Prioritized memory access explains planning and hippocampal replay

Marcelo G. Mattar
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540, USA
mmattar@princeton.edu

Nathaniel D. Daw
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540, USA
ndaw@princeton.edu

Abstract

To make decisions, animals must evaluate outcomes of candidate choices by accessing memories of relevant experiences. Recent theories suggest that phenomena of habits and compulsion can be reinterpreted as selectively omitting such computations. Yet little is known about the more granular question of which specific experiences are considered or ignored during deliberation, which ultimately governs decisions. Here, we propose a normative theory to predict not just whether but which memories should be accessed at each time to enable the most rewarding future decisions. Using nonlocal "replay" of spatial locations in hippocampus as a window into memory access, we simulate a spatial navigation task where an agent accesses memories of locations sequentially, in order of the expected utility of the computation: how much more reward would be earned due to better choices. We show that our theory offers a unifying account of a large range of hitherto disconnected findings in the place cell literature such as the balance of forward and reverse replay, biases in the replayed content, and effects of experience. We suggest that the various types of nonlocal events during behavior and rest reflect different instances of a single choice evaluation operation, unifying seemingly disparate proposed functions of replay including planning, learning and consolidation, and whose dysfunction may explain issues like rumination and craving.

1 Introduction

A hallmark of adaptive behavior is the effective use of experience to maximize reward (Sutton and Barto, 1998). In sequential decision tasks such as spatial navigation, actions can be separated from their consequences by many steps in space and time. Anticipating these consequences, so as to choose the best actions, thus often requires integrating multiple intermediate experiences from pieces potentially never experienced together (Daw and Dayan, 2014; Shohamy and Daw, 2015). For instance, planning may involve sequentially retrieving experiences to prospectively compose a series of possible future situations (Doll et al., 2015; Huys et al., 2015b; Kurth-Nelson et al., 2016). Recent research has explained many phenomena in which animals and humans either succeed in such prospective planning — or fail to, as in habits and compulsion — by suggesting they selectively engage or omit such computations as appropriate to the circumstances (via trading off two algorithms for value estimation, known as model-based and model-free) (Daw et al., 2005; Gillan et al., 2016; Keramati et al., 2011, 2016; Kool et al., 2016; Pezzulo et al., 2013). However, by focusing narrowly on whether or not to deliberate about the immediate future, this research largely fails to address which of the many possible experiences to consider in this evaluation process (though see: Cushman and Morris, 2015).

Relatedly, behavioral and neuroimaging data suggest that actions can also be evaluated by integrating experiences long before decisions are needed. In humans, for instance, future decisions in sequential tasks can be predicted from neural activity when relevant information is first learned (Wimmer and Shohamy, 2012) and also during subsequent rest (Gershman et al., 2014; Momennejad et al., 2017a) (Fig. 1a). Yet, this further highlights the selection problem: If actions can be evaluated long before they are needed, which experiences should the brain consider at each moment, and in what order, to set the stage for the most rewarding future decisions? These questions are likely central not just to healthy decisions, but to a variety of disordered including hallucations, craving, and rumination (Huys et al., 2015a; Smith et al., 2006). Addressing them requires a new, more granular theory predicting the patterns in which individual memories should be accessed to compute actions' values. Such patterns include prospective planning as a special case, but generalize to additional ways of computing values by accessing memories in different orders, at different times.

A valuable window into patterns of memory access is offered by the hippocampus, a structure known to play a critical role in memory for events and places (O'keefe and Nadel, 1978; Scoville and Milner, 1957). Neural activity recorded from hippocampal place cells during spatial navigation typically represents the animal's spatial position, though it can sometimes represent locations ahead of the animal (Diba and Buzsáki, 2007; Johnson and Redish, 2007; Pfeiffer and Foster, 2013). For instance, during "sharp wave ripple" events, activity might progress sequentially from the animal's current location towards a goal location (Diba and Buzsáki, 2007; Pfeiffer and Foster, 2013). These "forward replay" sequences predict subsequent behavior and have been suggested to support a planning mechanism that links actions to their deferred consequences along a spatial trajectory (Pfeiffer and Foster, 2013). However, analogously to the human evidence, remote activity in the hippocampus can also represent locations behind the animal (Ambrose et al., 2016; Davidson et al., 2009; Diba and Buzsáki, 2007; Foster and Wilson, 2006; Gupta et al., 2010), and even altogether disjoint, remote locations (especially during rest or sleep (Karlsson and Frank, 2009; Lee and Wilson, 2002)) (Fig. 1a). Indeed, the various conditions in which replay occurs have been suggested to reflect involvement in a range of distinct functions such as planning (Johnson and Redish, 2007; Pfeiffer and Foster, 2013), learning through credit assignment (Ambrose et al., 2016; Foster and Wilson, 2006; Johnson and Redish, 2005), memory retrieval (Jadhav et al., 2012) and consolidation (Carr et al., 2011; McClelland et al., 1995), and in forming and maintaining a cognitive map (Gupta et al., 2010; Tolman, 1948).

