_1 and then a path to g_2 (Fig. 5c, top-right). In contrast, if $p_{11} = 0.05$ (low), it replays the path to g_2 first (Fig. 5c, bottom). In general, in this model the priority (e.g., relative ordering and prominence) of replay of different possible routes, to both current and future goals, should depend parametrically on how often and how soon they are expected to be obtained.

4. Discussion

We have presented an RL account of prioritizing replay in order to build cognitive maps. The first contribution of this work is to formalize a hypothesis about how replay might be useful for building maps or routes separate from their reward value, giving the map hypothesis the same degree of formal specificity as the value hypothesis. In particular, our account distinguishes a set of candidate goal states in the environment and uses replay to learn shortest-path policies to each of them. For this, we introduced a cognitive map-like representation that we term the Geodesic Representation (GR), which learns the state-action value function from all states to each goal in a modified MDP where that goal is both terminal and rewarding. This separates "map" information (a collection of routes to possible goals, each equivalent to a Q function) from value information (which goals obtain, and how rewarding they each are), and suggests a role for replay in updating the former.

The second contribution of this work is to characterize which replay sequences would be adaptive if this were indeed the function of reply, and clarify how these predictions differ from the value view. To build replay sequences, we generalized the approach of Mattar and Daw [17], computing the expected utility of performing an experience update to the GR for any particular candidate replay. This takes advantage of the separation of map

from value by averaging per-goal expected utilities over a distribution of expected goals to yield an overall expected utility for replay. The expected utility of updates made to the GR rests in their shortening the lengths of paths from states to potential goals, rather than directly adjusting current decision variables such as Q-values in order to better harvest current rewards.

This decoupling of the current goals from the candidate future ones helps to capture a number of recent results that challenge the value view's tight coupling of replay content to choice behavior. In particular, two recent studies [32, 35] separately found that in maze navigation tasks with moving rewards, replay was systematically biased towards goal locations associated with the *previous* reward block rather than (as prioritized Q-value replay predicts) the *current* reward block. The current model captures these patterns naturally as arising from the fact that the distribution of candidate goals (hypothesized to drive replay) must be learned across multiple goal instances, across blocks or sessions, thus necessarily slower than the learning that drives within-block behavioral adjustment to each new goal.

Since the Mattar account is a special case of GR replay, the new model inherits many of the earlier model's successes: indeed, the two models coincide when goals are sparse, stable, and focused. The GR replay model also displays several new qualitative replay dynamics, related to the balancing of potential vs. current goals. These offer a range of predictions for new empirical tests. For instance, unlike replay for Q-values, replays to build a GR are predicted even before any rewards are received in an environment. Thus, for instance, our model predicts replay of paths during and after the initial unrewarded exposure to an environment in latent learning tasks [48, 27]. GR replay is also sensitive to the spatial statistics of the set of candidate goal states in the environment. Consequently, we predict that in environments with multiple potential goals, replay should focus on "central," topographically important states that are shared across the shortest paths to and between those goals. Moreover, we predict that the replay prioritization of goals, relative to each other and any currently active goals, should be modulated by their statistical properties – that is, if one goal is more common than the others, states associated with the path to it should be comparatively overweighted.

Accordingly, the model exposes two key potential avenues for future research in goal-directed navigation: what happens in the brain during latent learning of environment structure and the effect of goal dynamics on the development of cognitive map representations. Previous work has shown that

goals and routes play an important role in the population code in both the hippocampal formation and prefrontal cortex. For example, place cell remapping has been linked to the movements of goals within an environment [49] or the introduction of new goals to a familiar environment [50]. Similarly, grid cells distort their place fields upon discovery of goals in an environment [51]. Goal representations have been detected in rodent replay during awake rest in a flexible navigation task [5], as well as in human fMRI [52, 53, 54]. It remains to be seen how all these effects are modulated when the agent must arbitrate between multiple candidate choices, especially when incentives due to future options are pitted against the present reward structure.

On the theoretical level, our model offers a new perspective on the function of replay in navigation and beyond. It exposes deep but not previously obvious parallels between the value hypothesis and the map hypothesis, and in so doing addresses a high-level theoretical question in the replay literature: what does it mean for replay to build a cognitive map? We take the view that replay's role is to perform computations over memories, transforming them (here, by aggregating knowledge of local paths into plans for long-run routes) rather than simply strengthening, "consolidating" or relocating memories. In this respect, our model is spiritually connected with other views, such as complementary learning systems theory [55] and recent proposals (supported by human MEG experiments [56, 57]) suggesting that, beyond navigation, replay supports learning and remodeling of compositional schemas and structures more generally. Our teleological analysis enables us to reason about the value of replay, in terms of facilitating future reward gathering, and make precise predictions about prioritization. Although in the current model we do not yet consider non-spatial tasks or more general compositional structure, our work represents a first step in extending this type of analysis from value functions toward updating more abstract knowledge (here, maps) and points the way to extend this program in the future toward these other even more general domains.

Regarding the alternative view of replay as maintaining memory per se, our theory also provides a way to conceptualize even this as an active, prioritized computational process. In our simulations of these experiments, the GR underwent decay during each step of online behavior; this provided replay a formal role in terms of rebuilding and maintaining the GR, thereby preserving the accuracy of the agent's world model. Thus, even if the overall goal is simply to maintain a faithful representation of the local environment, there is still nontrivial computation implied in selecting which parts of that

representation are most important to maintain.

