A Model of Hippocampal Replay Driven by Experience and Environmental Structure Facilitates Spatial Learning

Nicolas Diekmann[1,2], Sen Cheng[1,2]*

1 Institute for Neural Computation, Faculty of Computer Science, Ruhr University Bochum, Germany
2 International Graduate School of Neuroscience, Ruhr University Bochum, Germany

*** sen.cheng@rub.de**

Abstract

Replay of neuronal sequences in the hippocampus during resting states and sleep play an important role in learning and memory consolidation. Consistent with these functions, replay sequences have been shown to obey current spatial constraints. Nevertheless, replay does not necessarily reflect previous behavior and can construct never-experienced sequences. Here we propose a stochastic replay mechanism that prioritizes experiences based on three variables: 1. Experience strength, 2. experience similarity, and 3. inhibition of return. Using this prioritized replay mechanism to train reinforcement learning agents leads to far better performance than using random replay. Its performance is close to the state-of-the-art, but biologically unrealistic, algorithm by Mattar & Daw (2018). Importantly, our model reproduces diverse types of replay because of the stochasticity of the replay mechanism and experience-dependent differences between the three variables. In conclusion, a unified replay mechanism generates diverse replay statistics and is efficient in driving spatial learning.

Introduction

Humans and other animals continuously make decisions that may hugely impact their well-being, be that shortly after emitting a choice or much later. To successfully optimize their behavior, animals have to must be able to correctly credit choices with resulting consequences, adapt their future behavior, and remember what they have learned. The hippocampus is known to play a critical role in the formation of new memories and retrieval of memories as evidenced by the famous case of patient H.M. (Corkin et al., 1997) and others (Wilson et al., 1995; Rosenbaum et al., 2004). In rats and mice, damage to the hippocampus is known to impair spatial learning and memory (Morris et al., 1982; Deacon et al., 2002). An important phenomenon linked to learning and memory found in the hippocampus is that of "replay" (Buhry et al., 2011). As an animal navigates within an environment, so-called place cells (O'Keefe and Dostrovsky, 1971) in the hippocampus are sequentially activated. Later, during awake resting states and during sleep, compressed reactivation of the aforementioned sequences can be observed within events of high-frequency neural activity which are known as sharp wave/ ripples (Buzsáki, 1989) (SWRs). These sequences preferentially start at the current position of the animal (Davidson et al., 2009) and can occur in the order observed during behavior as well as in the reverse order (Diba and Buzsáki, 2007).

Consistent with proposed function of replay in learning, Widloski and Foster (2022) showed that in a goal-directed navigation task replay sequences obey spatial constraints, when barriers in an environment change on a daily basis. By contrast, other studies suggest that replay is not limited to just reflecting the animal's previous behavior. For instance, replay can represent shortcuts that the animals had never taken (Gupta et al., 2010). Replay during sleep appeared to represent trajectories through reward-containing regions that the animals had seen, but never explored (Ólafsdóttir et al., 2015). Following a foraging task, replay sequences resembled random walks, i.e., their statistics

was described by a Brownian diffusion process, event though the preceding behavioral trajectories did not (Stella et al., 2019). How can the hippocampus support the generation of such a variety of replay statistics and in what way do the sequences facilitate learning? Here we address these questions by studying replay and its effect on learning using computational modelling.

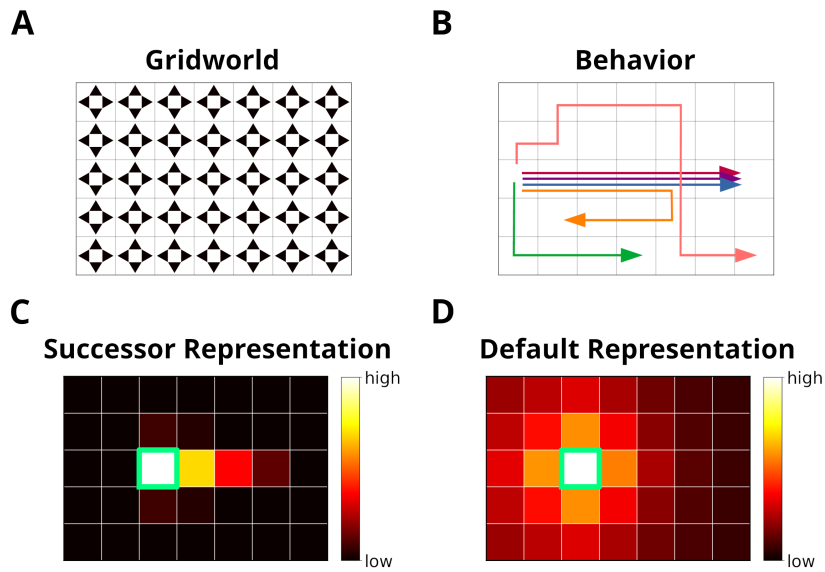

Figure 1. The grid world. A: An example of a simple grid world environment. An agent can transition between the different states, i.e. squares in the grid, by moving in the four cardinal directions depicted as arrows. **B**: Example trajectories of an agent moving in the grid world. **C**: The successor representation (SR) for one state (green frame). Note that the SR depends on the agent's behavior. **D**: The default representation (DR) for the same state as in C. In contrast to the SR, the DR does not depend on the agent's behavior and is equivalent to the SR given random behavior.

We adopt the reinforcement learning framework (Sutton and Barto, 2018) (RL), which formulates the problem of crediting choices with resulting consequences as an agent interacting with its environment and trying to maximize the expected cumulative reward. A common way of solving this problem involves the learning of a so-called value function, which maps pairs of environmental states and actions to expected future rewards, and then choosing actions that yield the highest value. A popular algorithm that can learn such a function is Q-learning (Watkins, 1989) in which the value function is referred to as Q-function and expected future rewards are referred to as Q-values. The Q-function is updated using the so-called temporal-difference (TD) error which is computed from an experience's immediate reward and Q-function estimates of future reward (see Methods section for more details). While RL has traditionally been used to solve technical control problems, it has recently been adopted to model animal (Bathellier et al., 2013) and human (Redish et al., 2007; Zhang et al., 2018) behavior.

RL generally requires many interactions with the environment, which results in slow and inefficient learning. Interestingly, replay of stored experiences, i.e. interactions with the environment, greatly improves the speed of learning (Lin, 1992). The function of experience replay in RL has been linked to that of hippocampal replay in driving spatial learning (Johnson and Redish, 2005). Mattar and Daw (2018) proposed a model of hippocampal replay as the replay of experiences which have the highest "utility" from a reinforcement learning perspective, meaning that the experience which yields the highest behavioral improvement is reactivated. In their model, the agent's environment is represented as a so-called grid world, which discretizes space into abstract states between which the agent can move using the four cardinal actions (Fig.1A). During experience replay, experiences associated with the different transitions in the environment can be reactivated by the agent. Their model accounts for multiple observed replay phenomena. However, computing the update utility in their model

requires evaluating how the agent's policy would change for all stored experiences. It is unclear how the brains of animals could compute or even approximate these utilities.

Here we introduce a model of hippocampal replay which is driven by experience and environmental structure, and therefore does not require biologically implausible computations like the one proposed by Mattar and Daw (2018). Nevertheless, our replay mechanism facilitates learning in a series of spatial navigation tasks and comes close to the performance of Mattar and Daw (2018)'s model. In simulations we show how the variety of hippocampal replay statistics can emerge from the variables that drive our model.

Results

Using Structural Knowledge and the Statistics of Experience to Prioritize Replay

We propose a model for the prioritization of experiences which we call *Spatial structure and Frequency-weighted Memory Access*, or SFMA for short. The model was conceived of with simplified grid world environments in mind (Fig.1). Each node in the grid corresponds to an environmental state. During behavior a reinforcement learning agent transitions between nodes and stores these transitions as experience tuples $e_t = (s_t, a_t, r_t, s_{t+1})$. Here, s_t is the current state, a_t is the action executed by the agent, r_t is a reward or punishment received after transitioning and s_{t+1} is the next state.

Reactivating stored experiences is viewed as analogous to the reactivation of place cells during replay events in the hippocampus. Assuming an experience e_t has just been reactivated, each stored experience e is assigned a priority rating $R(e|e_t)$ based on its strength $C(e)$, its similarity $D(e|e_t)$ to e_t and inhibition $I(e)$ applied to it (Fig. 2A):

$$R(e|e_t) = C(e)D(e|e_t)[1 - I(e)] \tag{1}$$

The experience strength $C(e)$ is modulated by the frequency of experience and reward.

The similarity $D(e|e_t)$ between e and e_t reflects the spatial distance between states, taking into account structural elements of the environment, such as walls and other obstacles. This rules out the use of the Euclidean distance. A possible candidate would be the successor representation (SR) (Dayan, 1993). Given the current state state s_j, the SR represents as a vector the likelihood of visiting the other states s_i in the near future (Fig.1C). This likelihood is commonly referred to as the discounted future occupancy. Since states separated by a barrier are temporally more distance to each other than without barrier their expected occupancy is reduced more by discounting. Unfortunately, the SR suffers from two disadvantages. First, the SR depends on the agent's behavior, i.e. the agent's policy, which will distort the distance relationships between states unless that agent behaves randomly. Second, if the structure of the environment changes, the SR has to be relearned completely, and this may be complicated further by the first problem. These problems were recently addressed by Piray and Daw (2021) in the form of the default representation (DR) (Fig.1D). Unlike the SR the DR does not depend on the agent's policy but rather on a default policy, which is a uniform action policy, i.e. random behavior, and the DR was shown to reflect the distance relationships between states even in the presence of barriers. Importantly, Piray and Daw (2021) demonstrated that the DR can be efficiently updated if the environmental structure changes using a low-rank update matrix. We chose to base experience similarity on the DR due to its advantageous features.