Here, we develop a normative theory to predict not just whether but which memories should be accessed at each time to enable the most rewarding future decisions. Our framework, based on the DYNA reinforcement learning (RL) architecture (Johnson and Redish, 2005; Ludvig et al., 2017; Sutton, 1990), views planning as learning about values from remembered experiences, generalizing and reconceptualizing work on trade-offs between model-based and model-free controllers (Daw et al., 2005; Keramati et al., 2011). We derive from first principles the utility of retrieving each individual experience at each moment to predict what memories a rational agent ought to access to lay the groundwork for the most rewarding future decisions. This utility, formalized as the increase in future reward resulting from such memory access, is shown to be the product of two terms: a gain term which prioritizes states behind the agent when an unexpected outcome is encountered; and a need term which prioritizes states ahead of the agent that are immediately relevant. Importantly, this theory at present investigates which experience among all would be most favorable in principle; it is not intended as (but helps point the way toward) a mechanistic or process-level account of how the agent might efficiently find them.

To test the implications of our theory, we simulate a spatial navigation task where an agent generates and stores experiences which can be later retrieved. We show that an agent that accesses memories sequentially and in order of utility produces patterns of sequential state consideration that resemble place cell replay, and reproduces qualitatively and with no parameter fitting a wealth of empirical findings including (i) the existence and balance between forward and reverse replay; (ii) the content of replay; and (iii) effects of experience. Thus, we propose the unifying view that all patterns of replay during behavior, rest, and sleep reflect different instances of a more general state retrieval operation that integrates experiences across space and time to propagate value and guide decisions.

2 Results

We consider a class of sequential decision tasks where an agent must decide in each situation (state; e.g. location in a spatial task) which action to perform with the goal of maximizing its expected future reward. The optimal course of action (policy) consists of selecting the actions with highest expected value (Q-value). The value of an action is defined as the expected utility, or cumulative discounted reward, from taking that action and following the optimal policy thereafter. Optimal decision making, therefore, requires the agent to estimate action values as accurately as possible for maximizing total reward.

We consider the fundamental evaluation operation in RL, a *Bellman backup*, which locally updates the value of an action in some state as the immediate payoff received when executing that action plus the discounted expected value of the resulting successor state. When applied to behavior, following the actual choice of an action and the observation of its reward and successor state, this operation corresponds to the standard temporal difference (TD) learning update, the basis of prominent theories of dopaminergic experiential learning (Schultz et al., 1997). Our account includes this "model-free" learning, but also allows for additional Bellman backups to be performed nonlocally, by retrieving the experience of executing an action and transitioning between states (i.e., a quantum of remembered or simulated experience), as in the DYNA framework (Sutton, 1990). Note that we refer to the information processed in a backup — a state, action, reward and successor state — as a "memory" but it could arise from two sorts of mnemonic representations: either a record of an individual experienced event, or a sample drawn from a learned model of the task. Such a learned transition model (formally, the estimated conditional distribution of successor states given a predecessor state and action, akin to a cognitive map) is also a (more semantic) memory of task experience, though expressed as a statistical summary rather than an individual event record. Since here we consider only fixed, deterministic spatial tasks, there is no variability between events, and these approaches coincide.

Stringing together these backup operations over a sequence of states and actions computes expected value over a trajectory; different "model-based" algorithms for computing values (such as tree search and value iteration) amount to a batch of many such backup operations, performed in different orders (Daw and Dayan, 2014; Sutton and Barto, 1998). Because this process can compose behavioral sequences or trajectories from pieces not experienced together, it can discover consequences missed by TD learning, which assess actions only in terms of their directly experienced outcomes (Schultz et al., 1997). Importantly, retrieving an experience allows the agent to update the value of any action in terms of its outcome, including in particular actions available in states other than the current state of the agent.

In contrast to previous theories, which largely considered whether or not to fully compute candidate actions' expected rewards by an extended, iterative deliberation (Daw et al., 2005; Keramati et al., 2011), we offer a more granular analysis by considering the (approximately) optimal scheduling of individual steps of value computation. To analyze which backups the agent should execute in any situation, we derive the utility of any specific Bellman backup – updating an action value at a target, potentially nonlocal, state – as the product of a *gain* and a *need* term (see Methods). The gain term quantifies the net increase in discounted future reward expected from a policy change at the target state — that is, it measures how much more reward the agent can expect to harvest following any visit the target state, due to what it learns from the update. This value depends, in turn, on whether the update changes the agent's policy, meaning that (in contrast to other prioritization heuristics considered in AI; (Moore and Atkeson, 1993; Peng and Williams, 1993; Schaul et al., 2015)), the theory predicts asymmetric effects of positive and negative prediction errors due to their differential effect on behavior (Fig. 1d).

To determine priority, the gain term is multiplied by the need term, which quantifies the number of times the agent is expected to harvest the gain by visiting the target state in the future. Here, earlier visits are weighted more heavily than later visits due to temporal discounting. This weighting implies that the need term prioritizes the agent's current state, and others likely to be visited soon (Fig. 1e).

Mathematically, the product of gain and need is the utility of a backup which updates an action value in any part of the environment (computed under a myopic approximation, i.e. treating each backup individually and one at a time). Overall priority thus depends on the balance between these two imperatives. For instance, a backup that has no effect on behavior has zero utility even if the target state is expected to be visited in the future (because it has zero gain, despite high need). Similarly, the need for a backup is zero if a state is never expected to be visited again, even if this backup would greatly impact behavior at the that state (because it has zero need, despite high gain).