Relatedly, the analysis of replay prioritization in terms of its value (and the resulting empirical predictions about goal statistics) is a main distinction between our work and other theories of replay that are more focused on memory per se. For instance, Zhou et al. [58] recently extended a successful descriptive model of memory encoding and retrieval, the Temporal Context Model (TCM), to encompass replay, viewed as associative spreading and strengthening over associations formed during encoding. Though it has a different rationale and goals, this model makes broadly similar predictions to Mattar's and the current one; this likely relates to technical similarities owing to the fact that TCM's associations actually coincide with the SR [59, 60], which constitutes the need term in the RL-based models. Differences in the models' predictions are thus likeliest to arise for situations where replay patterns turn on gain, which quantifies the value of particular replays in serving the animal's (current or future) goals and is not naturally or directly a consideration in pure memory models (though see [61]).

Our model leaves open a number of issues that are opportunities for future theoretical work. First, as with Mattar and Daw [17], the GR replay account is not a mechanistic or process-level model of how replay is produced in the brain; instead we aim to unpack the principles driving replay, by characterizing how replay would behave, if it were optimized (through whatever process, exact or approximate) to serve the hypothesized goals. A biologically plausible implementation of geodesic replay prioritization would primarily require a tractable approximation for computing gain (which here, given our aims and following Mattar, we compute unrealistically by brute force enumeration of possible computations).

Second, our analysis is based on the GR, which we chose to expose the close algebraic relationship to Q-learning and the Mattar model. However, the spirit of our argument generalizes readily to similar map-like representations. The key feature of the GR for our purposes is that, unlike the classic SR, it is off-policy: that is, it learns paths that would be appropriate when generalizing to other goals. We have also recently explored a different SR variant [62], the Default Representation (DR), which accomplishes similar off-policy generalization and could equally serve as a target map in the replay context. Both of these representations achieve flexibility over goals by addressing a restricted setting in which goals are terminal (i.e., the map learns to plan to one goal, or choose between goals, rather than how best to visit a series of goals as in the full RL setting). To the extent this is undesir-

able, it can be addressed in another variant by maintaining a set of SRs for different policies ("generalized policy improvement" [40, 63]). Finally, both the GR prioritization and Q-value prioritization frameworks are so far only well-defined in the tabular setting. It remains an exciting opportunity to understand how to import these ideas into RL with feature-based function approximation.

5. Methods

5.1. Derivation of one-step need and gain for the GR

Our derivation of the need and gain factorization for GR EVB follows the approach of Mattar and Daw [17]. First, we describe the notation. Throughout this section, s is the current state of the agent and g is the goal state under consideration. $\bullet'$ refers to $\bullet$ after learning, except for the state s' which is simply the successor state to s. $H(s, s')$ will be denoted $H_{ss'}$ and $G(s, a, s')$ will be denoted $G_{sas'}$. Similarly, the subscript will be dropped from π_g and $\pi(a|s)$ will be denoted π_{as}.

Recall from Equation 4 the definition of the GR state-value function:

$$H_{sg} \equiv \sum_a \pi_{as} G_{sag}$$

To reach our need-gain factorization, we start by considering the expected utility of performing a Bellman backup for H with respect to a single, fixed goal g. To that end, we examine the increase in value due to performing a learning update:

$$\begin{aligned} H'_{sg} - H_{sg} &= \sum_a \pi'_{as} G'_{sag} - \pi_{as} G_{sag} \\ &= \sum_a (\pi'_{as} - \pi_{as}) G'_{sag} + (G'_{sag} - G_{sag}) \pi_{as} \end{aligned} \tag{7}$$

Now, we use the environmental dynamics to observe that:

$$G_{sag} = P(g|s, a) + \gamma \sum_{s' \neq g} P(s'|s, a) H_{s'g} \tag{8}$$

and therefore:

$$G'_{sag} - G_{sag} = \gamma \sum_{s' \neq g} P(s'|s, a) \Big(H'_{s'g} - H_{s'g} \Big)$$

Plugging that into Eq. 7, we get:

$$H'_{sg} - H_{sg} = \sum_a (\pi'_{as} - \pi_{as}) G'_{sag} + \gamma \pi_{as} \sum_{s' \neq g} P(s'|s,a) \Big(H'_{s'g} - H_{s'g} \Big)$$

Note the recursive term $H'_{s'g} - H_{s'g}$ in the right-hand side. We can iteratively unroll this recursion, yielding:

$$\begin{aligned} H'_{sg} - H_{sg} &= \sum_a (\pi'_{as} - \pi_{as}) G'_{sag} + \gamma \pi_{as} \sum_{s' \neq g} P(s'|s,a) \Big(H'_{s'g} - H_{s'g} \Big) \\ &= \sum_{x \in \mathcal{S} \setminus g} \sum_{i=0}^{\infty} \gamma^i P(s \to x, i, \pi) \sum_a (\pi'_{ax} - \pi_{ax}) G'_{xag} \end{aligned}$$