Since experience tuples contain two states – the current state s_t and next state s_{t+1} – we consider two ways of measuring the similarity between experiences. In the default mode we compare the current states of two experiences (Fig. 2B). In the reverse mode the current state of the most recently reactivated experience is compared with the next states of the other experiences. We found that default and reverse modes tend to generated different kinds of sequences.

After an experience e has been reactivated, inhibition $I(e)$ is applied to prevent the repeated reactivation of the same experience. Inhibition is applied to all experiences

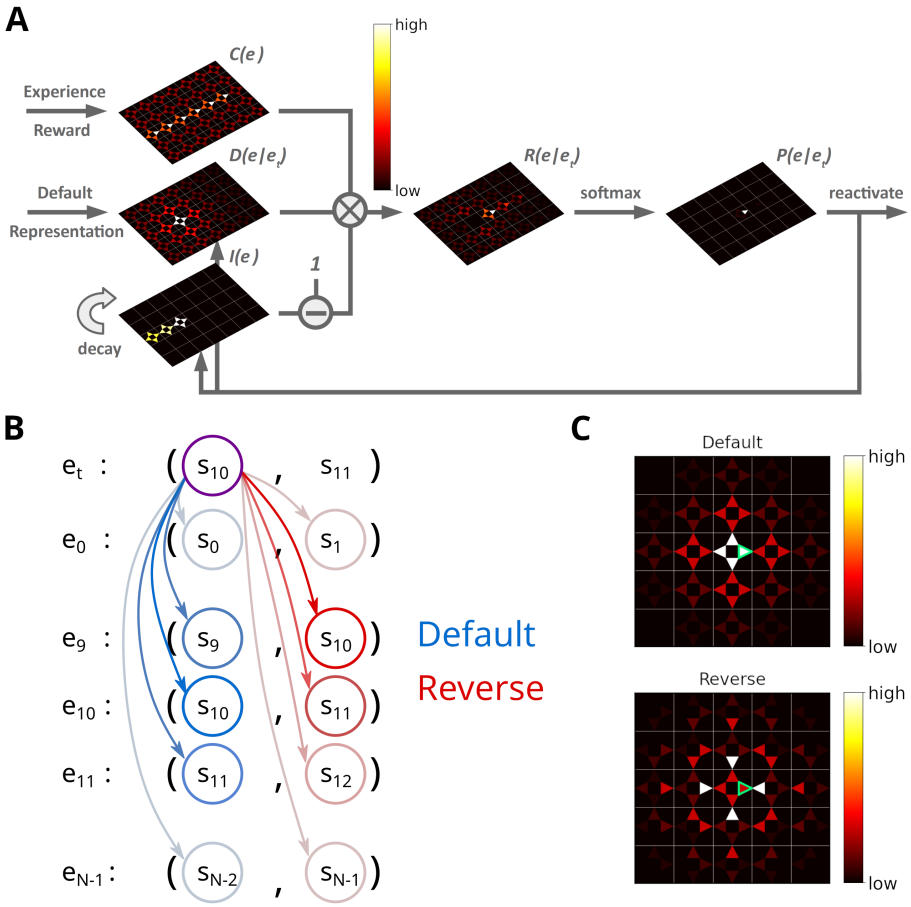

Figure 2. Illustration of the Spatial structure and Frequency-weighted Memory Access (SFMA) replay model. A: The interaction between the variables in our replay mechanism. Experience strength $C(e)$, experience similarity $D(e|e_t)$ and inhibition of return $I(e)$ are combined to form reactivation prioritization ratings. Reactivation probabilities are then derived from these ratings. **B**: Experience tuples contain two states: the current state and next state. In the default mode, the current state of the currently reactivated experience (violet circle) is compared to the current states of all stored experiences (blue arrows) to compute the experience similarity $D(e|e_t)$. In the reverse mode, the current state of the currently reactivated experience is compared to the next state of all stored experiences (red arrows). **C**: Example of the similarity of experiences to the currently reactivated experience (green arrow) in an open field for the default and reverse modes. Experience similarity is indicated by the colorbar. In the default mode, the most similar experiences are the current state or those nearby. In the reverse mode, the most similar experiences are those that lead to the current state.

sharing the same starting state s_t, i.e. inhibition is set to $I(e) = 1$, and decays in each timestep by a factor of $0 < \lambda < 1$. Inhibition values are maintained for one replay epoch and are reset when a new replay epoch is initiated.

The next experience to be reactivated is randomly chosen according to the activation probabilities $P(e|e_t)$, which are computed by applying a customized softmax function on $R(e|e_t)$ (see Methods). Sequences of replay are produced by iteratively activating and inhibiting individual experiences. Together, experience strengths and experience similarities guide replay while inhibition promotes the propagation of sequences.

To initiate replay we consider two cases: replay during awake resting states (online replay) is initiated at the agent's current position, whereas replay during sleep (offline replay) is initiated with an experience randomly selected from memory based on the relative experience strengths.

SFMA Facilitates Spatial Learning

We begin by asking what benefit a replay mechanism such as implemented in SFMA might have for spatial learning. To do so, we set up three goal-directed navigation tasks of increasing difficulty: a linear track, a square open field, and a labyrinth maze (Fig. 3A). Simulations were run for the default and reverse replay modes and compared to an agent trained without experience replay, an agent trained with random experience replay, and the state-of-the-art Prioritized Memory Access model (PMA) by Mattar and Daw (2018).

Using the reverse mode, our model clearly outperformed the agents trained without replay and with random replay (Fig. 3B). Importantly, performance was close to, even if slightly below, that of PMA. However, PMA requires accessing all experiences to determine how much they would improve the agent's policy, called the utility, and then reactivating the experience with the highest utility. The computation of the potential benefit of reactivating an experience is computationally infeasible in a biological neural network, a problem SFMA does not suffer from.

Learning performance of SFMA critically depends on the replay mode. The default mode yielded a performance that was much lower than the reverse mode and only slightly better than random replay. Considering its low performance, the default mode may appear undesirable, however, it is important for two reasons. First, it generates different types of sequences from the reverse mode (Fig. 3D and Supplementary Figure 1) and these sequences are more consistent with some experimental observations. While the reverse mode mostly generates reverse sequences in all situations, the default mode generates forward replay sequences at the beginning of trials and reverse sequences at the end of trials. This pattern of changing replay directions has been observed in experiments in familiar environments (Diba and Buzsáki, 2007). Also, as we will show below the default mode generates shortcuts replays as those found by Gupta et al. (2010), whereas the reverse mode does not. Second, the recent discovery of so-called pessimistic replay, i.e., replay of experiences which are not optimal from a reinforcement learning perspective, show that suboptimal replay sequences occur in human brains (Eldar et al., 2020) for a good reason (Antonov et al., 2022).

We therefore suggest that the reverse and default modes play distinct roles during the learning progress. Early in the learning session, when the environment is novel and TD errors are high, the reverse mode is preferentially used to acquire a successful behavior quickly. Indeed, in our simulations the reverse mode produces more updates that were optimal than did the default mode (Fig. 3C). The preponderance of reverse replay sequences during early learning is consistent with experimental observations in novel environments (Foster and Wilson, 2006) or after reward changes (Ambrose et al., 2016). Later in the learning session, when the environment has become familiar and TD errors are low, the default mode is preferentially used to make the learned behavior more robust. The default mode then accounts for mixed reverse and forward replay sequences in familiar environments. We provide a more detailed rationale for the default mode in the Discussions. We call this strategy of switching between the reverse and default modes the *dynamic* mode of SFMA. It yields a learning performance (Fig. 3B) and numbers of optimal updates (Fig. 3C) that are similar to the reverse mode.

In the following, we focus on the statistics of replay that SFMA generates in a range of different experimental settings.