In order to test the implications and properties of this theory, we simulate an optimal agent's behavior in a spatial navigation task (a "grid-world") where states are are locations in the environment and an action is a step of movement in one of the four cardinal directions. We assume that when the agent is paused (here, before starting a run and upon receiving a reward), it may access nonlocal memories, and that it does so sequentially in order of utility. By repeating this reactivation sequentially, value information may be propagated along spatial trajectories that may have never been traversed continuously by the agent. We also assume that this local operation is accomplished by place cell activity at the target location, allowing predictions of patterns of replay. Theories of model-free reward learning in navigational tasks typically assume that a hippocampal location representation is the input to a learned value mapping in striatum, updated

by dopaminergic prediction error (Foster et al., 2000). Trajectory replay in the hippocampus, which drives activation and plasticity throughout this system, is thus a substrate for value learning (Gomperts et al., 2015; Lansink et al., 2009; Meer and Redish, 2011).

We consider two distinct spatial environments (Fig. 1b). First, we simulate a linear track where the agent shuttles back and forth to collect rewards at the ends, a task widely used in studies of hippocampal replay (Fig. 1b, *left*). Second, we simulate an field with obstacles (walls) where the agent needs to move toward a reward placed at a fixed location, a task used extensively in previous RL studies of prioritized simulated experience (Peng and Williams, 1993; Sutton and Barto, 1998) (Fig. 1b, *right*).

2.1 Memory access and learning

Our first prediction was that prioritized memory access accelerates learning in a spatial navigation task. In both environments, we contrast an agent that accesses memories in a prioritized order with a baseline model-free agent that learns only by direct experience, and with an agent that simulates experiences drawn at random (original DYNA (Sutton, 1990)). In all cases, the number of steps required to complete an episode (the time between two receiving two rewards) is gradually reduced as the agent learns the task. Learning with prioritized experience replay, however, progresses faster due to the agent's ability to rapidly propagate value information along relevant trajectories (Fig. 1c). Notice that our theory predicts that a model-free agent is nonetheless able to learn this type of task, albeit in a slower fashion, in line with findings where the disruption of replay slows down learning without abolishing it completely (Jadhav et al., 2012).

2.2 Context-dependent balance between forward and reverse sequences

A major prediction of our theory is that the patterns of memory access are not random, but often involve adjacent locations to propagate value across spatial locations iteratively. In our simulations, we observed that, as in hippocampal recordings, replayed target states typically followed continuous sequences in either forward or reverse order (Fig. 2. In particular, our theory predicts two predominant patterns of backup, driven by the two terms of the prioritization equation. First, when an agent encounters a prediction error, this produces a large gain term behind the agent (Fig. 2a-f), reflecting the gain from propagating the new information to potential predecessor states (and, recursively, toward their predecessors). In this case, sequences tend to start at the agent's location and move backwards towards the start state (Fig. 2c,f). The need term, instead, tends to be largest in front of the agent (Fig. 2g-l); when it dominates, sequences tend to start at the agent's location and move forward toward the reward location (Fig. 2i,l). Importantly, such forward and backward activity is correlated with, but not simply identical to, the agent's recent and future past (as we quantify below); in fact, backups can even construct trajectories not previously traversed by the agent (Gupta et al., 2010). These findings largely reproduce the different patterns of reactivation observed in both humans and rodents, suggesting that they can be explained via the same prioritized operation applied in different situations.

Our theory also predicts that the patterns of memory access are not random, but often involve adjacent locations to propagate value across spatial locations iteratively. In our simulations, we observed that, as in hippocampal recordings, replayed target states typically followed continuous sequences in either forward or reverse order (Fig. 2. In particular, our theory predicts two predominant patterns of backup, driven by the two terms of the prioritization equation. First, when an agent encounters a prediction error, this produces a large gain term behind the agent (Fig. 2a-f), reflecting the gain from propagating the new information to potential predecessor states (and, recursively, toward their predecessors). In this case, sequences tend to start at the agent's location and move backwards towards the start state (Fig. 2c,f). The need term, instead, tends to be largest in front of the agent (Fig. 2g-l); when it dominates, sequences tend to start at the agent's location and move forward toward the reward location (Fig. 2i,l). These findings largely reproduce the different patterns of reactivation observed in both humans and rodents, suggesting that they can be explained via the same prioritized operation applied in different situations.

To quantify these observations in simulation, we classified each individual backup as *forward* or *reverse* by examining the next backup in the sequence. When a backed-up action was followed by a backup in that action's resulting state, it was classified as a *forward*. In contrast, when the state of a backup corresponded to the outcome of the following backed-up action, it was classified as *reverse*. Backups that did not follow either pattern were not classified in either category. We then followed standard procedures for identifying hippocampal replay events and assessed the significance of all consecutive segments of forward or reverse backups of length five or greater with a permutation test (Davidson et al., 2009; Diba and Buzsáki, 2007; Foster and Wilson, 2006).

