Neuron

Between planning and map building: Prioritizing replay when future goals are uncertain

Highlights

- A planning RL model of hippocampal replay is generalized for cognitive map building
- The model explains challenging "map-like" replay from slow goal-switching tasks
- The model also accounts for "planning-like" replay from fast goal-switching tasks
- The model unifies "map" and "planning" replay in terms of environment statistics

Authors

Yotam Sagiv, Thomas Akam, Ilana B. Witten, Nathaniel D. Daw

Correspondence

ysagiv@princeton.edu (Y.S.), ndaw@princeton.edu (N.D.D.)

In brief

Sagiv et al. develop a reinforcement learning model of hippocampal replay for cognitive map building. Their model explains recent challenging data that stymie a popular theoretical account while keeping its strengths. They lay the foundation for unifying the two dominant hypotheses of replay function (planning and map building) under the same umbrella.

Sagiv et al., 2025, Neuron *113*, 4278–4292
December 17, 2025
https://doi.org/10.1016/j.neuron.2025.09.021

CellPress

CellPress

Neuron

Article

Between planning and map building: Prioritizing replay when future goals are uncertain

Yotam Sagiv,[1,*] Thomas Akam,[2,3] Ilana B. Witten,[1,4] and Nathaniel D. Daw[1,5,6,*]
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2]Department of Experimental Psychology, University of Oxford, Oxford OX1 3EL, UK
[3]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London W1T 4JG, UK
[4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[5]Department of Psychology, Princeton University, Princeton, NJ 08544, USA
[6]Lead contact
*Correspondence: ysagiv@princeton.edu (Y.S.), ndaw@princeton.edu (N.D.D.)
https://doi.org/10.1016/j.neuron.2025.09.021

SUMMARY

Despite many empirical results about hippocampal replay, its computational function remains controversial. The "value" hypothesis contends that replay plans routes to current goals, while the "map" hypothesis holds that replay builds an abstract environmental representation, distinct from immediate goals. Data appear to support either view, though the planning hypothesis is particularly challenged by recent observations of replay lagging, rather than leading, animals learning to reach current goals. However, differentiating these ideas is difficult due to a lack of formal specificity, especially about the map hypothesis. We address these gaps by extending a prominent theory of planning to include routes to future as well as current goals: effectively a map. Whether replay prefers current goals, like planning, or others, like maps, then depends on their estimated likelihood of future relevance. This account reconciles both views with one another and with much data, revealing a deep relationship between the seemingly distinct hypotheses.

INTRODUCTION

Much attention has been paid to experience replay as a mechanism subserving learning and complex behavior. In particular, replay of nonlocal trajectories during sharp-wave ripples in hippocampal place cells[1–6] (and similar events elsewhere[7–9]) tantalizingly suggests a navigation-related computation.[5,10–16] However, replay's precise functional role—what it actually computes—remains a central question.

There are, broadly, two schools of thought: the "value hypothesis" suggests that replay facilitates planning or credit assignment to directly guide current or future choices.[3,5,16–26] Under this view, trajectory replay facilitates connecting candidate actions at some location with their rewarding consequences elsewhere in space (e.g., by updating a decision variable such as the value function). By contrast, the "map hypothesis" argues that replay builds (or maintains) some abstract representation of the environment (e.g., a "cognitive map" of its layout) and is not straightforwardly tied to subsequent behavior or reward.[1,4,27–32] Both interpretations have been argued to be consistent with data, though their differential predictions are often not obvious.

Recent theoretical work suggested an approach for improving the testability of these 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 replay event. This reasoning implies testable claims about how each replay should affect subsequent choices (by propagating reward information to distal choice points).[26,34] Furthermore, they argued that, given such a hypothesis about the effects of a replay on choices, one can derive a corresponding formal hypothesis about the *prioritization* of replayed trajectories. In particular, if this were the function of replay, then the brain would be expected to favor trajectories that maximize expected reward by best improving choices. This yields verifiable predictions (e.g., about the statistics of forward vs. reverse replay in different situations) that match a range of data.

Sharper empirical claims in turn permit falsification and refinement. Multiple authors using goal-switching tasks (notably Gillespie et al.[32] and Carey et al.[35]) subsequently reported that replayed trajectories tend to be systematically focused on past goals rather than current ones and thus lag rather than lead animals learning updated choice behavior. Such "paradoxical"[36] decoupling between behavioral learning and replay seems to disfavor the value hypothesis, which predicts that these quantities should track each other, and instead to support the map hypothesis. Here, we aim to explain these results by extending Mattar's approach to encompass the map hypothesis.

Indeed, although the value/map division appears intuitive, a challenge is that the map hypothesis remains incompletely

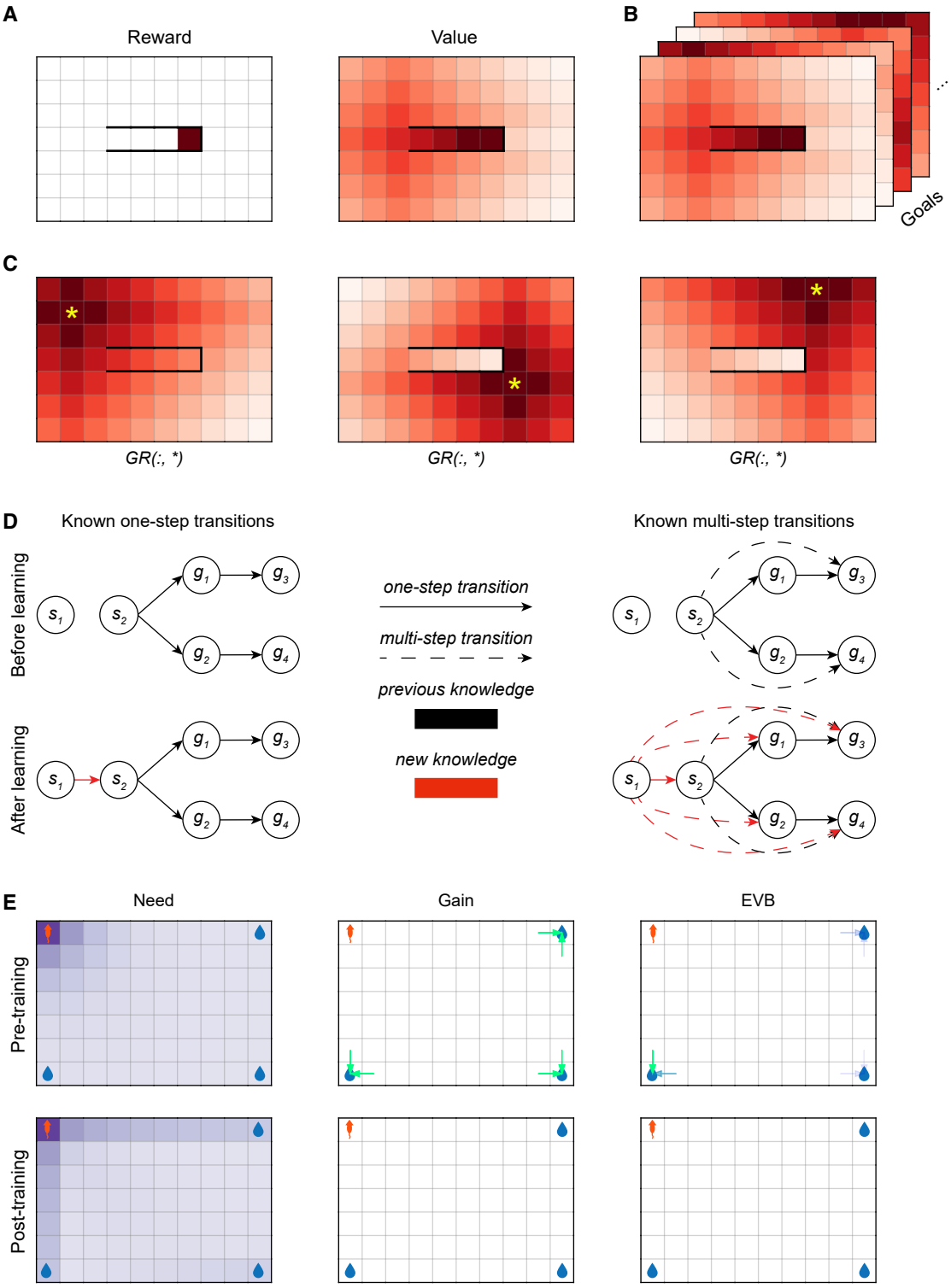

Figure 1. The GR

(A) Left: an open field with a corridor that encloses a single rewarded state. Right: the corresponding state value function.

(B) The GR is a stack of state-action value functions, one for each goal. For simplicity, we illustrate the value functions over states rather than over state-action pairs.

(C) Three different slices of the GR in the same environment. A yellow asterisk indicates the slice's corresponding goal state.

(legend continued on next page)

specified. Whereas RL theories offer a precise formalization of the value hypothesis, there has been less formal attention to the map hypothesis, starting with the question of what the "map" is. Outside the replay context, RL models generally operationalize the map as the environment's local connectivity (its "one-step" adjacency graph).[37] However, it seems unlikely that replay builds this representation, because replayed trajectories already reflect local connectivity, even immediately after encountering novel barriers.[11] Whatever the map is, a second open question is how replay builds it.

Our new account addresses these questions by extending Mattar's logic to a setting where the locations of rewards are unknown or dynamic. The analog of the value function (the target in the Mattar model) here is a *set* of value functions: one for each potential goal. This shifts computation from a narrow focus on the animal's current goals toward a more generally useful representation, in line with the map hypothesis. This set of value functions encodes a set of long-run routes toward many possible goals, similar to a successor representation (SR)[38] (though our variant is "off-policy" or insensitive to the animal's goals during learning).