Near-Homogeneous Exploration of an Open Environment Explains Random Walk Replays

We first investigated the replay statistics that our model produces in a simple environment without navigational goals similar to the experiment of Stella et al. (2019). To this end, we created a virtual square grid world environment of size 100×100 without rewards. For simplicity, we first set experience strengths to the same value, $C(e) = 1$, to reflect homogeneous exploration of the environment. All replays were generated using the default mode since the environment was familiar to the animal in the experiments,

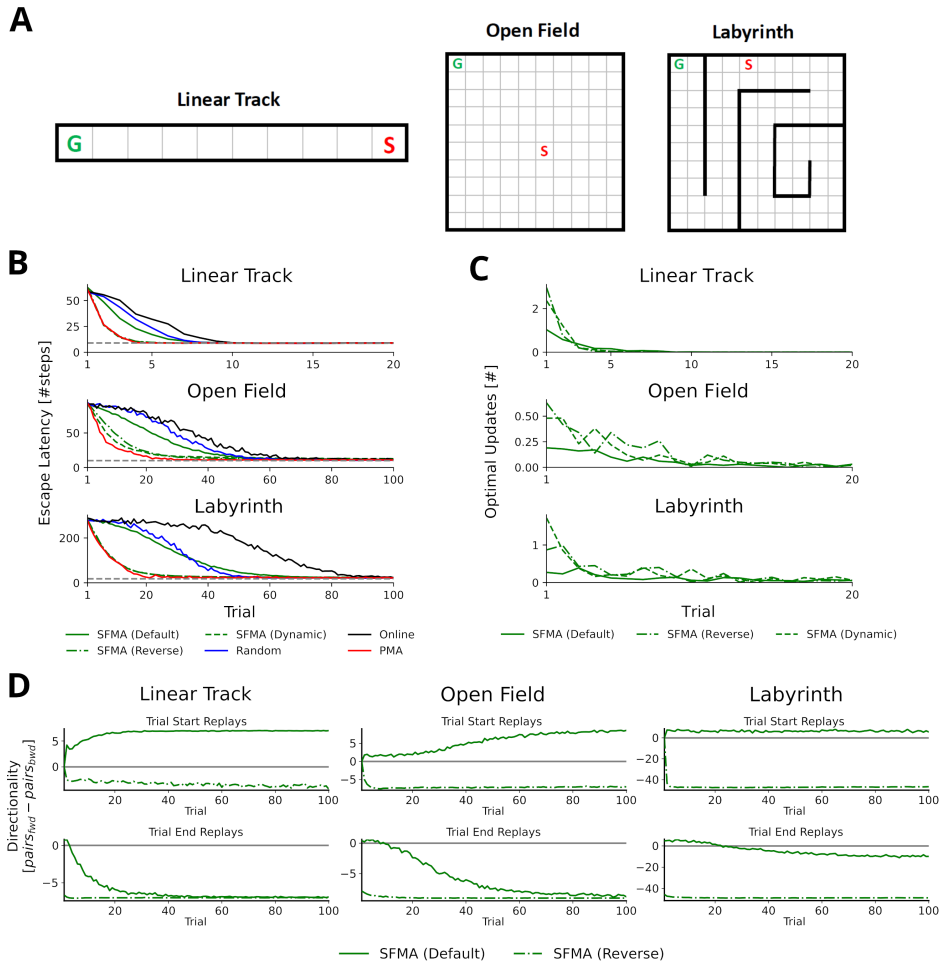

Figure 3. Statistics of replay has large impact on spatial learning. A: The three goal-directed navigation tasks that were used to test the effect of replay on learning: linear track, open field and maze. In each trial, the agent start at a fixed starting location S and has to reach the fixed goal location G. **B**: Performance for different agents measured as the escape latency over trials. Shown is the performance for an online agent without replay (black), an agent that was trained with random replay (blue), our SFMA model (green), and the Prioritized Memory Access model (PMA) by Mattar and Daw (2018) (red). The results of our SFMA model are further subdivided by the replay modes: default (solid), reverse (dash-dotted), and dynamic (dashed). Where the dashed and dash-dotted green lines are not visible they are overlapped by the red solid line. **C**: Reverse and dynamic replay modes produce more optimal replays while the default replay mode yields pessimistic replays. Shown is the number of optimal updates in the replays generated on each trial for different replay modes: default (solid), reverse (dash-dotted), and dynamic (dashed). Note, that in later trials there is a lack of optimal updates because the learned policy is close to the optimal one and any further updates have little utility. **D**: Directionality of replay produced with the default (solid) and reverse (dash-dotted) modes in the three environments. Directionality was measured at the difference between the number of forward (i.e. the next state in e_{t-1} is the current state in e_t) and reverse (i.e. the current state in e_{t-1} is the next state in e_t) pairs. The reverse replay mode produces replays with strong reverse directionality irrespective of whether replay was initiated at the start or end of a trial. In contrast the default mode produces replays with a small preference for forward directionality. After sufficient experience with the environment the directionality of replays is predominantly forward for replays initiated at the start and predominantly reverse for replays initiated at the end of a trial.

but using the reverse mode did not affect the results (Supplementary Figure 2).

With this setup the trajectories represented by the delayed experiences of the agent are visually similar to random walks (Fig. 4A). The displacement distribution of

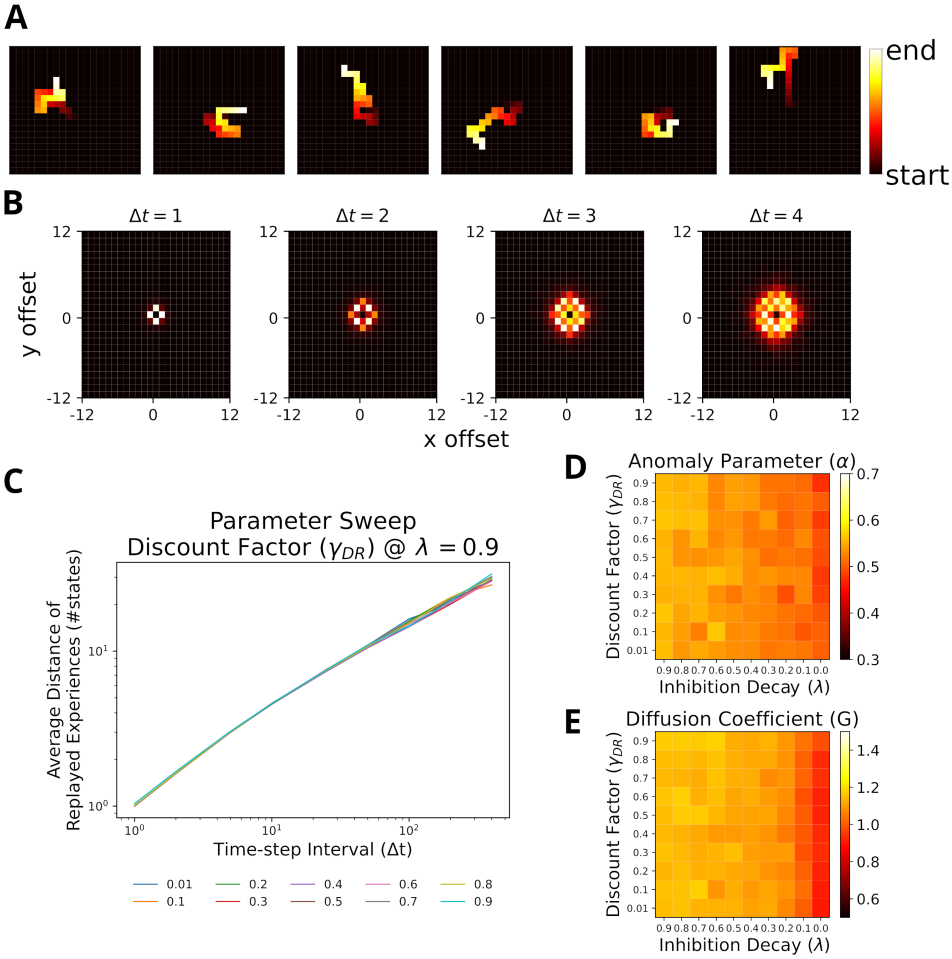

Figure 4. Replays resemble random walks across different parameter values for Default Representation (DR) discount factor and inhibition decay. A: Example replay sequences produced by our model. Reactivated locations are colored according to recency. **B**: Displacement distributions for four time steps (generated with $\beta_M = 5$). **C**: A linear relationship in the log-log plot between average distance of replayed experiences and time-step interval indicates a power-law. Lines correspond to different values of the DR's discount factor γ_{DR} as indicated by the legend. **D**: The anomaly parameters (exponent α of power law) for different parameter values of DR and inhibition decay. Faster decay of inhibition, which allows replay to return to the same location more quickly, yields anomaly parameters that closer resemble a Brownian diffusion process, i.e. closer to 0.5. **E**: The diffusion coefficients for different parameter values of DR and inhibition decay. Slower decay of inhibition yields higher diffusion coefficients.

replayed locations indicates that replay slowly diffuses from its starting location (Fig. 4B and Supplementary Figure 3). To analyze the trajectories more systematically, we used the Brownian Diffusion Analysis also used by Stella et al. In this analysis, a random walk is described by a power law relationship between the average distance between two replayed positions Δx and the time interval Δt, i.e., $\Delta x = G\Delta t^{\alpha}$ with $\alpha = 0.5$. Indeed, the simulated replays exhibits a linear relationship in the log-log-plot indicating a power law between the two variables (Fig. 4C). and the slope is close to the theoretical value for a random walk $\alpha = 0.5$. This result is robust across a large range of model parameters, the most relevant are the DR's discount factor γ_{DR} and inhibition decay λ, and even the range in our simulations, $\alpha \in [0.467, 0.574]$ (Fig. 4D), are a good match to the values reported by Stella et al. ($\alpha \in [0.45, 0.53]$). The values for the diffusion coefficient (Fig. 4E), which relate to the reactivation speed, are similarly robust and only affected when the decay factor is close to zero. Hence, our model

robustly reproduce the experimental findings.