In line with findings from the human and rodent literature on the linear track, we observed that replay (driven by the need term) extended in the forward direction before a run (Fig. 3a, *left*), providing information relevant for evaluating future trajectories. In contrast, replay extended in the reverse direction upon completing a run (driven by the gain term, Fig. 3a, *right*), providing a link between behavioral trajectories and their outcomes. Very few reverse sequences were observed prior to the onset of a run, and very few forward sequences were observed upon completion of a run, in line with previous findings (Diba and Buzsáki, 2007) (Fig. 3b).

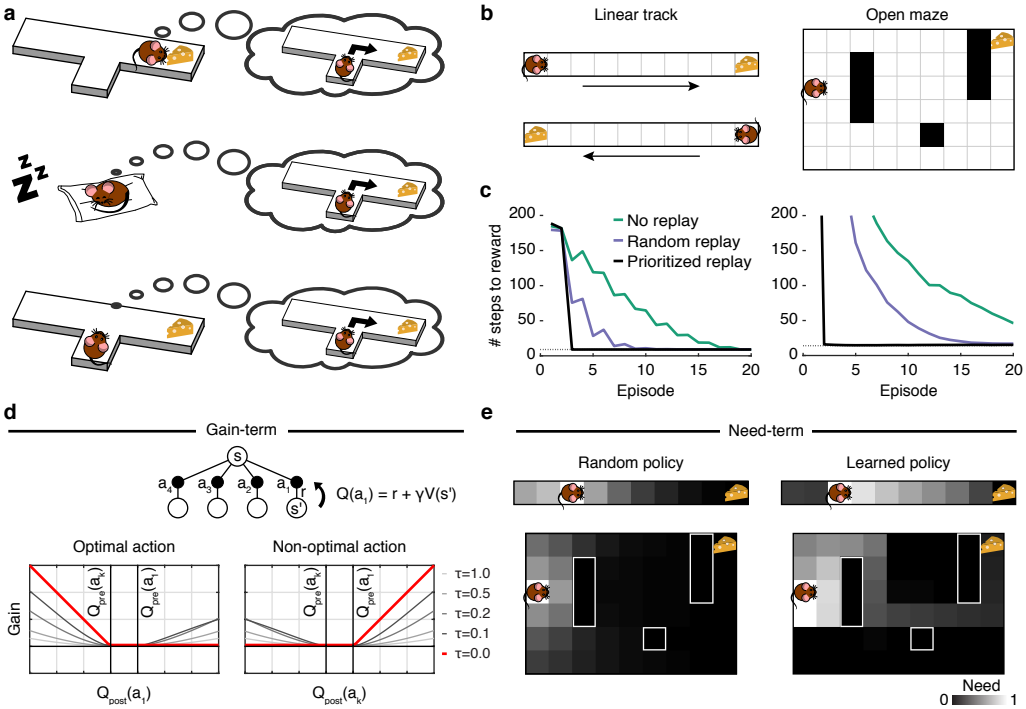

Figure 1: **Rational memory access through prioritized replay improves behavior** *(a)* Three ways an agent might learn, through replay, the relationship between an action and reward: *Top*: when reward is first encountered, through reverse reactivation; *Middle*: during sleep/rest through "offline" reactivation of the sequence; *Bottom*: prior to choice, by prospective (forward) activation. The last case is most often envisioned in theories of model-based deliberation, but replay of all three sorts occurs and human neuroimaging evidence suggests all can support decisions. *(b)* Grid-world environments simulated. *Left*: a linear track simulated as two disjoint segments (to reflect the unidirectionality of the hippocampal place code in this case) with rewards in opposite ends. *Right*: A two-dimensional maze with obstacles. *(c)* Replaying experiences according to the prioritization scheme speeds learning compared to learning without replay or with replay of randomly ordered experiences. *Left*: linear track; *Right*: field. *(d)* The gain term for updating a target action in a target state quantifies the increase, due to the update, in reward expected following a visit to the target state. *Top*: A schematic of a Bellman backup where the reactivation of experience $e = (s, a, r, s')$ propagates the one-step reward r and the value of s' to the state-action pair (s, a). *Bottom*: For a greedy agent (red line), the gain is nonzero either when the best action is found to be worse than the second best action (*left*, changing the policy to disfavor it) or when a suboptimal action is found to be the best action (*right*, changing the policy to favor it). The amount of gain increases with how much worse or better (respectively) is the old or new best action. Gray lines indicate that the gain term is more symmetric for more exploratory behavior (τ: softmax temperature parameter). *(e)* The need term from a particular target state corresponds to its expected future occupancy, measuring how imminently and how often reward gains will be harvested there. This is shown as a heat map over states, and also depends on the agent's future action choice policy, e.g. *Left*: Random policy (initially). *Right*: Learned policy, following training.