This SR-like representation, the geodesic representation (GR), captures a formal notion of a cognitive map.[39,40] That a map can be viewed as a generalized value function clarifies many issues. First, it goes beyond local information (adjacency and barriers): it measures long-run distance from each location to each goal, a nontrivial quantity that can be computed by trajectory replay, generalizing replay's putative role in building a single value function (Dyna-SR[41,42]). Such a map permits flexibly updating choice policies if goals change—without additional learning or computation. This connects the model to a longstanding empirical tradition that investigates animals' use of cognitive maps through goal-switching tasks such as latent learning or revaluation.[27,37,41] The same feature—that replay, if it builds a GR, can help reach future goals—also underlies a generalization of Mattar's priority metric to quantify which replays are most useful. Here, the value of replay in helping to reach current goals is combined with that for potential future goals, weighted according to learned beliefs about which goals are more likely.

Because this account generalizes Mattar's, it reconciles the value and map views. Both are recast as extreme cases of a spectrum of replay rules for settings where the animal expects different goals to be relevant, ranging from a focus on a single, current goal to a range of possible future goals. We show that this explains why paradoxical replay arises in tasks with periodic goal switching, where particular potential goals weigh heavily given the animals' experiences.[32,35] At the same time, since value replay arises as a special case of map replay when behavioral goals are more focused, the model explains why evidence for planning-focused replay has been reported in other settings where animals need to find many routes to a more stable, predominant goal.[5,11,43] The model suggests that these different replay regimes depend on the animal's beliefs about goal statistics and makes new predictions for experiments manipulating this factor.

RESULTS

The model

The GR

In RL models, a cognitive map (or "world model") is traditionally associated with the one-step transition function $T(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t)$ that captures how local actions a (directional steps) affect the current state s (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 takes substantial computation to find the optimal actions, e.g., by computing the long-run aggregate rewards resulting from different candidate actions. Formally, this is the state-action value $Q(s,a) = \mathbb{E}_{s' \sim P(s'|s,a)}[R(s') + \gamma \max_{a'} Q(s', a')]$. Previous theories[17] suggest that a goal of replay is to facilitate computing Q by aggregating reward over replayed trajectories. The value function is goal-specific: if the one-step rewards change (for example, if a reward moves), it must be recomputed. Thus, Q-learning (and similar methods) requires additional computation when goals change.

One way to address this limitation is to represent a map in terms of distances from start states to many goal states. Stated differently, instead of a single value function, an agent may maintain a set of value functions for many different rewards. One version of this idea is the SR.[38,44] We introduce a variant of it, the GR, inspired by Kaelbling,[45] 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 represents the shortest paths from each state in the environment to possible "goal" states (all other states or a subset).

Consider an episodic task in an environment with a single terminal state g containing a 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 (the terminal value 1 discounted by the steps to reach it; Figure 1A). Consequently, the optimal policy in this task maximizes return by minimizing this distance.

Accordingly, we define the GR as a stack of these Q value tables (Figure 1B), with each "page" in the stack encoding the state values in a different modified version of the underlying environment where the corresponding goal is assumed to be the only rewarding state (conferring a unit reward) and terminal. Note: these candidate goal states need not actually be terminal or rewarding in the task; these properties define a set of surrogate problems of finding shortest paths to each counterfactual goal. 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

(D) An agent learns how to reach several goals via a single learning step from s_1 to s_2. Black arrows: knowledge prior to learning; red arrows: knowledge gained after transitioning from s_1 to s_2; solid lines: one-step transitions; dashed lines: implied longer-horizon connections. Top row: agent's knowledge about the environment's one-step (left) and multi-step (right) transition structure before learning; bottom row: same, after learning.
(E) Visualizations of need, gain, and EVB in an open field. The agent starts in the top left corner with candidate goal locations in the other three corners. Top row: before replay; bottom row: after GR convergence. In the gain and EVB plots, arrow color and opacity indicate the value of the relevant metric.

associated Markov Decision Process (MDP). Letting $G(s,a,g)$ represent the "value" of action a in state s on page g of the GR, we define it as:

$$G(s,a,g) \equiv \mathbb{E}_{\pi_g}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{I}_{s_t=g}\middle| s_0 = s,\ a_0 = a\right], \qquad \text{(Equation 1)}$$

where g is any potential goal state, $\gamma \in [0,1)$ is a temporal discount rate, π_g is the optimal policy for reaching g (i.e., is reward-maximizing in the MDP where g is terminal and is the only rewarding state), and $\mathbb{I}_\bullet$ is the indicator function that is 1 if $\bullet$ is true and 0 otherwise. This definition encodes the above intuition: $G(s,a,g)$ 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 GR slices are illustrated in Figure 1C.

The GR may also be characterized by its Bellman equation:

$$G(s,a,g) = \mathbb{E}_{s'\sim P(s'|s,a)}\left[\mathbb{I}_{s'=g} + \gamma \mathbb{I}_{s'\neq g}\max_{a'} G(s',a',g)\right] \qquad \text{(Equation 2)}$$

Intuitively, if s' is the goal state g, then transitioning there accrues reward 1. Otherwise, the current value is γ times the value of taking the best available action in s'. This can be rewritten in vector form:

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

where $\odot$ denotes the elementwise product, $\mathbf{1}_{s'}$ is a one-hot vector at s', $\mathbf{0}_{s'}$ is 0 at s' and 1 everywhere else, and the max is taken separately for each goal state. This form of the equation demonstrates that distances to multiple goals can be updated through one learning step (e.g., via a vector of off-policy temporal-difference updates, one for each goal, as has been proposed for the SR[38,46]). Consider the setup in Figure 1D, where an agent knows how to get to goals $g_1, \ldots, g_4$ from state s_2 but not s_1. If the agent underwent the transition from $s_1 \rightarrow s_2$, they would learn, in a single step, how to reach all the goals already reachable from s_2.

With this learning rule, the GR itself is updated based on observing a state-action-successor tuple (s, a, s'). The off-policy nature of the update (a set of off-policy Q-learning updates, one for each candidate goal) means that the GR is not sensitive to either the exploratory policy or goals during learning. Thus, once a GR is learned, an agent can adapt to a new goal (e.g., the reward moved from one location to another) simply by switching which page $G(:,:,g)$ controls behavior. Such nimble switching is unlike Q-learning, and it implies that replay can have utility due to "pre-planning" to reach potential future goals. That is, replay can improve the agent's later choices even when goals have changed, earning additional reward.

Finally, the GR with a single goal is equivalent to Q-learning with a single, terminal reward. Thus the new theory generalizes its predecessor: from one goal to several mutually exclusive candidate goals, which may be available at different times. Although we concentrate on this case, the GR can easily be adapted to a more general class of reward functions: those containing multiple, simultaneously available terminal rewards of different magnitudes.

Prioritizing replay on the GR

Previous work[17] addressed prioritizing experience replay for Q value updating. To extend a theory of replay's function to a theory of replay content, Mattar proposed that replaying a state-action-state event (thus performing a Q value update there) should be prioritized according to its expected utility, i.e., the difference between the agent's expected return after vs. before the update. Replay can increase return if the update improves future choices. This "expected value of backup" (EVB) can be decomposed as a product of two terms, "need" and "gain." Need measures relevance: how often the agent expects to be in the updated state, while gain measures the change in that state's value (i.e., how much additional reward the agent expects to accrue if it visits the state, due to the update improving choice there). Different replay patterns then arise due to the balance between need and gain at different locations.

Since the GR aggregates a set of Q functions, EVB for it can also be decomposed into the product of need and gain terms. Generalizing Mattar's approach,[17] we first define an analog of the state value function for any goal g in the current setting:

$$H(s,g) \equiv \sum_a \pi_g(a|s) G(s,a,g). \qquad \text{(Equation 4)}$$

Here, π_g is the policy that tries to reach g. The expected improvement in H after backing up the experience $e_k = (s_k, a_k, s'_k)$ with respect to g factorizes (see STAR Methods):

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

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 via the update from $\bullet_{pre}$ to $\bullet_{post}$). $P(s \rightarrow s_k, i, \pi_{g,pre})$ 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 in Equation 5 measures how often the agent will reach the state being updated, s_k, given its current state s and its policy (precisely, it is the SR evaluated under π_g). Gain quantifies how much additional reward the agent should accumulate due to a change in policy due to the performed update. In summary, 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 (1) how relevant that experience is and (2) by the magnitude of the change induced by the update (Figure 1E).

So far, we have followed Mattar's definition of EVB for one value function. Since the GR comprises a set of value functions for multiple goals, and replay of an experience (via Equation 3) updates all of them, we need to aggregate their value into an overall EVB. We simply take the expectation (or more generally the expected discounted sum) of these per-goal EVBs under a distribution over goals: the agent's beliefs about which ones

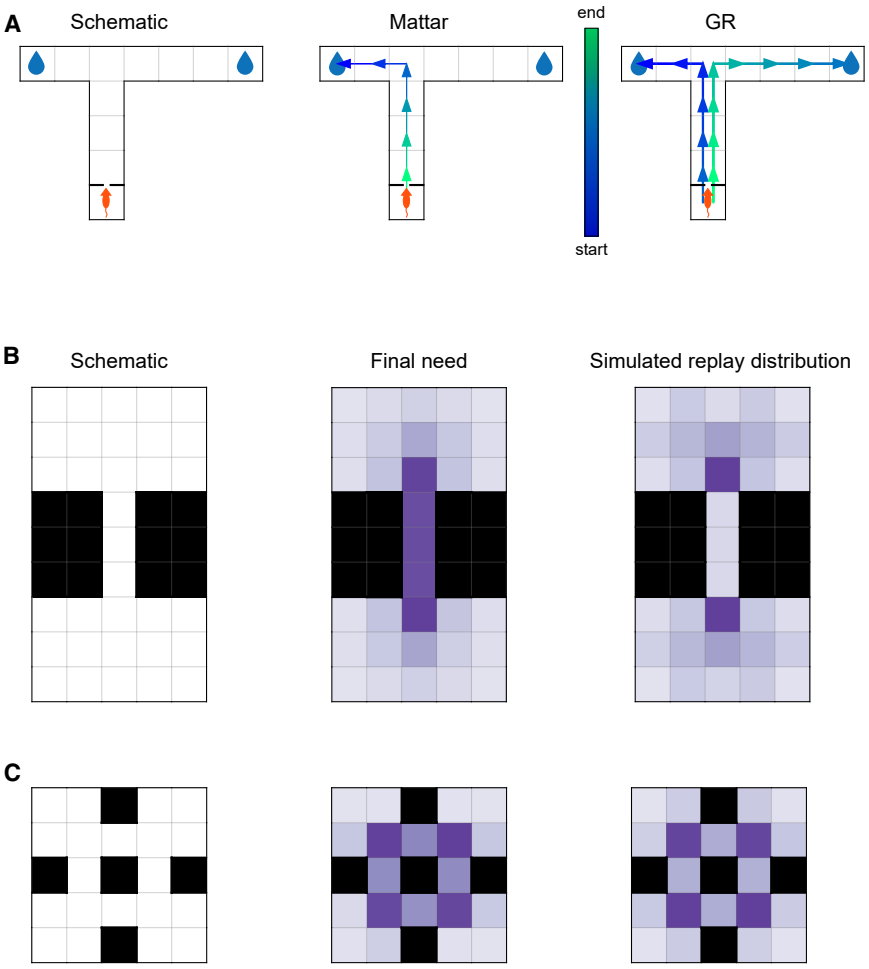

Figure 2. The GR supports replay to multiple goals and respects environmental structure