Since the results above were obtained assuming homogeneous explorations of the environment, we repeated our simulations with heterogeneous experience strengths (Supplementary Figure 4B). While the relationships in the log-log-plot seems to slightly deviate from linear for very large time-step intervals, the statistics still largely resemble a random walk (Supplementary Figure 4C). The range of values for $\alpha \in [0.410, 0.539]$ still covers the experimental observations, albeit shifted towards smaller values.

Stella et al. further reported that the starting locations of replay were randomly distributed across the environment and replay exhibited no preferred direction. Our model reproduces similar results in the case of homogeneous experience strengths (Supplementary Figure 5A & 5B) and heterogeneous experience strengths (Supplementary Figure 5C & 5D). The results of our simulations suggests that replay resembling random walks can be accounted for given near-homogeneous exploration of an open-field environment. If exploration is non-homogeneous, the statistics of a random walk hold only for short to medium time-step intervals.

Stochasticity Results in Shortcut Replays Following Stereotypical Behavior

Gupta et al. (2010) provided further evidence that replay does not simply reactivated previously experienced sequences by showing that during replay sequences occur that represent trajectories that animals were prevented from taking. These so-called shortcut sequences were synthesized from previously experienced trajectories. We constructed a simplified virtual version of the Gupta et al. experiment (Fig. 5A) to test whether and under which conditions our proposed mechanism can produce shortcut replays. In the experiment, animals exhibited very stereotypical behavior, i.e. running laps in one direction without running back. This kind of behavior is however not easily reproducible by a simple RL agent, which initially selects actions randomly. Therefore, the agent was forced to run one of three predefined patterns: right laps, alternating laps, and left laps (Fig. 5B). This allowed us to focus on the effect of different specific behavioral statistics on replay. The virtual agent was made to run 20 trials in one session and replays were simulated after each trial.

In the first 10 trials, the agents ran in one pattern, and in the next 10 trials, the agents used the same or a different pattern. To model Gupta et al.'s study, we ran simulations with the combinations right-left (RL), right-alternating (RA), and alternating-left (AL). In addition, the combination alternating-alternating (AA) served as a control condition.

In this case only the default mode could reproduce the experimental findings, while the reverse mode could not. The main cause lay within the large dissimilarity between the experiences at the decision point (D in Fig. 5A)and the experiences to the left and right of it, which in turn drove prioritization values to zero.

We found that our model produces shortcut-like replays for each run pattern combination (Figs. 5C & 5D). Shortcuts occurred in higher numbers for right-left and right-alternating combinations, and mainly in the last 10 trials in trials when the agent was running a left lap, i.e. every trial for right-left and every other trial for right-alternating (Supplementary Figure 6). For alternating-alternating and alternating-left combinations shortcuts occurred at lower number, but across the whole running session. This difference between combinations results from the balance of experience strengths associated with either of the laps and the center piece. For right-alternating and right-left combinations the experience strengths associated with right lap and center piece were similar and therefore replay, when initiated at the left lap's reward location, was not biased to reactivate experiences along the center piece. The number of shortcut replays then decreased as the experience strengths associated with the center piece increased relative to those associated with either of the laps. Since for alternating-left and alternating-alternating combinations the experience strengths associated with center piece already were higher relative to those associated with the laps an overall low but constant number of shortcut replays was produced.

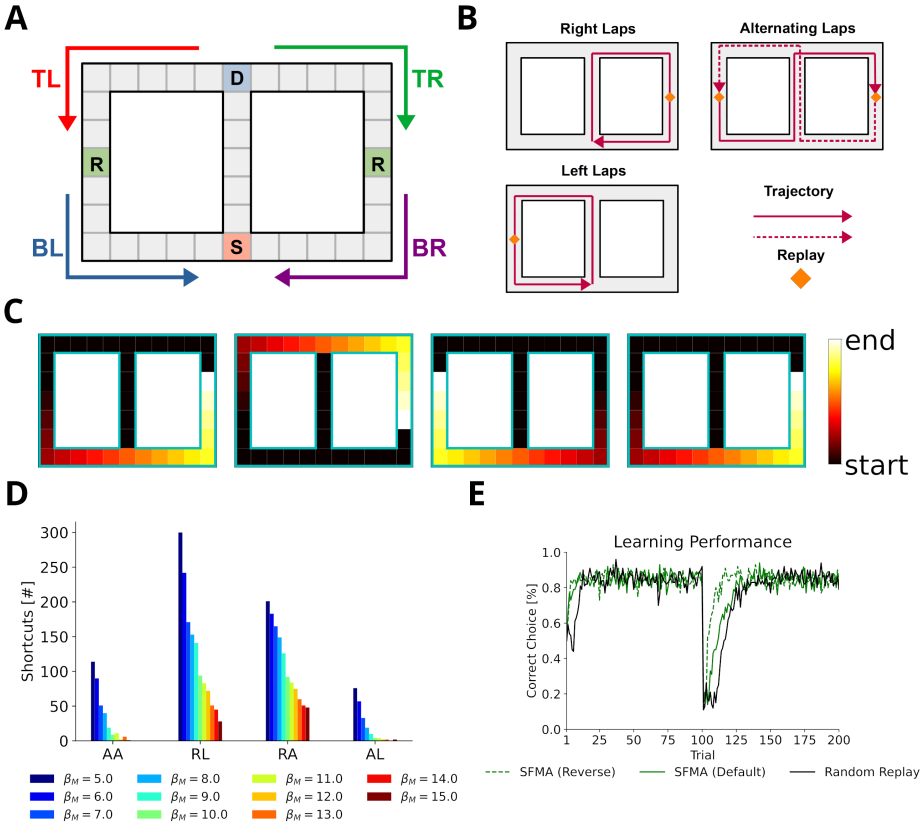

Figure 5. Replay of shortcuts results from stochastic selection of experiences and the difference in relative experience strengths. A: Simplified virtual version of the maze used by Gupta et al. (2010). The agent was provided with reward at specific locations on each lap (marked with an R). Trials started at bottom of the center corridor (marked with an S). At the decision point (marked with a D) the agent had to choose to turn left or right in our simulations. **B**: Running patterns for the agent used to model different experimental conditions: left laps, right laps and alternating laps. Replays were recorded at the reward locations. **C**: Examples of shortcut replays produced by our model. Reactivated locations are colored according to recency. **D**: The number of shortcut replays pooled over trials for different running conditions and values of the inverse temperature β_M. **E**: Learning performance for different replay modes compared to random replay. The agent's choice depended on the Q values at the decision point.

Across all conditions the number of shortcut replays decreased with increasing inverse temperature β_M, which transforms priority ratings to reactivation probabilities. The larger β_M, the more deterministic replay becomes. This result therefore suggest that shortcut replay is driven by stochasticity in the prioritization process and differences in relative experience strengths (Fig. 5D) of the maze's center piece and arms. These differences are in turn induced by the specific behavioral statistics.

Finally, we wanted to know if replay facilitates learning in this task. Since the agent cannot remember which lap it ran in a previous trial, it would not be able to learn with alternating rewarded locations. Therefore, we chose the right-left condition as the learning task. To deal with problematic behavior of simple RL agents mentioned above we adjusted our simulation to only consider the action selected at the decision point and let the agent run straight otherwise. We ran simulations in default and reverse mode. An agent trained with random experience replay was used as baseline. All agents solved the task and successfully learned the new goal location after the rewarded locations changed from right to left (Fig. 5E). Both default and reverse mode yielded faster learning than random experience replay. However, there was no large difference between the replay modes during the first half. These small differences in learning performance suggest that replaying strictly in reverse mode may not be necessary in tasks with

stereotypical behavior since the resulting experience strengths would naturally guide replays in forward or reverse depending on where replay started (Supplementary Figure 7).

Dynamic Structural Modulation Allows Replay to Adapt to Environmental Changes

Recently, Widloski and Foster (2022) reported that hippocampal replay dynamically adapts to daily changes in the structure of the environment. We argue that this type of replay can be accounted for by representing the environmental structure dynamically with the default representation (DR) (Piray and Daw, 2021). While the SR (Dayan, 1993) could also be used to represent the environment's structure, it cannot be updated efficiently in the face of structural changes and depends on the agent's policy. The DR does not suffer from these problems. Specifically, we employed the DR by separating the spatial representation into open 2D space and information about environmental barriers. Barrier information was provided when the environment changed (see Methods).

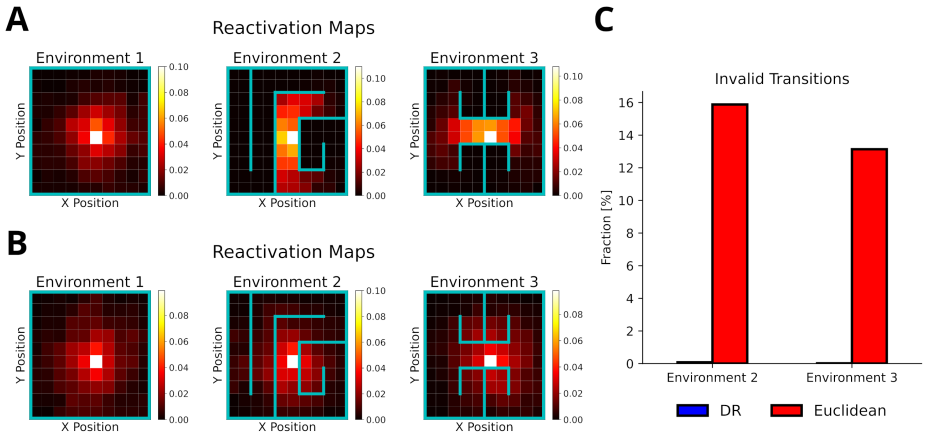

Figure 6. The default representation (DR) allows replay to adapt to environmental changes. A: Reactivation maps for replays recorded in three environments while the agent is located at the white square. Experience similarity is based on the DR with a discount factor $\gamma_{DR} = 0.1$. The colorbar indicates the fraction of reactivations across all replays. Note that replay obeys the current environmental boundaries. **B**: Same as in A, but experience similarity is based on the Euclidean distance between states. Note that replay ignores the boundaries. **C**: The fraction of invalid transitions during replay in different environments for the DR (blue) and Euclidean distance (red). While replay can adapt to environmental changes when using the DR, it does not when experience similarity is based on the Euclidean distance.