Since backups are local, $\pi'_{ax} - \pi_{ax} = 0$ for all x not equal to s_k, the start state of the backup. Thus we can simplify to:

$$H'_{sg} - H_{sg} = \begin{cases} \sum_{i=0}^{\infty} \gamma^i P(s \to s_k, i, \pi) \times \sum_a (\pi'_{as_k} - \pi_{as_k}) G'_{s_k ag} & \forall s_k \neq g \\ 0 & s_k = g \end{cases} \tag{9}$$

The case where $s_k \neq s^*$ is clearly a *need* $\times$ *gain* factorisation (first sum is a need term, second sum is a gain term). The case where $s_k = s^*$ is also such a factorisation but $gain = 0$ since there is no need to update transitions out of our goal-state with respect to our goal state (one does not need to navigate from g to g). To generalise this to a goal set of arbitrary size, we simply compute the expected value of backup as the mean goal-specific EVB averaged across the goal set under a distribution indicating their relative weights.

5.2. Elaborations on need and gain

5.2.1. Multi-step backups

We briefly note a special case of the need-gain computation. If the current step under consideration e_k is an optimal continuation of the previously replayed step e_{k-1} with respect to g, then we extend the one-step replay to a two-step replay (i.e., we update both $G(s_k, a_k, g)$ and $G(s_{k-1}, a_{k-1}, g)$). In general, if the previous sequence of replayed experiences constitutes an optimal $(n-1)$-step trajectory towards g, and e_k is an optimal continuation of that trajectory, we perform the full n-step backup. When doing this, the need is computed identically, but the gains are added across all the updated states. This has been shown to favor coherent forward replays [17].

5.2.2. Prospective need evaluation

In some scenarios, an agent may prefer to compute EVB not with respect to its current state s, but with respect to a potentially distinct set of other states (e.g., the set of starting states on the next trial) that we denote $\mathcal{S}_0$. This corresponds to the prioritization rule:

$$e^* = \arg\max_{e_k} \mathbb{E}_{g\sim P(g), s_0\sim P(s_0)}\Big[H_{post}(s_0, g) - H_{pre}(s_0, g)\Big] \tag{10}$$

where $P(s_0)$ defines a distribution over $\mathcal{S}_0$. Formally, the only change that needs to be made in order to facilitate this is to compute need in expectation over s_0:

$$\text{need}(s_k, g) = \mathbb{E}_{s_0\sim P(s_0)}\Big[\sum_{i=0}^{\infty} \gamma^i P(s_0 \to s_k, i, \pi_{g,pre})\Big] \tag{11}$$

5.2.3. Prioritization under a goal dynamics process

In Section 2, we motivated the GR by describing a scenario in which no goals are currently active, but the agent has some belief distribution about which states in the world could become active in the future. Here we describe a related, yet distinct, setup in which one goal is currently active, but the trial-by-trial evolution of goal activity is described by a Markovian transition matrix T_g (e.g., the goal dynamics process in Fig. 5).

Within such a paradigm, the per-goal EVB computation $\text{EVB}(e_k, g) \equiv H_{post}(s, g) - H_{pre}(s, g)$ does not change, but the way these are aggregated no longer involves computing the mean over a stationary goal distribution. Instead, they need to be aggregated over the dynamics process as a whole. To do this, we note that one reaps the benefits of performing a Bellman backup with respect to any goal only when that goal is active. As such, we can say that the total EVB under a given dynamics process for a fixed goal g is:

$$\text{DEVB}(e_k, g) = \sum_{t=0}^{\infty} \eta^t P(g_t = g)\text{EVB}(e_k, g)$$

where $P(g_t = g)$ is the probability that g is the active goal at trial t and $\eta \in [0, 1)$ is an episodic temporal discount factor operating at the trial-level timescale. Letting $\overrightarrow{\text{EVB}(e_k)}$ denote the vector of goal-specific EVB values for

e_k, we can compactly compute $\text{DEVB}(e_k, g)$ as:

$$\begin{aligned}\text{DEVB}(e_k) &= \overrightarrow{\text{EVB}(e_k)} \cdot \sum_{t=0}^{\infty} \eta^t P(g_t = g) \\ &= \overrightarrow{\text{EVB}(e_k)} \cdot \sum_{t=0}^{\infty} \eta^t T_g^t \vec{g_0} \\ &= \overrightarrow{\text{EVB}(e_k)} \cdot (I - \eta T_g)^{-1} \vec{g_0}\end{aligned}$$

The $(I - \eta T_g)^{-1}$ term may be thought of as the successor representation computed over the goal dynamics process.

5.3. Simulation details

We simulated a variety of "grid-world" environments – that is, deterministic environments in which an agent may move in each of the cardinal directions. We describe here structure shared across all simulations and then elaborate on each one in its respective section below.

In cases where behavior was simulated (i.e., the models of the Carey and Gillespie tasks), Q-learning agents selected actions via the softmax choice rule $\pi(a|s) \propto exp\big(Q(s,a)/\tau\big)$. In contrast, GR agents selected actions by first picking a goal to pursue (according to some independent behavioral module), and then executing the policy associated with that goal. That policy was usually a softmax policy $\pi_g(a|s) \propto exp\big(G(s,a,g)/\tau\big)$. Upon selecting action a in state s, the agent would transition to successor state s' and receive reward r (where relevant). In cases where no behavior was simulated (i.e., the asymmetric T-maze, the bottleneck/community graphs, and the prediction task), the agent simply sat at a fixed state and performed replay without selecting actions.