(A) In an asymmetric T-maze, 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 toward 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 analog of the community graph. Left: schematic; middle: asymptotic need; right: mean state replay across $n = 250$ simulations.

are likely to be relevant in the future. How this distribution is learned or constructed is an interesting question. We return to it later for different experimental tasks, since in practice such expectations will depend on animals' experiences. In this article, we make simple, task-dependent assumptions about these expectations to focus on the effects of goal uncertainty on replay. We envision these hand-constructed examples as standing in for a hierarchical learning account, in which animals learn a model of goal-switching dynamics from past experience with any particular task.[47–50] Regardless, an agent may then prioritize replay by picking the memory that maximizes the expected improvement from the current state s averaged over goals:

$$e^* = \underset{e_k}{\text{argmax}}\, \mathbb{E}_{g\sim P(g)}\left[H_{post}(s,g) - H_{pre}(s,g)\right]. \quad \text{(Equation 6)}$$

Once an experience is selected for replay, the GR is updated for *all* goals toward the target in Equation 3.

Simulation results

GR replay favors elements of routes shared between multiple goals

Since prioritized GR replay generalizes Mattar's account,[17] it retains its notable properties (e.g., coherent forward and reverse sequences and spatial initiation biases). Accordingly, we focus on exploring novel properties of GR replay, which would not be attributable to Q value replay for a single goal.

First, the GR allows for simultaneous planning across multiple candidate goals. To expose this distinction, we first consider a stylized situation. In Figure 2A, we simulated prioritized replay, using both Q-learning and GR agents, on an asymmetric T-maze in which the end of each arm contains a reward (and is a candidate goal), but one arm is shorter. We assume for simplicity that the reward states are terminal, there is a single, known start state, and the maze is viewed without online exploration, such as through a window.[21] The Q-learning agent (Figure 2A, middle) replays a path from the closest reward location to the starting location and then stops. By contrast, the GR agent (Figure 2A, right) replays paths from both the close and far goal locations to the start, in order of distance. This distinction illustrates that the Q-learning agent's objective is to build a reward-optimal policy—so it needs to learn only to reach the nearest 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 are both candidate goals and potential starting locations. Here, the target GR is an all-to-all map of shortest paths. One hallmark of GR-based replay is that, because priority is averaged over goals, replays are favored through locations that are shared between optimal routes between different starts and goals. For example, replay should be focused around bottleneck states, through which many optimal routes pass. We simulated GR replay while learning two environments with bottleneck states: a chamber with two large rooms connected by a narrow corridor (Figure 2B) and a four-room environment (based on Schapiro et al.'s[51,52] community graph) in which

each room can only be entered or exited via a single location (Figure 2C).

Recall that replay priority is the product of need and gain terms. Preference for bottleneck states arises algebraically from the need term, as can be seen in the middle figures, which plot need under the converged GR. In both graphs, the bottleneck states have the highest need since they are required for all paths 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. See Figure S1).

The analysis so far neglects the contribution of gain and of the progression of learning. These reflect the partly opposing contribution of another 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 (similarly, the door states in the community graph) are replayed relatively less when compared with their BC values. This is because (if the agent first learns to come and go from the exit state to other states in a room) multiple paths between the rooms can be bridged by one replay through the bottleneck. Stated differently, since a single GR update facilitates learning across many goals simultaneously, it is adaptive to learn about paths to goals within a given room, transfer that knowledge to the mouth of the bottleneck, and then carry it to the next room in a single replay sequence. Thus the model tends to favor the endpoints of bottlenecks, relative to the middles.

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

Recent studies[32,35] examining replay in mazes with dynamically changing goals have challenged the "value view." 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, e.g., if replay modifies Q values, and these dictate choices, then replay should, if anything, lead to changes in behavior. Next, 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 those goals are.

In one study by Gillespie et al.,[32] rats foraged for a reward in an eight-arm maze, in which a single arm stably dispensed a reward for a block of trials, but this target moved periodically (Figure 3A). In each block, rats thus had to first identify which arm was rewarding ("search" phase) and then repeatedly visit it once found ("repeat" phase). Three findings about the content of replay during the repeat phase challenge the value view.

(1) 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).
(3) The enrichment of a goal for replay persisted for several blocks after it had become irrelevant, decreasing smoothly over time.

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 reward estimate for each arm, used 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). Specifically, the GR $G(s,a,:)$ encodes eight long-run value functions, one for each of the reward port states. The agent also learns two one-step reward functions $R_{behav}(g)$ and $R_{replay}(g)$, representing, respectively, which of the reward ports *currently* contains reward and how likely each is to contain reward *overall across blocks*. To decide where to go, the agent chooses a goal g according to R_{behav}. g, in turn, specifies a corresponding value function $G(:,:,g)$ that implies a policy. G, R_{behav}, and R_{replay} are updated from online experience. G is furthermore updated over a set of replayed transitions at the end of each episode. These are selected to maximize the EVB, in expectation over the distribution of goals specified by R_{replay}. See Algorithm 1 and STAR Methods for details.

Algorithm 1. Lagged replay GR agent.

Require:
  R_{behav}, R_{replay} ▷ RW processes
  G ▷ Geodesic representation

for $t = 1, 2, 3, \ldots$ **do**
  $s = 0$ ▷ Start state
  $g_{true} \sim$ Env ▷ The (latent) currently active goal

  # Agent picks a goal and navigates to it
  $g_{agent} \sim \text{softmax}(R_{behav})$ ▷ The agent's chosen goal
  while s not terminal **do**
    $a \sim \text{softmax}(G(s,:,g_{agent}))$
    $s' \sim \text{Env}(s,a)$
    update$(G; s, a, s')$ ▷ Online GR learning
    $s \leftarrow s'$
  end while

(Continued on next page)

Algorithm 1. ***Continued***

```
  # Agent receives and integrates reward information
  r ~ Env(s, g_true)
  update(R_behav; s, r)                          ▷ High learning rate
  update(R_replay; s, r)                         ▷ Low learning rate
  decay(G)                                       ▷ Forgetting
  # Agent performs replay
  for k = 1, 2, 3, ..., k_max do
    s_k, a_k, s'_k ~ replay(R_replay)            ▷ Priority in expectation over R_replay
    update(G; s_k, a_k, s'_k)                    ▷ Offline GR learning
  end for
end for
```

The key insight explaining Gillespie et al.'s findings is the distinction between the two representations of the goals, R_{behav} and R_{replay}, which serve different purposes. While trial-by-trial choice behavior must nimbly track the current goal (i.e., the *within-block* reward function), replay should be prioritized in part to learn routes to locations where future rewards are likely to occur (i.e., the *across-block* goal distribution). Animals must estimate both of these functions from ongoing experience. In particular, they cannot know that the programmed across-block distribution of goals is uniform and have only relatively few samples from which to estimate it: around ten blocks per goal over the entire experiment. Even given a uniform prior, when multinomial probabilities are estimated from samples, previously experienced samples will be overweighted. In settings where these probabilities may also change, such learning leads to a recency-weighted average.[53–55]

Accordingly, estimating these functions, by definition, requires learning over two nested timescales: choice behavior is governed by tracking the current within-block goal, whereas estimating the distribution over goals requires accumulating experiences over many such blocks, i.e., with a much slower learning rate. Importantly, animals directly experience the rate of block switching, and it is established that organisms can adapt Rescorla-Wagner-style learning to match the environment's rate of change.[55,56] Thus, in our simulations we stylize such learning using two Rescorla-Wagner learning processes:

$$R_{behav}(g) \leftarrow R_{behav}(g) + \eta_{behav}\left(r - R_{behav}(g)\right)$$
$$R_{replay}(g) \leftarrow R_{replay}(g) + \eta_{replay}\left(r - R_{replay}(g)\right)$$

Here, $\eta_{behav}, \eta_{replay} \in [0, 1]$ are learning rates, g is the sampled goal state, and r is the empirically received reward. Both Rescorla-Wagner processes operate over the candidate goals, but with different learning rates (η_{behav} is higher to track the current goal, and η_{replay} is 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 highest), 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 timestep (see STAR Methods). This is a standard assumption in learning models, typically justified by the possibility of contingency change,[53,56] and has the effect of ensuring that learning continues in an 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 potential environmental change.

Prioritized GR replay from our agent qualitatively matched the patterns observed by Gillespie et al.[32] Within a block, GR replay displayed a bias for the previous-block goal arm. By contrast, a Q-learning agent using Mattar's prioritized replay scheme[17] preferred the current-goal arm (Figure 3B). Furthermore, consistent with the data, replay of the current-goal arm increased over the block for the GR agent while decreasing for the Q-learning agent (Figure 3C). Finally, following this peak in replay for the current goal (say, during block n), enhanced replay of it gradually waned over several blocks after it was deactivated (Figure 3D).