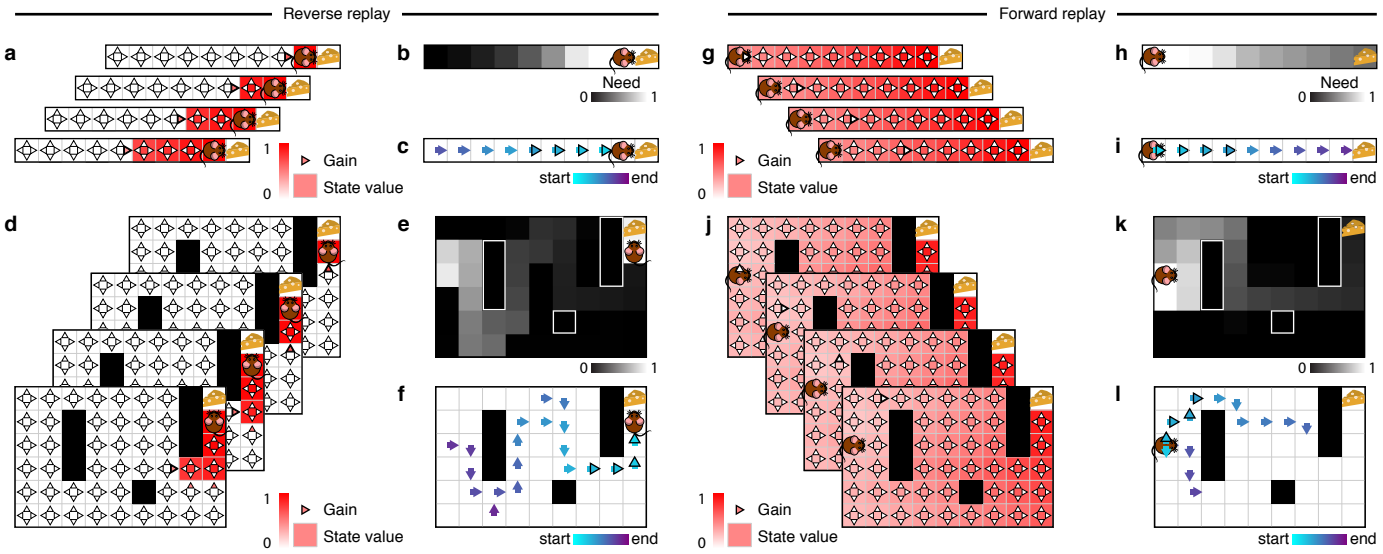

Figure 2: **Replay produce extended trajectories in forward and reverse directions.** *(a-f)* Example of reverse replay. *(g-l)* Example of forward replay. *(a,d)* Gain term and state values. Notice that the gain term is specific for each action, and that it may change after each backup due to its dependence on state values. Replaying the last action executed before finding an unexpected reward has a positive gain, since the corresponding backup will cause the agent to more likely repeat that action in the future. Once this backup is executed, the value of the preceding state is updated and replaying actions leading to this updated state will have a positive gain. Repeated iterations of this procedure leads to a pattern of replay that extends in the reverse direction. *(g,j)* If gain differences are smaller than need differences, the need term dominates and sequences will tend to extend in the forward direction. *(b,e,h,k)* Need term. Notice that the need term is specific for each state and doesn't change after each backup due to being fully determined by the current state of the agent. The need term prioritizes backups near the agent and extends forwards through states the agent is expected to visit in the future. In the field, the need term is responsible for sequences expanding in a depth-first manner as opposed to breadth-first. *(c,f)* Example reverse sequences obtained in the linear track (c) and open field (f). *(i,l)* Example forward sequences obtained in the linear track (i) and open field (l). Notice that forward sequences tend to follow agent's previous behavior but may also find new paths towards the goal.

Our claim that both forward and reverse replay may arise from the same prioritized operation may seem at odds with the general view that forward and reverse sequences have distinct functions (planning and learning, respectively (Ambrose et al., 2016; Diba and Buzsáki, 2007)). Evidence for this distinction has been argued to come from the observation that reverse and forward replay have different sensitivities to reward context: in rodents navigating a linear track, the rate of reverse replay is increased when the animal encounters an increased reward, and decreased when the animal encounters a decreased reward. In contrast, the rate of forward replay is similar despite increases or decreases in reward (Ambrose et al., 2016; Singer and Frank, 2009). Our hypothesis is instead that planning and learning are better understood as different variants of the same operation, i.e. using backups (in different orders) to propagate reward information over space and time. Yet, a hallmark of the model's gain term (arising from its definition in terms of policy change; Fig. 1d, and distinguishing our prioritization hypothesis from others that simply trigger update on any surprise (Moore and Atkeson, 1993; Peng and Williams, 1993; Schaul et al., 2015)) is asymmetric effects of increases vs. decreases in reward. This effect is predicted to arise only for reverse replay occurring at the end of a run, when the gain term is large and, therefore, dominates the utility of the backup.

We investigated the differential response of these two types of replay to increases or decreases in reward by simulating two conditions (i) an (*increased reward*) condition where the reward encountered by the agent was four times larger in half of the episodes, and (ii) a (*decreased reward*) condition where the encountered by the agent was zero in half of the episodes. The number of forward events was approximately equal in the increased reward setting regardless of the reward received. In contrast, the number of reverse events was much larger upon receiving a larger reward than upon receiving a conventional reward (Fig. 3c,d). This effect was driven both by an increase in the rate of reverse replay for larger rewards, and a decrease for conventional (1x) rewards, as observed experimentally (Ambrose et al., 2016). In the decreased reward setting, the number of forward events was also approximately equal in every lap regardless of the reward received. In contrast, the number of reverse events was much smaller upon receiving no reward than upon receiving a conventional reward (Fig. 3e,f). This effect was driven both by an decrease in the rate of reverse replay when the reward was 0, and an increase when the reward was conventional (1x), again replicating empirical findings (Ambrose et al., 2016).