For replay simulation, the agent was forced to perform a fixed number of replay steps (number depending on task) prioritized by its corresponding EVB metric. If the GR or Q-value table had converged, replay was cut off to avoid nonsense replay steps being emitted. Due to the determinism of the environment, both agents updated their internal representations with learning rate $\alpha = 1$. Unless otherwise mentioned, α was used as the learning rate for both learning due to online behaviour and due to replay.

5.4. Asymmetric T-maze

In the asymmetric T-maze, Q-value and GR agents were placed at the start state (the bottom of the stem of the T), behind an impassable wall. The

GR agent was placed in a reward-free environment and assigned the terminal states at the end of each arm of the T-maze as candidate goals. The goal distribution was uniform. In contrast, the Q-learning agent was placed an an environment where the terminal states at the end of each arm of the T-maze each conferred a reward of 1/2. Both agents were simulated using a temporal discount rate $\gamma = 0.95$, though the precise value of this parameter does not noticeably affect the results.

5.4.1. Bottleneck chamber/Community graph

In the bottleneck chamber, two 5x3 chambers were connected by a 3x1 corridor. The GR agent was assigned every state as a possible start state with a uniform distribution (and so performed replay prospectively over every state in the environment). It was also assigned every state as a candidate goal state, again with a uniform distribution. The "final need" plot is the mean need, taken across all starting locations, after GR convergence for a single simulation (and so may display asymmetries associated with tie-breaking). In contrast, the "simulated replay distribution" is computed by simulating replay until convergence many times ($n = 200$) and counting across every step of replay, across every simulation, where individual replay steps are initiated.

In the community graph maze, four 2x2 chambers were connected by 1x1 corridors. Simulations and analysis were otherwise conducted as in the bottleneck chamber.

5.4.2. Modeling the Gillespie task

In our model of the Gillespie [32] task, agents were placed in the starting state of an eight-arm maze (state diagram available at Supp. Fig. A.7). On every trial, a single arm dispensed a unit reward, and would continue to do so until it was visited fifteen times; once this threshold was reached, a new rewarding arm was pseudorandomly selected from the remaining seven arms. Analysis was conducted using both GR and Q-learning agents simulated over $n_s = 200$ sessions, each composed of $n_t = 200$ trials.

Since we are largely not interested in the timestep-by-timestep evolution of the learning dynamics of each agent, and instead in how they perform replay conditioned on the arms they have visited, neither agent actually performed a timestep-by-timestep action choice process. Instead, at the beginning of every trial, each agent selected an arm to navigate to and was then handed the optimal sequence of actions to be executed in order to reach

that arm's associated goal state. During this online behaviour phase, the agent updated its internal Q-value matrix or GR in accordance with the states, actions, successor states, and rewards it observed. In practice, this does not qualitatively affect the replay dynamics emitted by either agent and simply standardizes the length of each trial (e.g., skipping the exploratory phase in which the subject may go back and forth through the arm, or running into the walls, before it realizes that such motion is not productive). Both agents selected their navigational goal arm via the softmax choice rule $\pi(\text{arm}) \propto exp\big(V(\text{arm})/\tau\big)$ implemented over per-arm values learned with the Rescorla-Wagner algorithm:

$$V_{new}(\text{chosen arm}_t) = V_{old}(\text{chosen arm}_t) + \eta\Big(R_t - V_{old}(\text{chosen arm}_t)\Big)$$

Here, we use $\eta = 1$ and $\tau = 0.3$.

After each trial, the Q-learning agent was forced to perform three replay steps (i.e., the distance from the start state of the maze to any goal state). To do this, it used the prioritization procedure from Mattar and Daw [17]. In order to incentivize replay within a block, after each time-step a weak forgetting procedure was applied to the agent's Q-values, multiplying the whole Q-matrix by a fixed factor $c_{forg} = 0.95$. The Q-learning agent performed policy updates under a softmax rule over the underlying Q-values with a separate temperature parameter $\tau_{policy} = 0.1$ (this is not important for behaviour due to the action sequence specification described earlier, but *is* important for the computation of gain which is dependent on the change in the agent's policy due to the update). Finally, we assumed that updates due to replay had a lower learning rate $\alpha_{replay} = 0.7$ than online behaviour.

The GR learning was simulated in a largely similar fashion, with some extra details due to the additional need to specify a goal distribution for replay prioritization. The per-arm behavioural value learning was identical to the Q-learning agent (i.e., softmax choice rule with $\tau = 0.3$ over values learned with $\eta = 1$ Rescorla-Wagner, constant decay of the GR every time-step with $c_{forg} = 0.95$). Furthermore, the GR agent also performed policy updates under a softmax rule with temperature parameter $\tau_{policy} = 0.1$. However, in addition to the per-arm behavioural values, the GR agent used Rescorla-Wagner to learn independent per-arm values in order to derive a replay goal distribution. We suggest that this process reflects a desire to learn the long-run statistical properties of where goals appear in the world, and as such employs a much lower learning rate than its behavioural counterpart. Here,

we used $\eta_{replay} = 0.30$ for learning the replay-associated values and a softmax rule with $\tau_{replay} = 0.20$ to convert them into probabilities. At the end of each trial, these replay-associated values themselves underwent forgetting by a constant factor equal to 0.90.