To test whether our proposed replay mechanism can adapt to changes of the environment's structure, we simulated replays in three distinct consecutive environments (Fig. 6A) and evaluated the occurrence of replay crossing environmental barriers. All replays were generated using the default mode, but using the reverse mode did not affect these results (Supplementary Figure 8). Our proposed replay mechanism successfully adapts to changes in the environment's structure as illustrated by replays diffusing along barriers in the reactivation maps (Fig. 6A) and shown by the absence of invalid transitions (Fig. 6C). In contrast, when experience similarity is based on the Euclidean distance between locations, replay does not conform or adapt to the changing environmental structure (Fig. 6B, C).

Preplay of Unvisited Locations Requires Structural Modulation and Visual Exploration

Ólafsdóttir et al. (2015) reported the preplay of trajectories through a part of a T-maze that the animal could see, but not physically explore, and that contained a visible reward. Our model can provide an intuitive explanation for the reported phenomenon.

First, the use of DR explains why the preplay of unvisited locations is possible in principle. The DR allows for the representation of the environment's structure provided it is observable. Still, these experiences would have zero experience strength in our model and therefore would not be reactivated. However, during the reward cueing phase, the animals were reported to have paid attention to the reward cue. We view this as a form of visual exploration that increases the strength of experiences along the attended arm and updates the reward information related to the experiences.

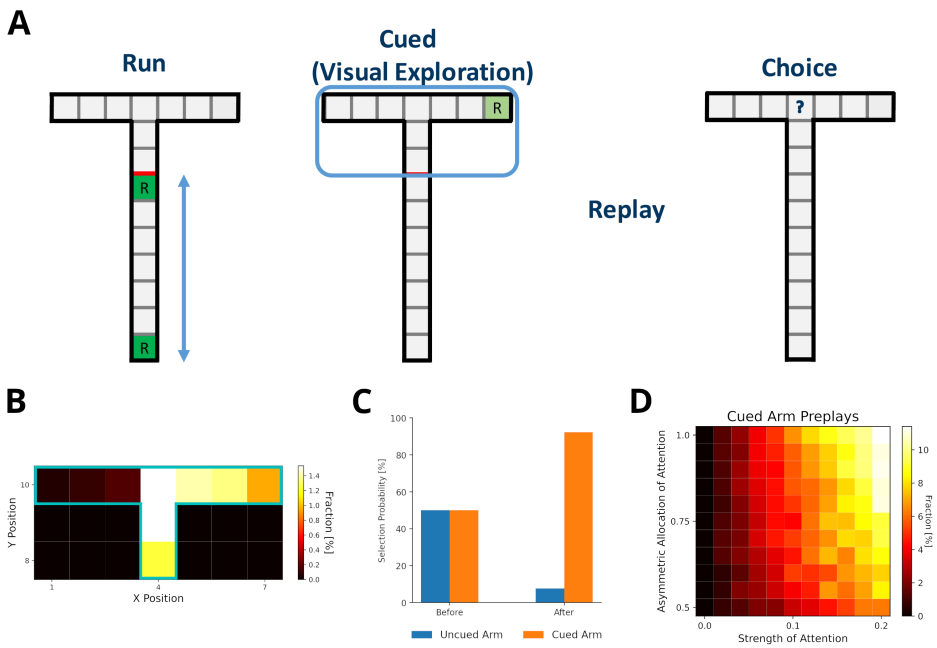

Figure 7. Preplay of cued, but unvisited, locations can be explained by visual exploration. A: Schematic of the simplified virtual maze and task design used to model study by Ólafsdóttir et al. (2015). First, in an initial run phase, the agent ran up and down the stem of the maze. Second, in the following visual exploration phase, the agent visually explores a cued part of the maze and increases the strength of the associated experiences. Third, experiences are reactivated according to their priority scores. Fourth, agents are given access to the entire T-maze. The fraction of choosing the two arms in the test phase are compared. **B**: Fractions of reactivated arm locations. Experiences associated with the cued arm (right) are preferentially reactivated. **C**: The fractions of choosing either arm before and after training the agent with preplay. Before preplay the agent shows no preference for one arm over the other. The agent preferentially selects the cued arm over the uncued arm after preplay. **D**: The percentage of cued-arm preplay (out of all arm preplays and stem replays) for different amounts of attention paid to the cued arm vs. the uncued arm (asymmetric allocation of attention) and experience strength increase relative to actual exploration (attention strength).

We set up a virtual version of the Ólafsdóttir et al. task (Fig. 7A) in which the agent first runs up and down the maze stem ten times with rewards provided at each end of the stem (run phase). In the cued phase, we increased experience strengths of the blocked arms by a small amount relative to actual exploration (attention strength), which was larger for the cued than the uncued arm (asymmetric allocation of attention). Then experiences were preplayed. Template matching was used to detect the preplay of cued and uncued arms, and replay of the stem. All preplays were generated using the default mode, but results were similar when using the reverse mode (Supplementary Figure 9). We found that our model indeed produces preplays of the cued arm and only rarely of the uncued arm (Fig. 7B). As expected, the attention strength as well as the asymmetric allocation of attention are key to produce preplays (Fig. 7D). The higher the value of these two parameters, the more often the cued arm is preplayed.

Furthermore, Ólafsdóttir et al. reported that animals expressed a behavioral preference for the cued arm when they were given physical access to the maze and the

barrier blocking the arms was removed. Our models shows the same preference for the cued arm after training an agent with simulated replays and testing it in the T-maze (7C).

Discussion

We introduced a replay mechanism (SFMA) that accounts for a variety of different replay statistics reported in the literature. The core aspect of SFMA is the prioritization of experiences according to the frequency at which they occur, the likelihood with which they are rewarded, and the structure of the environment. Furthermore, reactivated experiences are inhibited in the prioritization process to drive the generation of sequential activations. The statistics of these sequences depends on whether SFMA operates in the default or reverse mode. Surprisingly, despite these differences both modes are consistent with a diverse experimental results – with the notable exception of shortcut replays, which arise only in the default mode. However, the two modes have very different impact on spatial learning, which is much faster when using the reverse mode. In the following we discuss some details of our modeling assumptions and results.

The Components of the Spatially and Frequency-weighted Memory Access (SFMA) Model

The prioritization of experiences during replay is well supported by experimental evidence. Behavioral data suggests that animals use spatial representations to quickly adapt to changes in the environment (de Cothi et al., 2020). Structural modulation of hippocampal replay is reported in the literature (Wu and Foster, 2014; Ólafsdóttir et al., 2015; Widloski and Foster, 2022). The hippocampus and related structures, e.g. entorhinal cortex (EC), are commonly known for containing cells that selectively fire for spatial features (O'Keefe and Dostrovsky, 1971; Hafting et al., 2005; O'Keefe and Burgess, 1996). We chose to represent the environment's structure with the default representation (DR) (Piray and Daw, 2021), which is a special case of the successor representation (Dayan, 1993). The latter has been proposed to explain firing properties of place and grid cells (Stachenfeld et al., 2017), and the DR gave an account for boundary cells in EC. Furthermore, we separated the representation of space into a general representation of 2D space and information about environmental barriers. General representations of task structure in EC have been proposed to be forming hippocampal place cell code together with sensory information (Whittington et al., 2020).

The frequency of experience (Kudrimoti et al., 1999; O'Neill et al., 2008) as well as reward (Singer and Frank, 2009; Pfeiffer and Foster, 2013) have been shown to influence the content of replay. Experience has been incorporated in other models. For instance, the Mattar and Daw (2018) model uses the Successor Representation which depends on the agent's previous behavior.

While the main purpose of inhibition is that it enables the generation of sequences, it can be also viewed as a way of shifting priorities from experiences that were already using during learning to other experiences and, hence, ensuring the propagation of reward information. An interesting question pertains to whether a similar inhibition mechanism is utilized in the hippocampus. It was shown in modeling studies that the generation of sequences in a network does not strictly require inhibition (Bayati et al., 2015). However, forms of inhibition have been used to account for replay as well as preplay sequences in biologically plausible models (Chenkov et al., 2017; Azizi et al., 2013). We propose that the existence of inhibition implemented in our model could be tested by measuring the excitability of place cells following their reactivation during replay events.