2.3 Statistics of replayed locations: current position, goals, and paths

We have so far shown that prioritized memory access applied in different conditions may give rise to forward and reverse sequences, in line with both planning and learning (viewed as "pre-planning" to take account of the remote consequences of newly learned information) respectively. In addition to directionality, our theory predicts that the replay should also be biased toward relevant locations in the environment such as the agent's current position and the reward sites. In the model, these general biases arise from the average over many situations of its more specific experience- and situation-dependent replay, which is patterned due to the influence of particular locations like reward sites on both the need and gain terms.

Empirically, hippocampal events have an overall tendency to reflect a path that begins at the animal's current location (Davidson et al., 2009; Diba and Buzsáki, 2007; Foster and Wilson, 2006), a phenomenon termed 'initiation bias'. In contrast, no such bias is observed for the site where replay ends (Davidson et al., 2009) (Fig. 4b). Due to the need term, our simulations of the linear track reproduced these results, with replay events starting in locations exactly at or immediately behind the agent (Fig. 4a, *left*), and no such bias observed for locations where replay ended (Fig. 4a, *right*).

Sequential replay has also been shown to be biased toward goal and/or reward sites (Dupret et al., 2010; Pfeiffer and Foster, 2013) ("content biases") in line with its proposed role in route planning. To test these predictions, we simulated navigation in an open field environment, allowing a richer behavior where the planned trajectories are less constrained by the environment (Fig. 1d, *right*). We examined the existence of content biases by calculating the *activation probability* for each spatial location in the open field environment (the probability of a backup happening at each state within an episode). Backups tended to concentrate at or around the reward (goal) locations, in line with rodent recordings (Ólafsdóttir et al., 2015; Pfeiffer and Foster, 2013) (Fig. 4c). We quantified these results by calculating the probability of a backup happening at various distances from the agent and reward within an episode. In line with empirical evidence (Pfeiffer and Foster, 2013), backups were more likely than chance to happen near the agent and/or near the reward (though not exclusively (Jackson et al., 2006; Pfeiffer and Foster, 2013)), suggesting that replay encode trajectories that are useful for driving behavior towards a goal (Fig. 4d).

We then examined these replay probabilities separately within significant forward (Fig. 4e) and reverse events (Fig. 4g). In line with our previous results on the linear track (Fig. 3a), we observed that reverse events (which tend to happen when the agent reaches a reward) concentrate immediately behind the agent, due a combination of initiation bias and reward-location bias (Fig. 4g). Interestingly, forward events also have a slight tendency to represent locations near the reward, despite the fact that start locations were randomized in these simulations (Fig. 4e). In particular, locations corresponding to the final turn towards the reward were emphasized even more than locations near the reward, a consequence of the gain term being highest when there is a large effect on behavior. These results are consistent with empirical reports that reactivated place fields congregate around relevant cues (Singer and Frank, 2009).