These patterns arise in the model because the model's map is uniformly degraded by forgetting at each trial, though partly healed by online learning along the experienced path. Thus, when deciding which paths to replay at each step, the goal-conditioned EVBs are relatively uniform across the counterfactual goals. For this reason, replay's focus on those different goals is governed by their respective prevalence in the estimated across-block future goal distribution. Since this is being learned online, it reflects a recency-weighted average over reward experiences,[53–55] driving the initial focus on the most recent goal and the gradual incorporation and dominance of the latest one. The recency weighting can also be more directly appreciated in the gradual decline of goals' replay participation across blocks (Figure 3D).

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 (Figure 4A). Each day, 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 was biased toward the behaviorally non-preferred arm (Figure 4B).

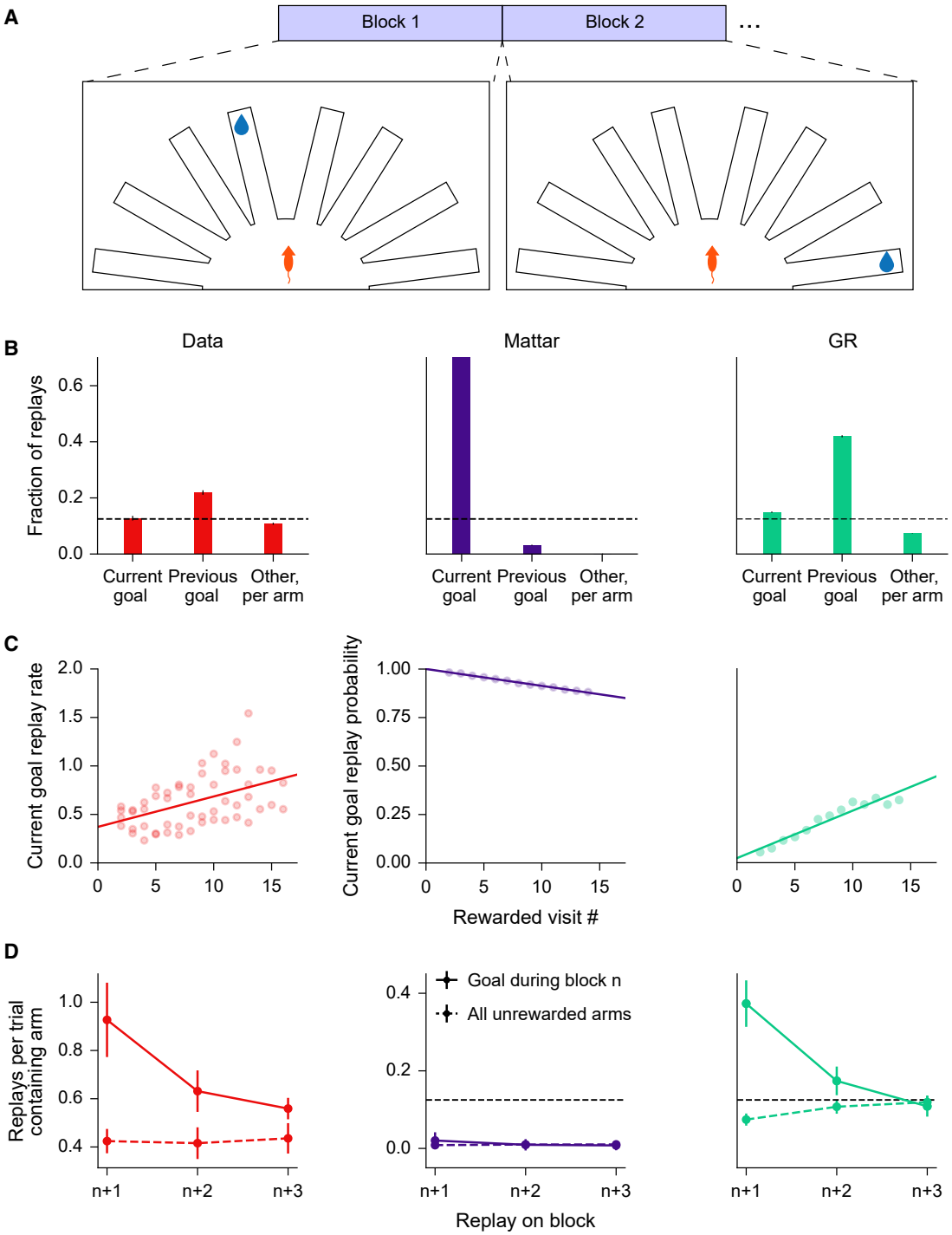

Figure 3. GR replay captures replay-behavior lag in Gillespie et al.

(A) Task schematic.

(B) Fraction of replays including 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.,[32] averaged across rats; middle: Q value replay; right: GR replay. Error bars indicate SEM.

(C) Rate of current-goal replay within a block as a function of rewarded visit number. Left: replotted data from Gillespie et al.,[32] averaged across rats; middle: Q value replay; right: GR replay.

(D) Rate of replay of the block *n* goal arm during subsequent blocks. Left: replotted data from Gillespie et al.,[32] averaged across rats, middle: Q value replay, right: GR replay. Error bars indicate SEM.

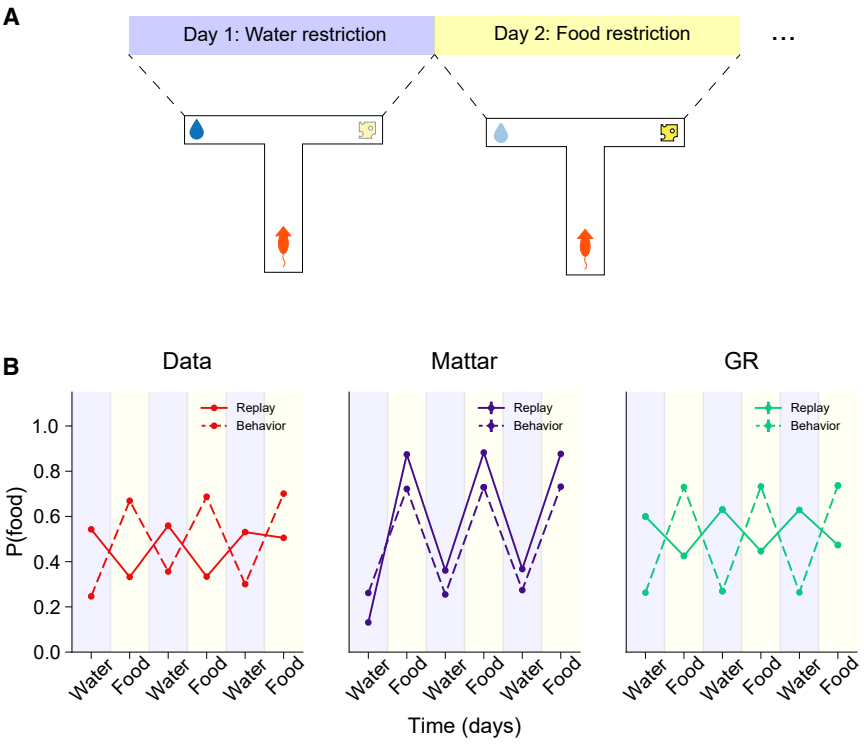

Figure 4. GR replay captures replay-behavior lag in Carey et al.

(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.,[35] middle: Q value replay/behavior, right: GR replay/behavior. Error bars indicate SEM.

We simulated the Carey task using a GR agent equipped with a fast-learning behavioral module and a goal distribution created by slow Rescorla-Wagner learning. The timescale for goal switching is slower than in the previous study (once rather than several times per day). For simplicity in simulation, we assume an appropriately calibrated goal-level learning rate. Relatedly, compared with the Gillespie study, rats had even fewer sessions—two or three for each goal—from which to estimate the goal distribution, making our assumption of learning by recency weighting even more plausible.

The effects of alternating food and water deprivation were realized by having asymmetric reward values for the two arms that switched between sessions.[57] Simulated replay recapitulated the mismatched behavior-replay pattern in the data (Figure 4B), again, because GR update priority is weighted by a goal distribution learned from across-block reward experiences. These favor maintaining the policy to visit the alternative goal, whose degradation by forgetting (hence, gain from replay) is not offset by online learning. By contrast, replay simulated from a Q-learning agent displayed a matched preference for the relevant reward in both behavior and replay (Figure 4B).

GR replay is predictive when prioritized under true transition dynamics

Previously, we showed that GR replay can recapitulate paradoxically lagged replay when goals are prioritized under a slow, block-level inference process that assumes recently encountered goals are more likely to recur in future. What if goals are instead prioritized under different statistical expectations about switching? We consider a different task with a more structured goal-switching pattern.[5] In contrast to those discussed above, this study has been taken to support the value or planning view of replay: in this setting, replay is biased toward the animals' current goal, is associated with improved navigation to it, and is well captured by Mattar's model.[17] Next we show that our model extends to this case: when generated under task-appropriate assumptions about goal dynamics, replay is *predictively* focused on navigating to the current goal.

We modeled the foraging task introduced by Pfeiffer and Foster.[5] A rat is placed into an open field with a 6x6 grid of reward wells (Figure 5A). In each session, one well is the "Home" well (changing between sessions). Each trial consists of a Home phase (where the Home well is active) and a Random phase (where one of the 35 remaining wells is active). Thus, each trial consists of navigation from a random location to a frequently visited fixed goal, followed by navigation from there to a rarely visited dynamic goal. We focus on three results from studies involving variants of this task, which all suggest a proactive, current-goal planning function for replay.