The Role of the Default and Reverse Replay Modes

While the default mode accounts for all varieties of replay statistics considered, it performs considerably worse during learning than the reverse mode since it does not produce reverse sequences as reliably. Reverse sequences are important for learning because they can propagate rewards information back to preceding actions that led to the reward (Foster and Wilson, 2006). However, when replay is initiated at the start of a trial, the default mode tends to reactivate sequences with an overall forward direction. In principle, it is possible to consistently generate reverse sequences at the end of a trial in the default mode, but that requires sufficient experience with the environment. In contrast, the reverse mode will always produce reverse sequences irrespective of whether replays are initiated at the start or end of a trial, and does not require extensive experience with the environment. A downside of the reverse mode is that those state-action pairs that lead to the recently reactivated experience have to be known.

To benefit from the advantages of both replay modes, we propose that they could be play different roles during learning. There is support in the literature for this view. Foster and Wilson (2006) found that reverse replay can occur after even a single experience, which is predominantly generated by the reverse mode. Furthermore, frequency of reverse replay has been observed to increase in response to positive reward changes (Ambrose et al., 2016). Hence, the reverse mode might be used preferentially when the situation required learning. By contrast, Wikenheiser and Redish (2013) reported both forward and reverse replay sequences during awake resting phases occurring in roughly equal proportions (for a whole session). In a familiar environment, Diba and Buzsáki (2007) found that replay sequences in forward order at the beginning of a trial prior to running on a linear track and in reverse order at the end of a trial after running. These results are more consistent with the default mode and more balanced replay direction in familiar environments.

What could the function of the default mode be, if it does not speed up learning much compared to random replay? A possibility could be a supportive function in the RL process. For instance, in the context of deep RL it is well known that multi-layer networks trained with sequences suffer from learning instabilities (Mnih et al., 2015). This instability is addressed by randomly replaying experiences to break the strong temporal correlations between sequential experiences. While not completely random, the default mode is more likely to break proper reverse and forward sequences. Another possible function that could be implemented by the default mode is the reactivation of experiences which otherwise would not (or only rarely) be reactivated in the reverse mode. Recent findings of pessimistic replay, i.e. the replay of sub-optimal experiences, in humans (Eldar et al., 2020) supports this. Interestingly, follow-up modeling work (Antonov et al., 2022) suggests that optimal replay occurs predominantly during initial trials and then switches to pessimistic replay. However, this switch is dependent on the *utility* of experiences (Mattar and Daw, 2018). The default mode could be viewed as enabling these pessimistic replays, and the switch from optimal to pessimistic replay corresponds to the dynamic mode in our model.

Alternatively, the default mode might not be directly related to RL, but serve a different function altogether. For instance, the sequences generated in the default mode might be important for maintaining synaptic connections which otherwise would not be maintained due to the specific distribution of reverse replays (Hebb, 1949; Caporale and Dan, 2008). Another function often attributed to hippocampal replay is that of planning (Diba and Buzsáki, 2007; Buckner, 2010; Buhry et al., 2011) by simulating possible future trajectories. The default mode could, in principle, support this function. In contrast, since the reverse mode almost exclusively produces reverse sequences, it could not support the prospective sequences needed for planning.

Related Work

Compared to other computational models that have been developed to model and understand hippocampal replay, SFMA has some desirable properties in terms of the range of experimental results explained and biological plausibility. The state-of-the-art

PMA model by Mattar and Daw (2018) proposes that replay prioritizes experiences according to their utility, which in turn is defined as the product of gain and need. In their model, gain represents the policy improvement gained if an experience were reactivated and need represents the expected future occupancy of the experience's current state. As mentioned in earlier sections, it is not clear how an agent can compute the gain without reactivating all stored experiences. In contrast, our model does not suffer from this problem, since the variables involved in its prioritization could be computed by a biological network. Inhibition of return could easily be implemented via local inhibitory recurrent connections and experience strength through strengthening of connections between cell assemblies, and Piray and Daw (2021) convingly argued the DR's connection to grid and border cells in the entorhinal cortex.

Furthermore, the implementation of PMA contains a few oddities. Because utility is defined as a product of gain and need, the gain must have a nonzero value to prevent all-zero utility values. This is achieved by applying a (small) minimum gain value. A problem arises for the special n-step update considered by their model for which gain is summed over n steps and which is the main driver of forward sequences. The minimum gain value is applied before summation. We argue that the minimum value should be applied after the summation of gain values. In addition, Mattar & Daw state that "EVB ties are broken in favor of shorter sequences", where EVB stands for "expected value of backup", i.e. the utility. However, in their implementation longer sequences are always favored in the case that gain is zero for all experiences. This is because when applying the aforementioned minimum gain value longer sequences will have higher gain than short sequences. When the computation of gain is changed to apply the minimum value after summation, the model's ability to produce sequences during offline replay is severely reduced (Supplementary Figure 10).

SFMA shares many features with PMA, so the two models could potentially be combined to mitigate their respective shortcomings. Sequences produced by our prioritization mechanism could be used to form gain estimates of reactivated experiences. These estimates could then be used to more efficiently prioritize the replays.

A different view of hippocampal replay was presented by Khamassi and Girard (2020). They argue that replay can be viewed as a form of bidirectional search in which learning is facilitated by alternating between phases of prioritized sweeping (Moore and Atkeson, 1993), which prioritizes experiences according to how their reactivation affects behavior, and trajectory sampling (Sutton and Barto, 2018), which explores potential routes by simulating trajectories through the environment. In simulations, they show that bidirectional search performs better than prioritized sweeping and trajectory sampling by themselves. Importantly, they also were able to reproduce the occurrence of shortcut replays. Their model is, in principle, able to generate random walk replays due to the fact that, in the absence of reward, replay would be dominated by the trajectory sampling phase under a uniform action policy. However, their model does give an account for how replay could adapt to environmental changes.

McNamee et al. (2021) proposed a generative model of replay which adapts replay statistics depending on the cognitive function that it supports. Unlike the models discussed above, this model focuses on states, i.e. locations, reactivated ignoring the combination of state and action. Their model produces random walks and they show that this replay statistics facilitates the learning of a successor representation of the environment. However, this model lacks a proper account for why replay should be able to quickly adapt to changes in the environment, and assumes generator for (random) sequences which is handcrafted for each simulation. Furthermore, their model does not seem to allow for differentiating forward from reverse sequence.

Conclusion

Our model gives a biologically plausible account of how the hippocampus could prioritize replay and produce a variety of different replay statistics, and fits well with contemporary views on replay. Furthermore and importantly, the replay mechanism facilitates learning in goal-directed tasks, and performs close to the current

state-of-the-art model without relying on biologically implausibly computations.

Materials and Methods

Reinforcement Learning

For our computational study we relied on reinforcement learning (RL) (Sutton and Barto, 2018) which is makes it possible to learn optimized sequences of choices based on sparse rewards. In RL, an agent interacts with an environment and tries to maximize the expected cumulative reward (Sutton and Barto, 2018):

$$R = \mathbb{E}\left[\sum_t \gamma^t r_t\right], \tag{2}$$

where γ is the so-called discount factor that weighs the importance of rewards that will be received in the future relative to immediate ones, and r_t is a real-valued reward received at time t. The RL agent must learn to choose those actions that maximizes the expected cumulative reward, the so-called optimal policy, through interactions with its environment. These interactions are generally formalized as experience tuples $e_t = (s_t, a_t, r_t, s_{t+1})$, where s_t and s_{t+1} denote the current state and next state, respectively. The action taken by the agent in state s_t is represented by a_t and the scalar reward received is represented by r_t. The critical problem in RL is crediting the right choices with leading to a reward that was received only (much) later.

A popular algorithm that solves this learning problem is Q-learning (Watkins, 1989) which uses a state-action value function $Q(s, a)$, also called the Q-function, to represent the cumulative future reward expected for taking action a in state s. The Q-function is updated from experiences in each time step t by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta\left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right] \tag{3}$$

Here, η denotes the learning rate. For all simulations the learning rate was set to $\eta = 0.9$. The term in brackets is known as the temporal difference error. RL generally requires many interactions with the environment, which results in slow learning. To address this problem Lin (1992) introduced experience replay, where interactions with the environment are stored as experiences e_t and reactivated for updating the Q-function later. In its simplest form, experience replay samples stored experiences uniformly at random. Learning from replay can be further improved by prioritizing experiences to be replayed (Sutton and Barto, 2018; Schaul et al., 2016).