5.4.3. Modeling the Carey task

In our model of the Carey [35] task, agents were placed in the starting state of a T-maze (see Supp. Fig. A.7 for a state diagram). On every trial, the goal states associated with each arm both dispensed rewards; the magnitude of these rewards depended on the session identity, with the arm corresponding to the restricted reward modality conferring a reward of 1.5 and the other arm conferring a reward of 1. For both Q-learning and GR agents, $n_a = 10$ virtual subjects were simulated, each undergoing $n_s = 6$ sessions of alternating water/food restriction that lasted $n_t = 200$ trials.

As in our simulation of the Gillespie task, our focus is on how these agents perform replay conditioned on their previous choices, rather than their moment-by-moment behavioural dynamics. As such, each agent simply selected an arm to navigate to and was then handed the optimal sequence of actions to be executed in order to reach that arm's associated goal state. During this online behaviour phase, the agent updated its internal Q-value matrix or GR in accordance with the states, actions, successor states, and rewards it observed. Both agents chose which arm to navigate to via the same softmax choice rule outlined in the previous subsection. For these simulations, we used the parameters $\eta = 1$ and $\tau = 0.5$.

After each trial, both the Q-learning and GR agents performed prioritized replay as outlined in the previous subsection. The Q-learning agent underwent forgetting with $c_{forg} = 0.75$. Its policy updates were performed under a softmax regime with $\tau_{policy} = 0.1$. The learning rate for replay was assumed to be lower than for online behaviour, with $\alpha_{replay} = 0.7$. The GR agent had the same values for these parameters. Furthermore, it had a replay-value learning rate of $\eta_{replay} = 0.1$, a softmax parameter $\tau_{replay} = 0.15$ for converting those values into a goal distribution, and no forgetting on the replay-associated arm values.

5.4.4. Modeling the prediction task

In our implementation of the prediction task, we simulated Q-learning and GR agents on the maze in Fig. 5 using the state diagram provided in

Supp. Fig. A.7. All agents began with their respective representation initialized at zero and performed replay until convergence. Both agents learned with a learning rate of $\alpha = 1$, and assumed a highly exploitative softmax behavioural policy with $\tau = 0.01$. The GR agent was assumed to know the true goal dynamics process and performed prioritized replay as described in Section 5.2.3, with an episodic discount factor equal to 0.9.

5.5. Data replotting

Replotting of data from Gillespie et al. [32] and Carey et al. [35] was performed by annotating the individual data points using WebPlotDigitizer 4.6 and then averaging as necessary.

Appendix A. Supplement

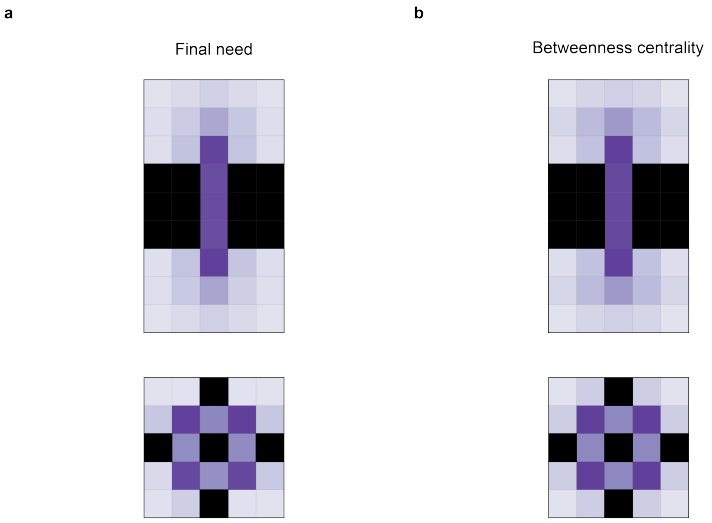

Figure A.6: **Need metric captures important elements of environment topology.** **(a)** Examples of final need, reproduced from Fig. 2. **(b)** Betweenness-centrality computed for the bottleneck chamber (top) and the community graph maze (bottom).