(1) Replay preferentially encodes the current Home well,[5] starting from the second trial (i.e., the first Home trial after Home was found).[43] We elide Pfeiffer's distinction between types of ripple events.
(2) Replay does not paradoxically favor the previous Random well compared with all other non-Home wells.[43]
(3) Home-containing replays are associated with improved performance in subsequent Home navigation, consistent with planning.[11]

The goal-switching pattern in this study differs from those discussed above in several ways that may affect animals' goal expectations and thus (in our model) their replay patterns. Although a new Home well is introduced for each session, the primary dynamic is the within-session alternation between Home and Random wells, with Home playing a sustained, predominant role. Since animals revisit the Home well every trial, they have considerable experience with this structure, and their behavior (latency and path length) demonstrates understanding of it. For instance, having discovered a new Home well, on the following trial they already return to it more readily than to the Random well.[43] Also, animals may be less prone here to exhibit recency bias in Random wells (compared with previous-block goals in the other tasks). Instead, we expect them to be closer to

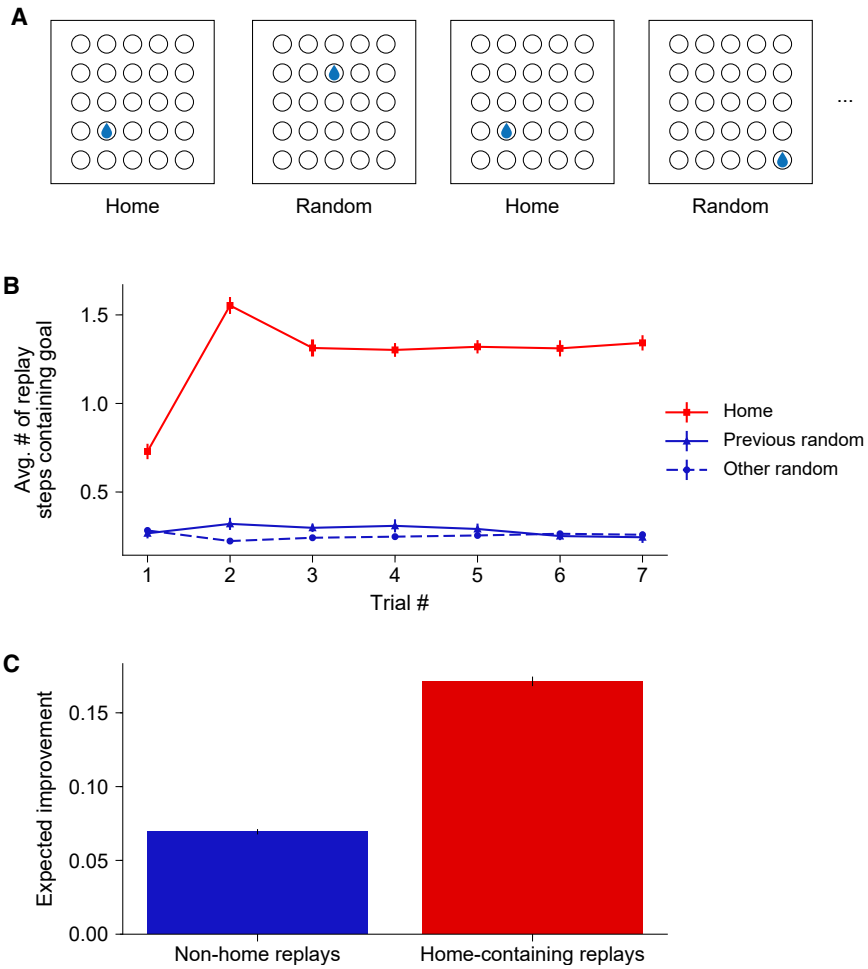

Figure 5. GR replay is coupled to current navigational goals when prioritized under rapidly alternating goal dynamics

(A) Schematic of the Pfeiffer and Foster[5] foraging task.

(B) Decomposition of states included in replay trajectories, plotted over time within a session. "Previous random" refers to the Random well selected on the previous trial. Error bars indicate SEM.

(C) Expected improvement (EI) in navigation efficiency due to Home-containing vs. Home non-containing replays before a Home trial. EI is measured as the change in expected occupancy of the Home well before vs. after execution of a candidate replay sequence. Error bars indicate SEM.

asymptotic understanding of the uniformity and stability of this distribution since they have considerable experience with long sequences of trials: more goal switches just in pretraining than the total animals experienced throughout the Carey or Gillespie studies. Another aspect of the study that affects the model is the size of the policy space: there are 35 routes from different wells to Home to learn and maintain, leaving less capacity for planning toward other future goals.

Accordingly, in contrast to our models of the Carey and Gillespie tasks, here we prioritized replay, assuming an understanding of the true, alternating goal dynamics (both the alternating structure and the uniform distribution of random goals—i.e., changing R_{behav} and R_{replay} in Algorithm 1 so that they are accurately computed under the goal dynamics, rather than being Rescorla-Wagner processes; see STAR Methods). This yielded replay with a proactive focus on the current goal, rather than lagged replay dynamics. That is, before Home trials, replay predictively focused on routes involving the Home well, starting immediately after it was discovered, with no paradoxical preference for the previous Random goal (Figure 5B). This reflects enhanced EVB for the Home goal due to its predominance in the veridical future expectancy, relative to the large, undifferentiated set of alternatives. Similarly, Home-containing candidate replay sequences were more effective in refining expected subsequent agent behavior than other sequences (Figure 5C).

GR replay trades off current and future goals by expectancy

We have emphasized that the current model extends the previous value view to, additionally, favor replay of routes to candidate goals expected in the future. This role of expectancy in weighting the utilities of replays with respect to different goals also implies one of the key predictions of this model for future empirical test: replay of current vs. other possible goals should be sensitive to learned expectations about their switching statistics.

To illustrate this, we simulated a dynamic maze navigation task with a detour replanning manipulation (Figure 6A). 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 rewarded, and given the goal on trial $t-1$, the rewarded goal on trial t was determined by a Markov process (Figure 6B). In this maze, the shortest path from the start state to either goal passes through the middle bottleneck state. Consequently, an agent executing this task should learn to route through that bottleneck regardless of the current goal. However, if on some trial it were to find that the middle bottleneck was inaccessible (e.g., if the path to it was blocked by a wall), it would need to replan using replay to build new policies for reaching each goal. Furthermore, if goal 1 is the 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 the need to visit the alternative will be.

We simulated this setup using 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} (Figure 6C, top left). By 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

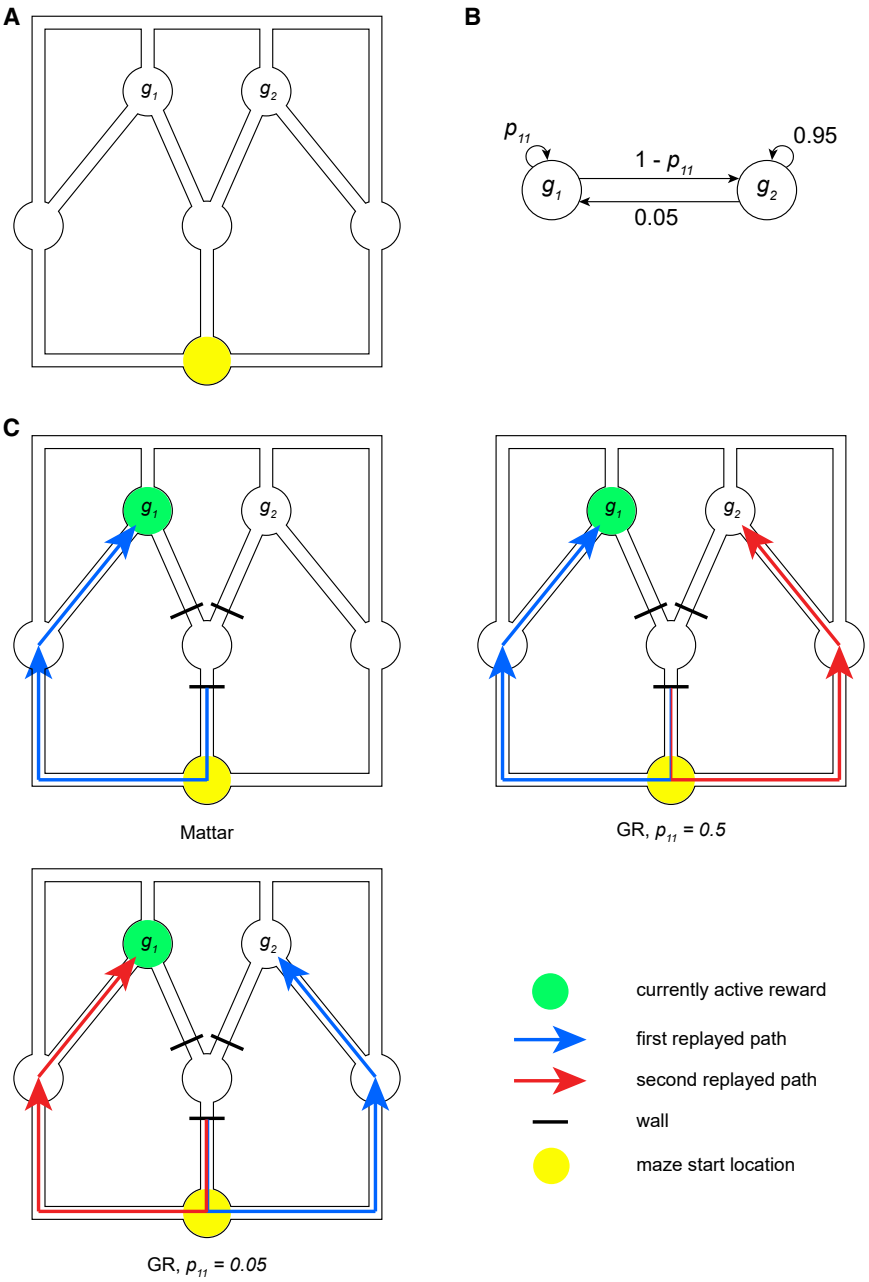

Figure 6. GR replay trades off future and current goals

(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}.

(Figure 6C, top right). By contrast, if $p_{11} = 0.05$ (low), it replays the path to g_2 first (Figure 6C, bottom). In general, in this model the priority (i.e., relative ordering and prominence) of replay of different possible routes, for both current vs. future goals, should depend parametrically on how often and soon they are expected to be obtained.