Spatial structure and Frequency-weighted Memory Access (SFMA)

For our simulations, we employed the Dyna-Q architecture in combination with grid world environments in a similar fashion to Mattar and Daw (2018). We view the reactivation of specific experiences as analogous to the reactivation of place cells in the hippocampus during awake and sleeping resting states. Like the original Dyna-Q architecture, our model stored experiences for all environmental transitions in a memory module. However, we stored additional variables for later prioritization of experiences during replay: The experience strength $C(e_i)$, the experience similarity $D(e_i|e_j)$ and inhibition of return $I(e_i)$. The experience strength is a count of how often a certain environmental transition was experienced. Following an experienced transition e_i the experience strength was updated as follows:

$$C(e_i) \leftarrow C(e_i) + 1 \tag{4}$$

Additionally, experience strength was modulated by rewards experienced during behavior. If reward was encountered the strength of experiences e_i was increased

weighted by their similarity to e_r according to the following update rule:

$$C(e_i) \leftarrow C(e_i) + rD(e_i|e_r) \tag{5}$$

where r is the scalar reward encountered and e_r is the corresponding experience. The similarity between two experiences e_i and e_j was encoded by $D(e_i|e_j)$ and was based on the Default Representation (DR) (Piray and Daw, 2021). Specifically, we used the DR to represent open space and barrier information separately. The DR representing open space was computed directly from the state-state transition model T which represents the default policy:

$$D = (X - \gamma_{DR}T)^{-1} \tag{6}$$

Here, X is the identity matrix and γ_{DR} is the DR's discount factor. The DR's discount factor was set to $\gamma_{DR} = 0.1$ unless stated otherwise. Barrier information was provided during initialization as a list of invalid transitions and could be updated in case of environmental changes. The low rank matrix update of the DR was computed for these invalid transitions. Inhibition $I(e_i)$ was used to prevent the repeated reactivation of the same experience and applied following reactivation of an experience to all experiences e_i that share the same starting state s_t:

$$I(e_i) \leftarrow 1 \tag{7}$$

Inhibition terms decayed in each timestep with a factor λ:

$$I(e_i) \leftarrow \lambda I(e_i) \tag{8}$$

Inhibition was reset between replay epochs and the decay factor was set to $\lambda = 0.9$ unless stated otherwise.

Replays were initiated in two different ways depending on whether they were simulated for awake resting states or sleep. Awake replay was initiated at the agent's current state, sleep replay was initiated in a randomly chosen state based on the relative experience strengths:

$$p(e) = \frac{C(e)}{\sum_t C(e_i)} \tag{9}$$

Given the most recently reactivated experience e_t, an experience e was stochastically reactivated based on the priority ratings $R(e|e_t)$:

$$R(e|e_t) = C(e)D(e|e_t)[1 - I(e)] \tag{10}$$

The activation probabilities $P(e|e_t)$ were computed by applying a customized softmax function to the normalized priority ratings:

$$P(R_i) = \frac{e^{\beta_M R_i} - 1}{\sum_t (e^{\beta_M R_j} - 1)} \tag{11}$$

where β_M is the inverse temperature. $\beta_M = 9$, unless stated otherwise. Since in the usual softmax function priority ratings of zero would yield non-zero reactivation probabilities, we subtracted one from each exponential in our customized softmax function. The replay epoch ended, i.e. replay stopped, when either a maximum number N of replayed experiences was reached or the priority ratings for all experiences were below a threshold $\theta = 10^{-6}$.

Since each experience tuple contains two states, i.e. the current and next state, that could be compared to one another to determine similarity of experiences $D(e|e_t)$, we explored two different replay modes. The default mode, which computes similarity based on the two current states, and the reverse mode, which compares the next state of experience e to the current state of experience e_t (Fig. 2B).

Model Implementation

We implemented our model as well as all of our simulations and analysis code in Python 3 using the CoBeL-RL framework (Walther et al., 2021) (`https://doi.org/10.5281/zenodo.5741291`).

Algorithm 1 Spatially and Frequency-weighted Memory Access (SFMA)

Require: e_t (Replay initiated)
for t = 1:N **do**
  for experience e **do**
    Compute priority rating $R(e|e_t) = C(e)D(e|e_t)[1 - I(e)]$.
  end for
  if $\max R < \theta$ **then**
    Stop replay.
  end if
  Compute reactivation probabilities $P(e|e_t)$.
  Choose next experience e_{t+1} to reactivate.
  Reduce/decay inhibition for all stored experiences.
  Inhibit experience e_{t+1}.
end for

Simulation: Navigation Tasks

Spatial learning performance was tested in three grid world environments: a linear track of size 10×1 states, an open field of size 10×10 states and a labyrinth of size 10×10 states. Each environment had one specific goal, which yielded a reward of 1, and starting state. We trained agents which use SFMA in default, reverse, and dynamic modes. The discount factor used for learning was $\gamma = 0.99$. In the last mode, the agent dynamically chose either the default or reverse mode based on the cumulative temporal-difference error experienced in the current trial. The higher the cumulative temporal-difference error was, the higher was the probability of choosing the reverse replay mode. For comparison we also trained agents that learned from online experience only, with random replay, and Prioritized Memory Access (Mattar and Daw, 2018) (PMA). To account for the different sizes and structure of the environments, the number of steps per trial as well as the length of replays were different for each environment (Table 1). Since the model by Mattar and Daw (2018) uses additional replay at the start of each trial, we split the replay length in half. Furthermore, during action selection we masked actions which would lead the agent to same state. Reasons for this were twofold: First, by masking action we sped up learning for all agents by removing pointless actions. Second, it mitigated the unfair advantage of PMA which in its prioritization ignores experiences which lead into the same state. To test the optimality of SFMA's different replay modes we computed for each step in each replay whether the reactivated experience was optimal, that is, it had the highest utility as defined by Mattar and Daw (2018). For simplicity we assumed uniform need as Antonov et al. (2022) did.

Environment	Trials	Steps per Trial	Replay Length
Linear Track	20	100	10
Open Field	100	100	10
Labyrinth	100	300	50

Table 1. Training settings for each task.

We further analyzed the directionality of replay in the default and reverse modes depending on when in a trial it was initiated, i.e. start and end of a trial. Simulations were repeated with the same training settings listed above except that the number of trials for the linear track was extended to 100. Furthermore, additional replays were generated at the start and end of each trial but they did not contribute to learning. The directionality of these additional replays was measured by the difference between the number of forward and reverse pairs of consecutively reactivated experiences e_{t-1} and e_t. A pair e_{t-1} and e_t was considered to be forward when the next state of e_{t-1} was the current state of e_t, reverse when the current state of e_{t-1} was the next state of e_t and unordered otherwise.

Simulation: Random Walk Replay

To replicate the experiment by Stella et al. (2019) we chose an open-field grid world with 100×100 states. The large size of this environment allowed us to simulate replays of different lengths. We considered two cases: Homogeneous experience strengths and heterogeneous experience strengths, which were highest at the center of environment and decreased with distance to half the maximum value at the borders. In each simulation a total 50 replays of length 500 were generated. At such a large length, replays can easily reach the environment's borders if initialized near the borders. We therefore initialized replays exclusively at the environment's center. Simulations were repeated for different values of DR discount factor γ_{DR} and inhibition decay λ:

$$\gamma_{DR} \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\} \tag{12}$$

$$\lambda \in \{0.00, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\} \tag{13}$$

We applied the Brownian Diffusion Analysis used by Stella et al. (2019) to the output of our simulations. Furthermore, we tested the randomness of replay starting positions as well as their directions. We did so by initializing 1000 replays and restricted replay length to five. Values for the DR discount factor and inhibition decay were fixed at $\gamma_{DR} = 0.9$ and $\lambda = 0.9$, respectively. The distribution of starting positions was computed by counting how often a given state occurred as the current state of initially replayed experiences. For the distribution of initial replay directions we first divided the environment into spatial bins of size 20 by 20 states. Direction was divided into 90° bins, i.e. bins centered at 0°, 90°, 180° and 270°. For each spatial bin we then counted how often replay direction fell into respective direction bins. Lastly, we computed displacement distributions at different time steps, i.e. $\Delta t = 1, 2, 3, 4$, relative to initially replayed positions in a 25×25 states window. This was done by averaging the replayed position relative to the start of replay.

Simulation: Shortcut Replay

For this simulation we simplified the maze used by Gupta et al. (2010) as an 11×7 states figure-eight-maze-like grid world. Reward locations on either side were reduced from two to one location. Since our main interest lay in the effect of specific stereotypical running behavior expressed by the animals on replay, basic running patterns were predefined:

1. From starting position to left side's reward location.
2. From starting position to right side's reward location.
3. From left side's reward location to left side's reward location.
4. From right side's reward location to right side's reward location.
5. From left side's reward location to right side's reward location.
6. From right side's reward location to left side's reward location.

From these running patterns we created sessions consisting of 20 trials. Each session was characterized by one pattern during the first 10 trials, which could change to a different pattern for the remaining 10 trials, e.g. 10 right laps followed by 10 left laps. Running patterns 1 and 2 were only used in the initial trial of a session. Patterns 3 and 4 were used when the agent was supposed to run laps on the same side, and patterns 5 and 6 were used when the agent was supposed to alternate between the two reward locations. We considered the following pattern combinations (used patterns are listed in parenthesis ordered numerically):

- **RL**: Right-Left (from patterns 2, 3, 4 and 6).
- **RA**: Right-Alternating (from patterns 2, 4, 5 and 6).

- **AL**: Alternating-Left (from patterns 2, 3, 5 and 6).
- **AA**: Alternating-Alternating (Control Condition; from patterns 2, 4, 5 and 6).

After each trial 200 replays were generated and stored for later analysis. Since reactivation probabilities are controlled by the softmax function's inverse temperature parameter β_M, we also repeated each simulation for a range of values, i.e. $\beta_M \in \{5, 6, ..., 15\}$. To detect shortcut replays we first generated state sequence templates for each corner of the maze (i.e. TL, TR, BL and BR) in forward and backward direction. A template was considered a match if the number of mismatches was at most two. If two corners that together would form a shortcut (e.g. TL in backward direction and TR in forward direction) were detected in a replay and were separated by 8 ± 2 states they were considered to form a shortcut. The different possible shortcuts were:

- From TL (backward) to TR (forward).
- From TR (backward) to TL (forward).
- From BL (forward) to BR (backward).
- From BR (forward) to BL (backward).