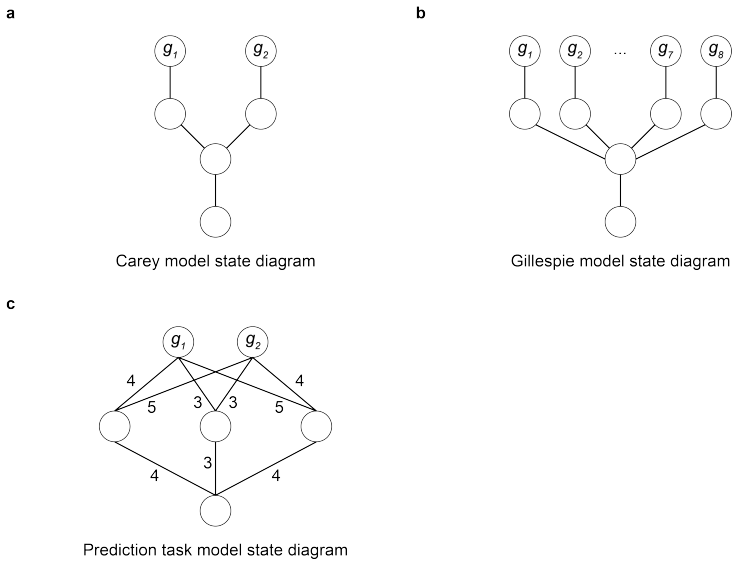

Figure A.7: **State diagrams for the simulations in Sections 3.2 and 3.3. (a)** State diagram for our model of Carey et al. [35] **(b)** State diagram for our model of Gillespie et al. [32] (c) State diagram for our model of the prediction task in Section 3.3. Numbers indicate distances (e.g., the left bottleneck state requires four steps to reach g_1), which are implemented through intermediate states (not shown).

References

[1] J. O'Keefe, L. Nadel, The hippocampus as a cognitive map, Clarendon Press ; Oxford University Press, Oxford : New York, 1978.

[2] T. J. Davidson, F. Kloosterman, M. A. Wilson, Hippocampal Replay of Extended Experience, Neuron 63 (4) (2009) 497–507. doi:10.1016/j.neuron.2009.07.027.

[3] K. Diba, G. Buzsáki, Forward and reverse hippocampal place-cell sequences during ripples, Nature Neuroscience 10 (10) (2007) 1241–1242. doi:10.1038/nn1961.

[4] M. F. Carr, S. P. Jadhav, L. M. Frank, Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval, Nature Neuroscience 14 (2) (2011) 147–153. doi:10.1038/nn.2732.

[5] B. E. Pfeiffer, D. J. Foster, Hippocampal place-cell sequences depict future paths to remembered goals, Nature 497 (7447) (2013) 74–79. doi:10.1038/nature12112.

[6] D. J. Foster, M. A. Wilson, Reverse replay of behavioural sequences in hippocampal place cells during the awake state, Nature 440 (7084) (2006) 680–683. doi:10.1038/nature04587.

[7] D. R. Euston, M. Tatsuno, B. L. McNaughton, Fast-Forward Playback of Recent Memory Sequences in Prefrontal Cortex During Sleep, Science 318 (5853) (2007) 1147–1150. doi:10.1126/science.1148979.

[8] A. Peyrache, M. Khamassi, K. Benchenane, S. I. Wiener, F. P. Battaglia, Replay of rule-learning related neural patterns in the prefrontal cortex during sleep, Nature Neuroscience 12 (7) (2009) 919–926. doi:10.1038/nn.2337.

[9] K. Kaefer, M. Nardin, K. Blahna, J. Csicsvari, Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching, Neuron 106 (1) (2020) 154–165.e6. doi:10.1016/j.neuron.2020.01.015.

[10] A. S. Gupta, M. A. van der Meer, D. S. Touretzky, A. D. Redish, Hippocampal Replay Is Not a Simple Function of Experience, Neuron 65 (5) (2010) 695–705. doi:10.1016/j.neuron.2010.01.034.

[11] J. Widloski, D. J. Foster, Flexible rerouting of hippocampal replay sequences around changing barriers in the absence of global place field remapping, Neuron 110 (9) (2022) 1547–1558.e8. doi:10.1016/j.neuron.2022.02.002.

[12] D. J. Foster, J. J. Knierim, Sequence learning and the role of the hippocampus in rodent navigation, Current Opinion in Neurobiology 22 (2) (2012) 294–300. doi:10.1016/j.conb.2011.12.005.

[13] X. Wu, D. J. Foster, Hippocampal Replay Captures the Unique Topological Structure of a Novel Environment, Journal of Neuroscience 34 (19) (2014) 6459–6469. doi:10.1523/JNEUROSCI.3414-13.2014.

[14] G. de Lavilléon, M. M. Lacroix, L. Rondi-Reig, K. Benchenane, Explicit memory creation during sleep demonstrates a causal role of place cells in navigation, Nature Neuroscience 18 (4) (2015) 493–495. doi:10.1038/nn.3970.

[15] I. Gridchyn, P. Schoenenberger, J. O'Neill, J. Csicsvari, Assembly-Specific Disruption of Hippocampal Replay Leads to Selective Memory Deficit, Neuron 106 (2) (2020) 291–300.e6. doi:10.1016/j.neuron.2020.01.021.

[16] K. Kay, J. E. Chung, M. Sosa, J. S. Schor, M. P. Karlsson, M. C. Larkin, D. F. Liu, L. M. Frank, Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus, Cell (2020) S0092867420300611doi:10.1016/j.cell.2020.01.014.

[17] M. G. Mattar, N. D. Daw, Prioritized memory access explains planning and hippocampal replay, Nature Neuroscience 21 (11) (2018) 1609–1617. doi:10.1038/s41593-018-0232-z.

[18] C. S. Lansink, P. M. Goltstein, J. V. Lankelma, B. L. McNaughton, C. M. A. Pennartz, Hippocampus Leads Ventral Striatum in Replay of Place-Reward Information, PLoS Biology 7 (8) (2009) e1000173. doi:10.1371/journal.pbio.1000173.