DISCUSSION

We propose an RL account of prioritizing replay in order to build cognitive maps. First, we formalize a hypothesis about how replay might be useful for building maps separate from current value, giving the map hypothesis the same formal specificity as the value hypothesis and framing it in the same terms. Our account distinguishes a set of candidate goal states and uses replay to learn shortest-path policies to each. For this, we introduced a cognitive map-like representation, the GR, which learns the state-action value functions in a set of modified MDPs where each goal is terminal and rewarding. This separates map information (a collection of Q functions for possible goals) from value information (which goals obtain) and suggests a role for replay in updating the former. In this way, value- and map-like replay arise as different extreme cases of a more general account, depending on the importance of current vs. hypothetical future goals.

Second, we characterize which replay sequences would be adaptive if this were the function of replay and clarify how these predictions differ from the value view. For this, we generalized the approach of Mattar and Daw,[17] computing the expected utility of an update to the GR from any 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 shortening the paths from states to potential goals, rather than adjusting current decision variables to better harvest current rewards.

This decoupling of the current and candidate future goals helps to capture results that challenge the value view's coupling of replay content to choice behavior. Two recent studies[32,35] found that in mazes with moving rewards, replay was biased toward goals associated with the *previous* reward block rather than (as prioritized Q value replay predicts) the *current* rewards. Our model captures these patterns as arising from the fact that the distribution of candidate goals (hypothesized to drive replay) must be learned across multiple goal

instances, i.e., across blocks or sessions, thus necessarily slower than the learning that drives within-block behavioral adjustment to each new goal.

On our account, in both tasks, this recency bias reflects misestimation of environmental dynamics, putatively due to limited sampling. Given limited experience with the objective uniformity and stability of the goal-switching distributions, for rational learners, goal estimates will be recency-weighted.[40,48,53,54] Given more experience, we assume that the agent can learn the true dynamics of the environment. In this case, recent goals would not have special status, and replay will be more predictively focused on true current goals. Our model thus explains why replay in some experiments has seemed more consistent with the value account: if some goal predominates over alternatives for a sustained period, if the task is complex enough that new paths to this goal remain uncomputed and crowd out priority for precomputing other paths, and if animals have enough experience to understand this, then the current model effectively reduces to the Mattar model, and replay focuses on planning to the current, primary goal. Accordingly, we showed that GR replay is able to recapitulate critical results in the Pfeiffer and Foster foraging task.[5,11,43] Together, these results unify the two dominant hypotheses regarding the function of replay: under the new model, the propensity of replay to encode the future vs. the past depends on the pattern of goal changes, the number and complexity of goals, and the agent's understanding of them. Our account suggests many experiments that could test the effect of these factors. For example, our hypothesis predicts that an altered version of the Gillespie task where the goal arm changed every trial with some predictable structure (e.g., rotating clockwise, together with a larger set of arms to strain capacity) would elicit predictive replay dynamics in line with Pfeiffer and Foster rather than retrospective replay.

Several simplifying assumptions here are opportunities for future work. First, to illustrate how replay depends on goal expectations, we used task-specific, hand-constructed rules for estimating goal expectancy. These assumptions are broadly consistent with what animals could be expected to learn about the objective goal dynamics of each task and with models and experiments showing that the brain gradually learns the structure and dynamics of environmental change.[40,47–50,54,58] This includes volatility parameters that capture how quickly events change and govern learning rates, grounding the values of those hand-tuned parameters in our simulations. A fuller account would combine these lines of work, nesting our replay rule under a task-agnostic learner that would estimate the structure and timescales of goal switching.

A related point is that prioritized replay only has empirical relevance if animals are operating in a data- and compute-limited regime. If the online experience is prodigious or the replay budget sufficiently cheap, then there is no need to judiciously prioritize replay. We suggest that, in these experiments, animals are not operating in this regime. Replay events are sparse, as is task experience relative to task complexity. We again use a stylized model (a fixed replay budget per trial), reflecting the general finding that across tasks, replays tend to occur a few at a time, mostly between trials. Again, a fuller model could replace this simplifying assumption with more detailed computations, trading off the opportunity costs and benefits of replay.[59] Notwithstanding, in the model, the replay budget balances against other free parameters (learning rates, task complexity, forgetting, and time discounting) to drive more or less focus on counterfactual goals, mainly via the goal-conditioned gain computations and by the balance between different goals in the goal-marginalized EVB. Our work thus suggests that task stability, complexity, and the opportunity cost of planning are all key knobs for future experiments to gain better insight into these trade-offs.

Since the Mattar account is a special case of GR replay, the new model inherits the earlier model's successes: the two models coincide when goals are sparse, stable, and focused. The GR replay model also displays several new qualitative replay dynamics, related to balancing potential vs. current goals. These offer a range of predictions for new tests. For instance, unlike replays 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.[27,60] GR replay is also sensitive to the structure and statistics of candidate goal states. Consequently, we predict that in environments with multiple potential goals, replay should focus on "central" states that are shared across many goals' paths. Moreover, we predict that the replay prioritization of goals, relative to each other and currently active goals, should be modulated by their statistical properties—that is, if one goal is more common, states associated with the path to it should be overweighted.

On the theoretical level, our model offers a new perspective on the function of replay in navigation and beyond. It exposes new parallels between the value and map hypotheses and addresses a high-level theoretical question: 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 (by aggregating local connectivity knowledge into plans for long-run routes) rather than simply strengthening or relocating them. Our model is thus spiritually connected with other views, such as complementary learning systems theory[61] and recent proposals (supported by human magnetoencephalography experiments[62,63]) suggesting that, beyond navigation, replay supports learning of compositional structures more generally. Although we do not yet extend to this degree of task generality, our teleological analysis enables us to reason about the value of replay for facilitating future reward gathering and make precise predictions about prioritization.

Regarding the alternative view of replay as maintaining memory per se, our theory conceptualizes even this as a prioritized computation. In our simulations, the GR underwent decay during each step of online behavior, which gives replay a role in continually rebuilding the GR. Thus, even if the overall goal is simply to maintain a faithful representation of the local environment, there is still nontrivial computation in selecting which parts 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. Zhou et al.[64]

extended a successful descriptive model of memory encoding and retrieval, the temporal context model (TCM), to encompass replay, viewed as associative spreading over memories. Despite its different rationale, this model makes similar predictions to Mattar's and ours; this may be because TCM's associations coincide with the SR,[65,66] which is also the need term in the RL models. Differences in the models' predictions will thus likeliest arise for situations where replay turns on gain, which quantifies the value of replays in serving the animal's goals and is not naturally a consideration in pure memory models. Zhou addresses these issues (like Talmi et al.[67]) by assuming that more rewarding and punishing events form stronger memories. This may lead to distinct predictions, because gain in our model treats better- vs. worse-than-expected events asymmetrically.[68] Similar considerations distinguish our model from Bakermans et al.,[69] who also suggest replay builds paths to particular goals. That model has no machinery for reasoning about the impact of different goal-switching dynamics or future goal distributions, which we predict are key for understanding differences between tasks. Meanwhile, Chen and van der Meer[36] build on complementary learning theory to suggest that "paradoxical replay" helps to build representations when experience is unbalanced. This perspective is broadly consistent with our view and further refined by our model's more RL-driven analysis about in which tasks these considerations have bite.

Our model leaves open a number of other issues for future work. First, as with Mattar's,[17] our account is not a process-level model of how replay is produced. Instead, we unpack the principles driving replay by characterizing how replay would behave if it were optimized (through whatever process) to serve the hypothesized goals. A biologically plausible implementation of replay prioritization would require a tractable approximation for computing gain (which, following Mattar, we compute unrealistically by brute force enumeration).

Second, our analysis is based on the GR. We chose it because it facilitates directly generalizing the Mattar model, but our broader argument generalizes to similar temporally abstract or goal-conditioned value representations.[40,44,70,71] One key feature of the GR (and an appealing alternative, the default representation[40]) is that, unlike the classic SR, it is off-policy: that is, it learns paths that would be appropriate when generalizing to other goals. Finally, both the GR and Q value prioritization frameworks are so far only well-defined in the tabular setting. It remains an exciting opportunity to understand how to reconcile these ideas with function approximation.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Nathaniel Daw (ndaw@princeton.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request.
- All original code has been deposited at Zenodo and is publicly available at 10.5281/zenodo.17042347 as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by U.S. Army Research Office grant W911NF-16-1-0474 (N.D.D.); Wellcome Trust career development award 225926/Z/22/Z (T.A.); and National Institutes of Health grants R01MH121093 (N.D.D.), R01DA047869 (I.B.W.), U19NS123716 (I.BW.), DP1MH136573 (I.B.W.), and R01MH135587 (N.D.D.).

AUTHOR CONTRIBUTIONS

Conceptualization, Y.S., I.B.W., and N.D.D.; methodology, Y.S., I.B.W., and N.D.D.; modeling, Y.S. and T.A.; analysis, Y.S.; writing – original draft, Y.S., T.A., I.B.W., and N.D.D.; writing – review & editing, Y.S., T.A., I.B.W., and N.D.D.; funding acquisition, I.B.W. and N.D.D.; supervision, I.B.W. and N.D.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Derivation of one-step need and gain for the GR
  - Multi-step backups
  - Prospective need evaluation
  - Prioritization under a goal dynamics process
  - General simulation details
  - Asymmetric T-maze simulation
  - Bottleneck chamber/Community graph simulations
  - Modeling the Gillespie task
  - Modeling the Carey task
  - Modeling the Pfeiffer task
  - Modeling the prediction task
  - Data replotting

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2025.09.021.

Received: February 26, 2024
Revised: June 27, 2025
Accepted: September 16, 2025
Published: October 27, 2025

REFERENCES

1. O'Keefe, J., and Nadel, L. (1978). The Hippocampus as a Cognitive Map (Oxford University Press).
2. Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal Replay of Extended Experience. Neuron *63*, 497–507. https://doi.org/10.1016/j.neuron.2009.07.027.
3. Diba, K., and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. Nat. Neurosci. *10*, 1241–1242. https://doi.org/10.1038/nn1961.