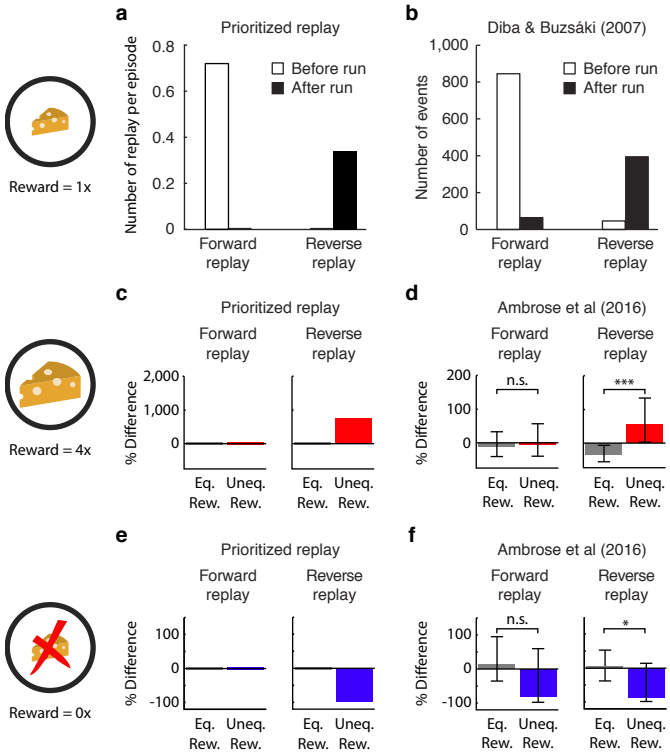

Figure 3: **Forward and reverse sequences happen at different times and are modulated asymmetrically by reward.** *(a)* Forward sequences tend to take place before the onset of a run while reverse sequences tend to take place after the completion of a run, upon receipt of reward. *(b)* Data from Diba & Buzsáki (2007), their Fig. 1C. *(c)* We simulated a task where, in half of the episodes, the reward received was 4x larger than baseline. *Left*: The number of forward events was approximately equal in every lap both when the rewards were equal, as well as when the rewards were 4x larger. *Right*: In contrast, the number of reverse events was approximately equal when the rewards were equal, but much larger upon receiving a larger reward in the unequal reward condition. *(e)* Data from Ambrose et al (2016), their Fig. 3E,H. *(d)* We simulated a task where, in half of the episodes, the reward received was zero. *Left*: The number of forward events was approximately equal in every lap both when the rewards were equal, as well as when the rewards were removed. *Right*: In contrast, the number of reverse events was approximately equal when the rewards were equal, but almost completely abolished upon receiving no reward in the unequal reward condition. *(f)* Data from Ambrose et al (2016), their Fig. 5C,F; note that the effect of reward for forward replay (left) is not significant.

Given the proposed involvement of replay in planning and learning, forward and reverse replay might differentially predict future and past behavior, respectively. We measured the probability that a given forward or reverse event would include locations visited by the agent in the future or past. In the open field, our simulations revealed that forward replay has a higher probability than chance of including states visited a few steps in the future, and a probability only slightly higher than chance to include states visited in the past (Fig. 4f). In contrast, reverse replay has a higher probability than chance of including states visited in the past, and a probability only slightly higher than chance to include states visited in the future (Fig. 4h). Therefore, replayed trajectories not only represent the agent and reward locations, but tend to correspond to the specific trajectories followed by the agent in either the past (reverse replay) or future (forward replay), again in line with rodent studies (Pfeiffer and Foster, 2013; Singer and Frank, 2009).

Lastly, we address the case of remote replay, where sequences correspond to spatial locations away from the animal (Davidson et al., 2009) or remote environments altogether (Karlsson and Frank, 2009). In particular, even during sleep — where the content of replay rarely corresponds to the location where the animal is sleeping — replay tends to represent rewarding areas of the environment in comparison to similar but unrewarding areas (Ólafsdóttir et al., 2015). In our model, biases in reactivation during rest can again be understood in terms of the same need- and gain-based prioritization (with need defined as expected future occupancy subsequently). We tested these predictions of sleep replay by simulating a T-maze with a reward placed at the end of one of the two arms (Fig. 4i), with the agent absent from the environment (see Methods). The proportion of backups corresponding to actions leading to the cued arm was much greater than the proportion of backups corresponding to actions leading to the uncued arm (Fig. 4j), in line with empirical results (Ólafsdóttir et al., 2015).

2.4 Diverging effects of familiarity and specific experiences

As a learning model, our theory also predicts characteristic effects of experience on the prevalence and location of replay. In particular, change in the need vs. gain terms predicts countervailing effects of task experience. As a task becomes well learned, prediction errors decrease, policies stabilize, and the gain expected due to replay decreases, favoring a reduction in overall replay (in the model, gain never vanishes entirely due to an assumption of persistent uncertainty because of the possibility of environmental change). At the same time, as behavior crystallizes, the need term becomes more focused along the particular routes learned by the agent (e.g., compare Fig. 1e, left and right). This predicts that, conditional on replay occurring, particular states are increasingly likely to participate.

The balance between these countervailing effects may help to explain apparent inconsistency in the replay literature, as both increases and decreases in replay have been reported, albeit using a range of different dependent measures and designs. Indeed, existing reports on the effect of familiarity on the incidence of replay have been difficult to reconcile (Buhry et al., 2011; Cheng and Frank, 2008; Jackson et al., 2006; O'Neill et al., 2008; Singer and Frank, 2009). While some studies report an increase in reactivation with time spent in the environment (Jackson et al., 2006; O'Neill et al., 2008), several others report a decrease with familiarity (Cheng and Frank, 2008; Diba and Buzsáki, 2007; Foster and Wilson, 2006; Singer and Frank, 2009). Specifically, the more time an animal spends between two place fields, the more the corresponding place cell pair is reactivated during sleep (consistent with focusing of need on these states (O'Neill et al., 2008)). In contrast, replay is more easily observed in novel than in familiar tracks (consistent with a decrease in gain overall (Foster and Wilson, 2006)), and the average activation probability is highest in novel environment (Cheng and Frank, 2008). It has been suggested that replay tends to increase within session with exposure, but decrease between sessions, but decrease across sessions as the animal becomes familiar with a novel environment (Buhry et al., 2011). This may reflect the additional effect of experience vs. computation on learning in our model. In particular, both need (favoring focused replay) and gain (opposing overall replay) are affected by actual experience in an environment, but only gain is affected by replay (e.g. during rest between sessions). This is because only experience can teach an agent about the situations it is likely to encounter, i.e. need), but value learning from replayed experience reduces subsequent gain.

We examined the effect of familiarity and specific experience on the incidence of replay. We simulated our agent in both environments and calculated the number of significant replay events as a function of experience (episode number). In line with previous reports (Foster and Wilson, 2006), we observed that the number of significant events is highest during the first episodes and decays gradually with experience (Fig. 5a). Fewer significant replay events in both forward and reverse directions were observed with increased experience (Fig. 5a, insets). Yet, when calculating the probability of an event including a specific state, we observed that it increased with number of visits in both environments (Fig. 5b), also in line with previous reports (O'Neill et al., 2008). Again, these two effects reflect the effect of experience on the two terms governing priority: while the gain term decreases with exposure (with the gradual reduction in prediction errors), the need term increases as the agent's trajectory becomes more predictable.