[19] A. C. Singer, L. M. Frank, Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus, Neuron 64 (6) (2009) 910–921. doi:10.1016/j.neuron.2009.11.016.

[20] S. N. Gomperts, F. Kloosterman, M. A. Wilson, VTA neurons coordinate with the hippocampal reactivation of spatial experience, eLife 4 (2015) e05360. doi:10.7554/eLife.05360.

[21] H. F. Ólafsdóttir, C. Barry, A. B. Saleem, D. Hassabis, H. J. Spiers, Hippocampal place cells construct reward related sequences through unexplored space, eLife 4 (2015) e06063. doi:10.7554/eLife.06063.

[22] R. E. Ambrose, B. E. Pfeiffer, D. J. Foster, Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward, Neuron 91 (5) (2016) 1124–1136. doi:10.1016/j.neuron.2016.07.047.

[23] M. Gruber, M. Ritchey, S.-F. Wang, M. Doss, C. Ranganath, Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events, Neuron 89 (5) (2016) 1110–1120. doi:10.1016/j.neuron.2016.01.017.

[24] M. C. Zielinski, W. Tang, S. P. Jadhav, The role of replay and theta sequences in mediating hippocampal-prefrontal interactions for memory and cognition, Hippocampus 30 (1) (2020) 60–72. doi:10.1002/hipo.22821.

[25] A. Singer, M. Carr, M. Karlsson, L. Frank, Hippocampal SWR Activity Predicts Correct Decisions during the Initial Learning of an Alternation Task, Neuron 77 (6) (2013) 1163–1173. doi:10.1016/j.neuron.2013.01.027.

[26] Y. Liu, M. G. Mattar, T. E. J. Behrens, N. D. Daw, R. J. Dolan, Experience replay is associated with efficient nonlocal learning, Science 372 (6544) (2021) eabf1357. doi:10.1126/science.abf1357.

[27] E. C. Tolman, Cognitive maps in rats and men., Psychological Review 55 (4) (1948) 189–208, place: US Publisher: American Psychological Association. doi:10.1037/h0061626.

[28] A. D. Grosmark, F. T. Sparks, M. J. Davis, A. Losonczy, Reactivation predicts the consolidation of unbiased long-term cognitive maps, Nature Neuroscience 24 (11) (2021) 1574–1585. doi:10.1038/s41593-021-00920-7.

[29] E. I. Moser, E. Kropff, M.-B. Moser, Place Cells, Grid Cells, and the Brain's Spatial Representation System, Annual Review of Neuroscience 31 (1) (2008) 69–89. doi:10.1146/annurev.neuro.31.061307.090723.

[30] L. Roux, B. Hu, R. Eichler, E. Stark, G. Buzsáki, Sharp wave ripples during learning stabilize the hippocampal spatial map, Nature Neuroscience 20 (6) (2017) 845–853. doi:10.1038/nn.4543.

[31] D. Schiller, H. Eichenbaum, E. A. Buffalo, L. Davachi, D. J. Foster, S. Leutgeb, C. Ranganath, Memory and Space: Towards an Understanding of the Cognitive Map, The Journal of Neuroscience 35 (41) (2015) 13904–13911. doi:10.1523/JNEUROSCI.2618-15.2015.

[32] A. K. Gillespie, D. A. Astudillo Maya, E. L. Denovellis, D. F. Liu, D. B. Kastner, M. E. Coulter, D. K. Roumis, U. T. Eden, L. M. Frank, Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice, Neuron 109 (19) (2021) 3149–3163.e6. doi:10.1016/j.neuron.2021.07.029.

[33] R. S. Sutton, Dyna, an integrated architecture for learning, planning, and reacting, ACM SIGART Bulletin 2 (4) (1991) 160–163. doi:10.1145/122344.122377.

[34] T. A. Krausz, A. E. Comrie, A. E. Kahn, L. M. Frank, N. D. Daw, J. D. Berke, Dual credit assignment processes underlie dopamine signals in a complex spatial environment, Neuron 111 (21) (2023) 3465–3478.e7. doi:10.1016/j.neuron.2023.07.017.

[35] A. A. Carey, Y. Tanaka, M. A. A. van der Meer, Reward revaluation biases hippocampal replay content away from the preferred outcome, Nature Neuroscience 22 (9) (2019) 1450–1459. doi:10.1038/s41593-019-0464-6.

[36] N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nature Neuroscience 8 (12) (2005) 1704–1711. doi:10.1038/nn1560.

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

[38] E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, Predictive representations can link model-based reinforcement learning to model-free mechanisms, PLOS Computational Biology 13 (9) (2017) e1005768. doi:10.1371/journal.pcbi.1005768.

[39] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, S. J. Gershman, The successor representation in human reinforcement learning, Nature Human Behaviour 1 (9) (2017) 680–692. doi:10.1038/s41562-017-0180-8.

[40] A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. van Hasselt, D. Silver, Successor Features for Transfer in Reinforcement Learning, arXiv:1606.05312 [cs]ArXiv: 1606.05312 (Apr. 2018).

[41] L. P. Kaelbling, Learning to Achieve Goals (1993) 5.