4. Carr, M.F., Jadhav, S.P., and Frank, L.M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. Nat. Neurosci. *14*, 147–153. https://doi.org/10.1038/nn.2732.
5. Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79. https://doi.org/10.1038/nature12112.
6. Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature *440*, 680–683. https://doi.org/10.1038/nature04587.
7. Euston, D.R., Tatsuno, M., and McNaughton, B.L. (2007). Fast-Forward Playback of Recent Memory Sequences in Prefrontal Cortex During Sleep. Science *318*, 1147–1150. https://doi.org/10.1126/science.1148979.
8. Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S.I., and Battaglia, F.P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat. Neurosci. *12*, 919–926. https://doi.org/10.1038/nn.2337.
9. Kaefer, K., Nardin, M., Blahna, K., and Csicsvari, J. (2020). Replay of Behavioral Sequences in the Medial Prefrontal Cortex during Rule Switching. Neuron *106*, 154–165.e6. https://doi.org/10.1016/j.neuron.2020.01.015.
10. Gupta, A.S., van der Meer, M.A.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. Neuron *65*, 695–705. https://doi.org/10.1016/j.neuron.2010.01.034.
11. Widloski, J., and Foster, D.J. (2022). Flexible rerouting of hippocampal replay sequences around changing barriers in the absence of global place field remapping. Neuron *110*, 1547–1558.e8. https://doi.org/10.1016/j.neuron.2022.02.002.
12. Foster, D.J., and Knierim, J.J. (2012). Sequence learning and the role of the hippocampus in rodent navigation. Curr. Opin. Neurobiol. *22*, 294–300. https://doi.org/10.1016/j.conb.2011.12.005.
13. Wu, X., and Foster, D.J. (2014). Hippocampal Replay Captures the Unique Topological Structure of a Novel Environment. J. Neurosci. *34*, 6459–6469. https://doi.org/10.1523/JNEUROSCI.3414-13.2014.
14. de Lavilléon, G., Lacroix, M.M., Rondi-Reig, L., and Benchenane, K. (2015). Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. Nat. Neurosci. *18*, 493–495. https://doi.org/10.1038/nn.3970.
15. Gridchyn, I., Schoenenberger, P., O'Neill, J., and Csicsvari, J. (2020). Assembly-Specific Disruption of Hippocampal Replay Leads to Selective Memory Deficit. Neuron *106*, 291–300.e6. https://doi.org/10.1016/j.neuron.2020.01.021.
16. Kay, K., Chung, J.E., Sosa, M., Schor, J.S., Karlsson, M.P., Larkin, M.C., Liu, D.F., and Frank, L.M. (2020). Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus. Cell *180*, 552–567.e25. https://doi.org/10.1016/j.cell.2020.01.014.
17. Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci. *21*, 1609–1617. https://doi.org/10.1038/s41593-018-0232-z.
18. Lansink, C.S., Goltstein, P.M., Lankelma, J.V., McNaughton, B.L., and Pennartz, C.M.A. (2009). Hippocampus Leads Ventral Striatum in Replay of Place-Reward Information. PLoS Biol. *7*, e1000173. https://doi.org/10.1371/journal.pbio.1000173.
19. Singer, A.C., and Frank, L.M. (2009). Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus. Neuron *64*, 910–921. https://doi.org/10.1016/j.neuron.2009.11.016.
20. Gomperts, S.N., Kloosterman, F., and Wilson, M.A. (2015). VTA neurons coordinate with the hippocampal reactivation of spatial experience. eLife *4*, e05360. https://doi.org/10.7554/eLife.05360.
21. Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. eLife *4*, e06063. https://doi.org/10.7554/eLife.06063.
22. Ambrose, R.E., Pfeiffer, B.E., and Foster, D.J. (2016). Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. Neuron *91*, 1124–1136. https://doi.org/10.1016/j.neuron.2016.07.047.
23. Gruber, M.J., Ritchey, M., Wang, S.F., Doss, M.K., and Ranganath, C. (2016). Post-learning Hippocampal Dynamics Promote Preferential Retention of Rewarding Events. Neuron *89*, 1110–1120. https://doi.org/10.1016/j.neuron.2016.01.017.
24. Zielinski, M.C., Tang, W., and Jadhav, S.P. (2020). The role of replay and theta sequences in mediating hippocampal-prefrontal interactions for memory and cognition. Hippocampus *30*, 60–72. https://doi.org/10.1002/hipo.22821.
25. Singer, A.C., Carr, M.F., Karlsson, M.P., and Frank, L.M. (2013). Hippocampal SWR Activity Predicts Correct Decisions during the Initial Learning of an Alternation Task. Neuron *77*, 1163–1173. https://doi.org/10.1016/j.neuron.2013.01.027.
26. Liu, Y., Mattar, M.G., Behrens, T.E.J., Daw, N.D., and Dolan, R.J. (2021). Experience replay is associated with efficient nonlocal learning. Science *372*, eabf1357. https://doi.org/10.1126/science.abf1357.
27. Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208. https://doi.org/10.1037/h0061626.
28. Grosmark, A.D., Sparks, F.T., Davis, M.J., and Losonczy, A. (2021). Reactivation predicts the consolidation of unbiased long-term cognitive maps. Nat. Neurosci. *24*, 1574–1585. https://doi.org/10.1038/s41593-021-00920-7.
29. Moser, E.I., Kropff, E., and Moser, M.B. (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. Annu. Rev. Neurosci. *31*, 69–89. https://doi.org/10.1146/annurev.neuro.31.061307.090723.
30. Roux, L., Hu, B., Eichler, R., Stark, E., and Buzsáki, G. (2017). Sharp wave ripples during learning stabilize the hippocampal spatial map. Nat. Neurosci. *20*, 845–853. https://doi.org/10.1038/nn.4543.
31. Schiller, D., Eichenbaum, H., Buffalo, E.A., Davachi, L., Foster, D.J., Leutgeb, S., and Ranganath, C. (2015). Memory and Space: Towards an Understanding of the Cognitive Map. J. Neurosci. *35*, 13904–13911. https://doi.org/10.1523/JNEUROSCI.2618-15.2015.
32. Gillespie, A.K., Astudillo Maya, D.A., Denovellis, E.L., Liu, D.F., Kastner, D.B., Coulter, M.E., Roumis, D.K., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. Neuron *109*, 3149–3163.e6. https://doi.org/10.1016/j.neuron.2021.07.029.
33. Sutton, R.S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. SIGART Bull. *2*, 160–163. https://doi.org/10.1145/122344.122377.
34. Krausz, T.A., Comrie, A.E., Kahn, A.E., Frank, L.M., Daw, N.D., and Berke, J.D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. Neuron *111*, 3465–3478.e7. https://doi.org/10.1016/j.neuron.2023.07.017.
35. Carey, A.A., Tanaka, Y., and van der Meer, M.A.A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. Nat. Neurosci. *22*, 1450–1459. https://doi.org/10.1038/s41593-019-0464-6.
36. Chen, H.T., and van der Meer, M.A.A. (2024). Paradoxical replay can protect contextual task representations from destructive interference when experience is unbalanced. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.09.593332.
37. Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711. https://doi.org/10.1038/nn1560.
38. Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Comput. *5*, 613–624. https://doi.org/10.1162/neco.1993.5.4.613.
39. Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat. Neurosci. *20*, 1643–1653. https://doi.org/10.1038/nn.4650.

40. Piray, P., and Daw, N.D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. Nat. Commun. *12*, 4942. https://doi.org/10.1038/s41467-021-25123-3.

41. Russek, E.M., Momennejad, I., Botvinick, M.M., Gershman, S.J., and Daw, N.D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput. Biol. *13*, e1005768. https://doi.org/10.1371/journal.pcbi.1005768.

42. Momennejad, I., Russek, E.M., Cheong, J.H., Botvinick, M.M., Daw, N.D., and Gershman, S.J. (2017). The successor representation in human reinforcement learning. Nat. Hum. Behav. *1*, 680–692. https://doi.org/10.1038/s41562-017-0180-8.

43. Pfeiffer, B.E. (2022). Spatial learning drives rapid goal representation in hippocampal ripples without place field accumulation or goal-oriented theta sequences. J. Neurosci. *42*, 3975–3988. https://doi.org/10.1523/JNEUROSCI.2479-21.2022.

44. Barreto, A., Dabney, W., Munos, R., Hunt, J.J., Schaul, T., van Hasselt, H., and Silver, D. (2018). Successor Features for Transfer in Reinforcement Learning. Preprint at arXiv.

45. Kaelbling, L.P. (1993). Learning to Achieve Goals. IJCAI, 2, 1094–1098.

46. Gardner, M.P.H., Schoenbaum, G., and Gershman, S.J. (2018). Rethinking dopamine as generalized prediction error. Proc. Biol. Sci. *285*, 20181645. https://doi.org/10.1098/rspb.2018.1645.

47. Courville, A.C., Gordon, G.J., Touretzky, D., and Daw, N. (2003). Model uncertainty in classical conditioning. Advances in Neural Information Processing Systems. 16.

48. Courville, A.C., Daw, N.D., and Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. Trends Cogn. Sci. *10*, 294–300. https://doi.org/10.1016/j.tics.2006.05.004.

49. Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. *20*, 251–256. https://doi.org/10.1016/j.conb.2010.02.008.

50. Gershman, S.J., Norman, K.A., and Niv, Y. (2015). Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. *5*, 43–50. https://doi.org/10.1016/j.cobeha.2015.07.007.

51. Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. Nat. Neurosci. *16*, 486–492. https://doi.org/10.1038/nn.3331.

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

53. Kakade, S., and Dayan, P. (2002). Acquisition and extinction in autoshaping. Psychol. Rev. *109*, 533–544. https://doi.org/10.1037/0033-295x.109.3.533.

54. Mathys, C., Daunizeau, J., Friston, K.J., and Stephan, K.E. (2011). A bayesian foundation for individual learning under uncertainty. Front. Hum. Neurosci. *5*, 39. https://doi.org/10.3389/fnhum.2011.00039.

55. Piray, P., and Daw, N.D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. Nat. Commun. *12*, 6587. https://doi.org/10.1038/s41467-021-26731-9.

56. Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221. https://doi.org/10.1038/nn1954.

57. Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. Trends Cogn. Sci. *10*, 375–381. https://doi.org/10.1016/j.tics.2006.06.010.