For our additional learning simulations an RL agent was trained in the environment. Since a simple RL agent like the one employed is not capable of expressing the same stereotypical running behavior, we restricted selectable actions such that the agent can only run forward. The only relevant choice was therefore limited to the maze's decision point. If the agent picked the correct choice it received a reward of 1 for that trial. Furthermore, since the agent is not capable of learning in conditions with alternating reward arms for lack of working memory, we restricted simulations to the Right-Left condition. Simulations were repeated for agents trained with SFMA in default and reverse mode as well as random replay. To better differentiate the effect of replay type, we allowed but one replay after each trial and increased the number of trials per session to 200. The discount factor used for learning was $\gamma = 0.99$.

Simulation: Adaptive Replay

To evaluate our model's ability to adapt to structural changes in an environment we set up three structurally different environments and presented them successively to our model. Replays were simulated in each environment and analyzed by computing the fraction of invalid transitions. Invalid transitions were defined as two consecutively replayed experiences whose current states were separated by a barrier. To evaluate the importance of the representation chosen for experience similarity we performed all simulations using the Euclidean distance for the experience similarity, i.e.,

$$D(e_i|e_j) = \exp\left(-\|s_i - s_j\|\right) \tag{14}$$

Simulation: Preplay

We simplified the experimental setup from Ólafsdóttir et al. (2015) as a 5×7 T-maze-like grid world. A barrier was inserted to separate stem from arms. The simulation was divided into three phases. First, in the run phase the agent was made to run up and down the stem, and rewards were provided at the two ends of the stem. Second, in the cue phase we replicated the attention paid to the arms by the animals as an increase of experience strengths at the arms. Visual exploration was associated with a lower is experience strengths than physical exploration by a factor we call attention strength. The values focused on for the strength of attention was:

$$\{0.0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2\} \tag{15}$$

The right arm was chosen arbitrarily as the cued arm that contained a reward and therefore was allocated more attention than the other, uncued arm (asymmetric allocation of attention). For the asymmetric allocation of attention, we focused on values greater than 0.5 as they best reflect the attention reported:

$$\{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 1.0\} \tag{16}$$

The increase of experience strengths for cued was computed by multiplying attention strength and asymmetry with the increase that would be expected if the agent actually traversed the cued arm. The increase of experience strengths for the uncued arm was compuated similarly but using the inverse symmetry, i.e. $1 - symmetry$. For each combination of values for attention strength and asymmetric allocation of attention we simulated 5000 preplays of length six. Preplay sequences were identified with template matching. Templates were prepared for stem, cued, and uncued arms for forward and reverse sequences. In a final choice phase, the value combination (0.75 and 0.14 for asymmetric allocation of attention and attention strength, respectively) which yielded a preplay fraction closest to the one reported by Ólafsdóttir et al. (2015), i.e. 7.37, was used to train the agent's Q-function. The discount factor used for learning was $\gamma = 0.99$. The choice at the decision point was compared between before and after training assuming a softmax action selection policy with an inverse temperature parameter of $\beta = 20$. The action that leads into a wall and the action that leads back into the stem were ignored since we wanted to focus on the agent's preference for the cued vs the uncued arm.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 419037518 – FOR 2812, P2.

References

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(5):1124–1136.

Antonov, G., Gagne, C., Eldar, E., and Dayan, P. (2022). Optimism and pessimism in optimised replay. *PLOS Computational Biology*, 18(1):e1009634.

Azizi, A. H., Wiskott, L., and Cheng, S. (2013). A computational model for preplay in the hippocampus. *Frontiers in Computational Neuroscience*, 7.

Bathellier, B., Tee, S. P., Hrovat, C., and Rumpel, S. (2013). A multiplicative reinforcement learning model capturing learning dynamics and interindividual variability in mice. *Proceedings of the National Academy of Sciences*, 110(49):19950–19955.

Bayati, M., Valizadeh, A., Abbassian, A., and Cheng, S. (2015). Self-organization of synchronous activity propagation in neuronal networks driven by local excitation. *Frontiers in Computational Neuroscience*, 9.

Buckner, R. L. (2010). The Role of the Hippocampus in Prediction and Imagination. *Annual Review of Psychology*, 61(1):27–48.

Buhry, L., Azizi, A. H., and Cheng, S. (2011). Reactivation, Replay, and Preplay: How It Might All Fit Together. *Neural Plasticity*, 2011:1–11.

Buzsáki, G. (1989). Two-stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience*, 31(3):551–570.

Caporale, N. and Dan, Y. (2008). Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience*, 31(1):25–46.

Chenkov, N., Sprekeler, H., and Kempter, R. (2017). Memory replay in balanced recurrent networks. *PLOS Computational Biology*, 13(1):e1005359.

Corkin, S., Amaral, D. G., González, R. G., Johnson, K. A., and Hyman, B. T. (1997). H. M.'s Medial Temporal Lobe Lesion: Findings from Magnetic Resonance Imaging. *The Journal of Neuroscience*, 17(10):3964–3979.

Davidson, T. J., Kloosterman, F., and Wilson, M. A. (2009). Hippocampal Replay of Extended Experience. *Neuron*, 63(4):497–507.

Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 5(4):613–624.

de Cothi, W., Nyberg, N., Griesbauer, E.-M., Ghanamé, C., Zisch, F., Lefort, J. M., Fletcher, L., Newton, C., Renaudineau, S., Bendor, D., Grieves, R., Duvelle, E., Barry, C., and Spiers, H. J. (2020). Predictive Maps in Rats and Humans for Spatial Navigation. preprint, Animal Behavior and Cognition.

Deacon, R. M., Bannerman, D. M., Kirby, B. P., Croucher, A., and Rawlins, J. P. (2002). Effects of cytotoxic hippocampal lesions in mice on a cognitive test battery. *Behavioural Brain Research*, 133(1):57–68.

Diba, K. and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, 10(10):1241–1242.

Eldar, E., Lièvre, G., Dayan, P., and Dolan, R. J. (2020). The roles of online and offline replay in planning. *eLife*, 9:e56911.

Foster, D. J. and Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680–683.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal Replay Is Not a Simple Function of Experience. *Neuron*, 65(5):695–705.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley.

Johnson, A. and Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, 18(9):1163–1171.

Khamassi, M. and Girard, B. (2020). Modeling awake hippocampal reactivations with model-based bidirectional search. *Biological Cybernetics*, 114(2):231–248.

Kudrimoti, H. S., Barnes, C. A., and McNaughton, B. L. (1999). Reactivation of Hippocampal Cell Assemblies: Effects of Behavioral State, Experience, and EEG Dynamics. *The Journal of Neuroscience*, 19(10):4090–4101.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3-4):293–321.

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

McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2021). Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nature Neuroscience*, 24(6):851–862.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moore, A. W. and Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13(1):103–130.

Morris, R. G. M., Garrud, P., Rawlins, J. N. P., and O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381(6581):425–428.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1):171–175.

O'Neill, J., Senior, T. J., Allen, K., Huxter, J. R., and Csicsvari, J. (2008). Reactivation of experience-dependent cell assembly patterns in the hippocampus. *Nature Neuroscience*, 11(2):209–215.

Pfeiffer, B. E. and Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79.

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

Redish, A. D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805.

Rosenbaum, R., McKinnon, M. C., Levine, B., and Moscovitch, M. (2004). Visual imagery deficits, impaired strategic retrieval, or memory loss: disentangling the nature of an amnesic person's autobiographical memory deficit. *Neuropsychologia*, 42(12):1619–1635.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2016). Prioritized Experience Replay. *arXiv:1511.05952 [cs]*. arXiv: 1511.05952.

Singer, A. C. and Frank, L. M. (2009). Rewarded Outcomes Enhance Reactivation of Experience in the Hippocampus. *Neuron*, 64(6):910–921.

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

Stella, F., Baracskay, P., O'Neill, J., and Csicsvari, J. (2019). Hippocampal Reactivation of Random Trajectories Resembling Brownian Diffusion. *Neuron*, 102(2):450–461.e7.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition.

Walther, T., Diekmann, N., Vijayabaskaran, S., Donoso, J. R., Manahan-Vaughan, D., Wiskott, L., and Cheng, S. (2021). Context-dependent extinction learning emerging from raw sensory inputs: a reinforcement learning approach. *Scientific Reports*, 11(1):2713.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK.

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183(5):1249–1263.e23.

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*.

Wikenheiser, A. M. and Redish, A. D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. *Hippocampus*, 23(1):22–29.

Wilson, B. A., Baddeley, A. D., and Kapur, N. (1995). Dense amnesia in a professional musician following herpes simplex virus encephalitis. *Journal of Clinical and Experimental Neuropsychology*, 17(5):668–681.

Wu, X. and Foster, D. J. (2014). Hippocampal Replay Captures the Unique Topological Structure of a Novel Environment. *Journal of Neuroscience*, 34(19):6459–6469.

Zhang, R., Zhang, S., Tong, M. H., Cui, Y., Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2018). Modeling sensory-motor decisions in natural behavior. *PLOS Computational Biology*, 14(10):e1006518.

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.