[42] M. P. H. Gardner, G. Schoenbaum, S. J. Gershman, Rethinking dopamine as generalized prediction error (2018) 10.

[43] A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, M. M. Botvinick, Neural representations of events arise from temporal community structure, Nature Neuroscience 16 (4) (2013) 486–492. doi:10.1038/nn.3331.

[44] A. C. Schapiro, N. B. Turk-Browne, K. A. Norman, M. M. Botvinick, Statistical learning of temporal community structure in the hippocampus: *STATISTICAL LEARNING OF TEMPORAL COMMUNITY STRUCTURE*, Hippocampus 26 (1) (2016) 3–8. doi:10.1002/hipo.22523.

[45] S. Kakade, P. Dayan, Acquisition and extinction in autoshaping., Psychological review 109 (3) (2002) 533.

[46] T. E. Behrens, M. W. Woolrich, M. E. Walton, M. F. Rushworth, Learning the value of information in an uncertain world, Nature neuroscience 10 (9) (2007) 1214–1221.

[47] Y. Niv, D. Joel, P. Dayan, A normative perspective on motivation, Trends in Cognitive Sciences 10 (8) (2006) 375–381. doi:10.1016/j.tics.2006.06.010.

[48] H. C. Blodgett, The effect of the introduction of reward upon the maze performance of rats., University of California Publications in Psychology 4 (1929) 113–134.

[49] H. Xu, P. Baracskay, J. O'Neill, J. Csicsvari, Assembly Responses of Hippocampal CA1 Place Cells Predict Learned Behavior in Goal-Directed Spatial Tasks on the Radial Eight-Arm Maze, Neuron 101 (1) (2019) 119–132.e4. doi:10.1016/j.neuron.2018.11.015.

[50] D. Dupret, J. O'Neill, B. Pleydell-Bouverie, J. Csicsvari, The reorganization and reactivation of hippocampal maps predict spatial memory performance, Nature Neuroscience 13 (8) (2010) 995–1002. doi:10.1038/nn.2599.

[51] C. N. Boccara, M. Nardin, F. Stella, J. O'Neill, J. Csicsvari, The entorhinal cognitive map is attracted to goals, Science 363 (6434) (2019) 1443–1447. doi:10.1126/science.aav4837.

[52] A. D. Ekstrom, M. J. Kahana, J. B. Caplan, T. A. Fields, E. A. Isham, E. L. Newman, I. Fried, Cellular networks underlying human spatial navigation, Nature 425 (6954) (2003) 184–188. doi:10.1038/nature01964.

[53] T. I. Brown, V. A. Carr, K. F. LaRocque, S. E. Favila, A. M. Gordon, B. Bowles, J. N. Bailenson, A. D. Wagner, Prospective representation of navigational goals in the human hippocampus, Science 352 (6291) (2016) 1323–1326. doi:10.1126/science.aaf0784.

[54] L. Howard, A. Javadi, Y. Yu, R. Mill, L. Morrison, R. Knight, M. Loftus, L. Staskute, H. Spiers, The Hippocampus and Entorhinal Cortex Encode the Path and Euclidean Distances to Goals during Navigation, Current Biology 24 (12) (2014) 1331–1340. doi:10.1016/j.cub.2014.05.001.

[55] J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, 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) (1995) 419–457, place: US Publisher: American Psychological Association. doi:10.1037/0033-295X.102.3.419.

[56] P. Schwartenbeck, A. Baram, Y. Liu, S. Mark, T. Muller, R. Dolan, M. Botvinick, Z. Kurth-Nelson, T. Behrens, Generative replay underlies compositional inference in the hippocampal-prefrontal circuit, Cell (2023) S0092867423010255doi:10.1016/j.cell.2023.09.004.

[57] Z. Kurth-Nelson, T. Behrens, G. Wayne, K. Miller, L. Luettgau, R. Dolan, Y. Liu, P. Schwartenbeck, Replay and compositional computation, Neuron 111 (4) (2023) 454–469. doi:10.1016/j.neuron.2022.12.028.

[58] Z. Zhou, M. J. Kahana, A. C. Schapiro, Replay as context-driven memory reactivation, preprint, Neuroscience (Mar. 2023). doi:10.1101/2023.03.22.533833.

[59] S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, P. B. Sederberg, The Successor Representation and Temporal Context, Neural Computation 24 (6) (2012) 1553–1568. doi:10.1162/NECO_a_00282.

[60] C. Y. Zhou, D. Talmi, N. Daw, M. G. Mattar, Episodic retrieval for model-based evaluation in sequential decision tasks (2023).

[61] D. Talmi, L. J. Lohnas, N. D. Daw, A retrieved context model of the emotional modulation of memory., Psychological Review 126 (4) (2019) 455–485, place: US Publisher: American Psychological Association. doi:10.1037/rev0000132.

[62] P. Piray, N. D. Daw, Linear reinforcement learning in planning, grid fields, and cognitive control, Nature Communications 12 (1) (2021) 4942. doi:10.1038/s41467-021-25123-3.

[63] A. Barreto, S. Hou, D. Borsa, D. Silver, D. Precup, Fast reinforcement learning with generalized policy updates, Proceedings of the National Academy of Sciences 117 (48) (2020) 30079–30087. doi:10.1073/pnas.1907370117.