58. McGuire, J.T., Nassar, M.R., Gold, J.I., and Kable, J.W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. Neuron *84*, 870–881. https://doi.org/10.1016/j.neuron.2014.10.013.

59. Agrawal, M., Mattar, M.G., Cohen, J.D., and Daw, N.D. (2022). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. Psychol. Rev. *129*, 564–585. https://doi.org/10.1037/rev0000309.

60. Blodgett, H.C. (1929). The effect of the introduction of reward upon the maze performance of rats. In University of California Publications in Psychology, *4* (University of California), pp. 113–134.

61. McClelland, J.L., McNaughton, B.L., and 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. Psychol. Rev. *102*, 419–457. https://doi.org/10.1037/0033-295X.102.3.419.

62. Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., and Behrens, T. (2023). Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. Cell *186*, 4885–4897.e14. https://doi.org/10.1016/j.cell.2023.09.004.

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

64. Zhou, Z., Kahana, M.J., and Schapiro, A.C. (2023). A unifying account of Replay as context-driven memory reactivation. eLife *13*, RP99931. https://doi.org/10.1101/2023.03.22.533833.

65. Gershman, S.J., Moore, C.D., Todd, M.T., Norman, K.A., and Sederberg, P.B. (2012). The Successor Representation and Temporal Context. Neural Comput. *24*, 1553–1568. https://doi.org/10.1162/NECO_a_00282.

66. Zhou, C.Y., Talmi, D., Daw, N., and Mattar, M.G. (2023). Episodic Retrieval for Model-Based Evaluation in Sequential Decision Tasks. Psychol Rev. *132*, 18–49.

67. Talmi, D., Lohnas, L.J., and Daw, N.D. (2019). A retrieved context model of the emotional modulation of memory. Psychol. Rev. *126*, 455–485. https://doi.org/10.1037/rev0000132.

68. Hunter, L.E., Meer, E.A., Gillan, C.M., Hsu, M., and Daw, N.D. (2022). Increased and biased deliberation in social anxiety. Nat. Hum. Behav. *6*, 146–154. https://doi.org/10.1038/s41562-021-01180-y.

69. Bakermans, J.J.W., Warren, J., Whittington, J.C.R., and Behrens, T.E.J. (2025). Constructing future behavior in the hippocampal formation through composition and replay. Nat. Neurosci. *28*, 1061–1072. https://doi.org/10.1038/s41593-025-01908-3.

70. Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In International Conference on Machine Learning (PMLR), pp. 1312–1320.

71. Eysenbach, B., Zhang, T., Levine, S., and Salakhutdinov, R.R. (2022). Contrastive learning as goal-conditioned reinforcement learning. Advances in Neural Information Processing Systems 35, 35603–35620.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python 3.9	Python.org	N/A

METHOD DETAILS

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:

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

Now, we use the environmental dynamics to observe that:

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

and therefore:

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

Plugging that into Equation 7, we get:

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

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

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

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:

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

The case where $s_k \neq g$ is clearly a *need* $\times$ *gain* factorisation (first sum is a need term, second sum is a gain term). The case where $s_k = g$ 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 generalize 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.

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]

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:

$$e^* = \underset{e_k}{\operatorname{argmax}} \mathbb{E}_{g\sim P(g), s_0 \sim P(s_0)} \left[H_{post}(s_0, g) - H_{pre}(s_0, g) \right], \quad \text{(Equation 10)}$$

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:

$$\text{need}(s_k, g) = \mathbb{E}_{s_0 \sim P(s_0)} \left[\sum_{i=0}^{\infty} \gamma^i P\left(s_0 \rightarrow s_k, i, \pi_{g,pre}\right) \right]. \quad \text{(Equation 11)}$$

Prioritization under a goal dynamics process

When first describing the GR, we motivated it 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 well-switching statistics in the Pfeiffer model or the goal dynamics process in Figure 6).

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:

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

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

General 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, Gillespie, and Pfeiffer tasks), Q-learning agents selected actions via the softmax choice rule $\pi(a|s) \propto \exp(Q(s, a)/\tau)$. 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 $\pi_g(a|s) \propto \exp(G(s, a, g)/\tau)$. 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 (exact 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 behavior and due to replay.

Asymmetric T-maze simulation

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.

Bottleneck chamber/Community graph simulations

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 analyses were otherwise conducted as in the bottleneck chamber.

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 Figure S2). 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 executed 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. This choice 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). Throughout online behavior, the agent updated its internal Q-value matrix or GR in accordance with the states, actions, successor states, and rewards it observed (i.e., in addition to learning from replay, the agent also learns from online experience; this, at least partly, serves to counteract decay and drive replay away from current goals).

Both agents selected their navigational goal arm by sampling from a bespoke behavioral module R_{behav} that tracked the value of the candidate arms. In particular, the agents sampled a goal to pursue via the softmax choice rule $\pi(\text{arm}) \propto \exp(R_{behav}(\text{arm})/\tau)$ applied to per-arm values learned with the Rescorla-Wagner algorithm:

$$R_{behav}(\text{chosen arm}_t) \leftarrow R_{behav}(\text{chosen arm}_t) + \eta(r_t - R_{behav}(\text{chosen arm}_t))$$

Here, we use $\eta = 1$ (due to the determinism of the reward schedule) and $\tau = 0.35$ (the value of τ essentially controls the rates of lapses during the repeat phase, which has a modest but generally insignificant effect on the extent to which states in the rewarded arm are forgotten in between repeated visits).

After each trial, both agents were required to perform three replay steps – i.e., the distance from the start state of the maze to any goal state. In general, the number of replay steps to perform per replay ''event'' is a free parameter in this model. Clearly, if this number were set ''too high'' then replay could simply learn (big chunks of) the entire Q-function/GR in a single event, obviating the need for prioritization. On the other hand, if it is ''too low'' then replay would not be able to effectively aid learning the Q-function/GR. We reasoned that, empirically, replay events tend to be sparse (i.e., the replay budget is tightly-controlled); accordingly, setting this parameter such that replay traces out single, but full, behavioral trajectories approximates a theoretically-justifiable ''sweet spot''.

The Q-learning agent performed the task using 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 $c_{forg} = 0.94$. The specific value of this parameter is not critical, though if it is close to 1 then there is too little forgetting to incentivize intelligent prioritization of replay (note that lower values of c_{forg} induce stronger forgetting, which serves to raise the gain parameter). The Q-learning agent performed policy updates under a softmax rule over the underlying Q-values with a separate temperature parameter $\tau_{policy} = 0.20$ (this is not important for behavior 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; in this case, lower temperature values mean that value updates more strongly affect future behavior, thereby increasing gain). Finally, we assumed that updates due to replay had a lower learning rate $\alpha_{replay} = 0.7$ than online behavior.

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 behavioral value learning was identical to the Q-learning agent (i.e., softmax choice rule with $\tau = 0.35$ over values learned with $\eta = 1$ Rescorla-Wagner, constant decay of the GR every time-step with $c_{forg} = 0.94$). Furthermore, the GR agent also performed policy updates under a softmax rule with temperature parameter $\tau_{policy} = 0.2$. However, in addition to the per-arm behavioral values, the GR agent also separately maintained a Rescorla-Wagner process R_{replay} to learn and maintain a distribution over goal arms for the sake of prioritizing replay. 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 behavioral counterpart. Here, the per-arm probabilities are initialized at $1/8$ (i.e., a uniform distribution over the candidate goal arms). Upon encountering a new active goal, these values are updated towards a target consisting of a one-hot vector encoding the location of the newly-discovered goal, with a slow learning rate of $\eta_{goal} = 0.05$. (It is straightforward to show that if initialized as a distribution, the Rescorla-Wagner rule will always maintain the learned values as a distribution so long as the target is a one-hot vector.)

Modeling the Carey task

In our model of the Carey[35] task, agents were placed in the starting state of a T-maze (see Figure S2 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, $n_a = 200$ virtual subjects were simulated, each undergoing $n_s = 6$ sessions of alternating water/food restriction that lasted $n_t = 20$ 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 behavioral 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 behavior 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 $\tau = 0.5$ (adjusted from the Gillespie value to more accurately match the behavioral lapse rate in the Carey data).

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 $c_{forg} = 0.94$. Its policy updates were performed under a softmax regime with $\tau_{policy} = 0.40$. The learning rate for replay was assumed to be lower than for online behavior, with $\alpha_{replay} = 0.7$. The GR agent had the same values for these parameters. Furthermore, it had a replay-value learning rate of $\eta_{goal} = 0.04$ (note that this is marginally lower than the Gillespie value to account for the slightly longer block length, but the exact value of this parameter is not critically important), which it used to learn a replay goal distribution analogously to the Gillespie agent.

Modeling the Pfeiffer task

In our implementation of the Pfeiffer and Foster[5] foraging task, $n_a = 25$ subjects underwent $n_s = 10$ sessions consisting of $n_t = 15$ trials. Trials were composed of an initial Home phase and a subsequent Random phase. The agent's behavior model sampled goals for the agent to pursue based on the true phase-specific goal distributions (i.e., sampling only Home on Home phases after Home is identified and uniformly on Random phases/before Home is identified). The agent prioritized replay using the approach outlined in "Prioritization under a goal dynamics process", under a goal dynamics matrix that encoded the task's alternating Home-Random structure, as well as the agent's joint posterior over the current trial phase and the location of the Home well. The episodic discount factor for this process is $\gamma_{episodic} = 0.95$.

As in the previous simulations, the GR agent performs policy updates under a softmax regime with $\tau_{policy} = 0.40$, underwent forgetting with $c_{forg} = 0.94$, and had online- and replay-associated learning rates of 1.0 and 0.7, respectively. The agent performed ten replay steps in between each trial (chosen arbitrarily; unlike in the previous two simulations, there is no fixed trial-length that could be picked).

Modeling the prediction task

In our implementation of the prediction task, we simulated Q-learning and GR agents on the maze in Figure 6 using the state diagram provided in Figure S2. 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 behavioral policy

with $\tau = 0.01$. The GR agent was assumed to know the true goal dynamics process and performed prioritized replay as described in ''Prioritization under a goal dynamics process'', with an episodic discount factor equal to 0.9.

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.