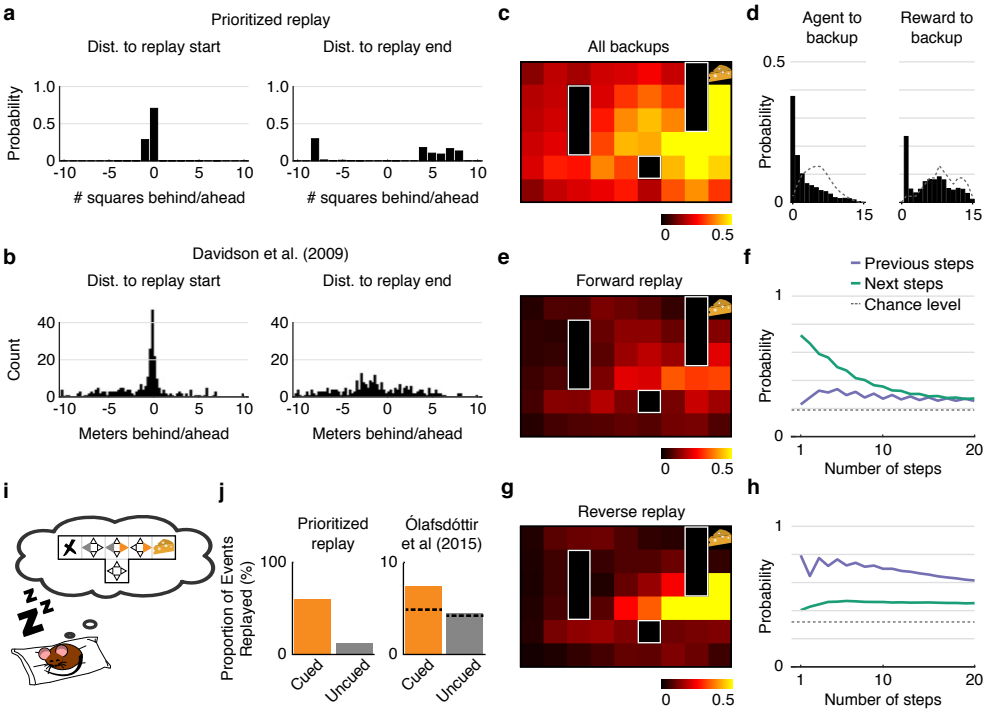

Figure 4: **Replay over-represents agent and reward locations and predicts subsequent and past behavior.** *(a) Left*: Distribution of start locations of replay trajectories relative to the agent's position and heading on the linear track. Negative distances indicate that the replayed trajectory starts behind the agent. All replay events in the linear track start at or immediately behind the agent's location. *Right*: Similar distribution for end locations of replay trajectories relative to the agent's position and heading on the linear track. Most replay events end slightly ahead or slightly behind the agent's location. *(b)* Data from Davidson et al (2009) showing the distribution of start (*Left*) and end (*Right*) locations of replay trajectories relative to the animal's position and heading on the track. *(c)* Activation probability across all backups within an episode. Colors represent the probability of a backup happening at each location within a given episode. Notice that backups are more likely to occur in locations near the reward. *(d)* Probability that a given backup happens at various distances from the agent (*left*) and from the reward (*right*) in the open field. Dotted lines represent chance levels. Notice that backups are substantially more likely to happen near the agent and/or near the reward than chance. *(e)* Activation probability across all backups within significant forward sequences. Forward replay tends to concentrate around turning points near the reward. Notice that because the random starting location, no initiation bias is observed in this plot. *(f)* How forward replay predicts future and previous steps in the open field. The lines indicate the probability that any given forward sequence within an episode contains the state the agent will/have occupied a given number of steps in the future/past. Dotted lines represent chance levels. Notice that forward replay is more likely to represent future states than past states. *(g)* Activation probability across all backups within significant reverse sequences. Reverse replay tends to concentrate near the reward. Notice that the higher activation probability for reverse events is due to a combination of a reward-location bias and initiation bias, given that reverse sequences tend to start near the reward, where the agent is. *(h)* How reverse replay predicts future and previous steps in the open field. The lines indicate the probability that any given reverse sequence within an episode contains the state the agent will/have occupied a given number of steps in the future/past. Dotted lines represent chance levels. Notice that reverse replay is more likely to represent past states than future states. *(i)* We simulated an agent in an offline setting (e.g. sleep) after exploring a T-maze and receiving a reward on the right (cued) arm. *(j) Left*: The proportion of backups corresponding to actions leading to the cued arm (red) is much greater than the proportion of backups corresponding to actions leading to the uncued arm (blue). *Right*: Data from Ólafsdóttir et al (2015) showing the proportion of spiking events categorized as preplay events for the cued and uncued arms. The grey dashed line shows the proportion of events expected by chance.

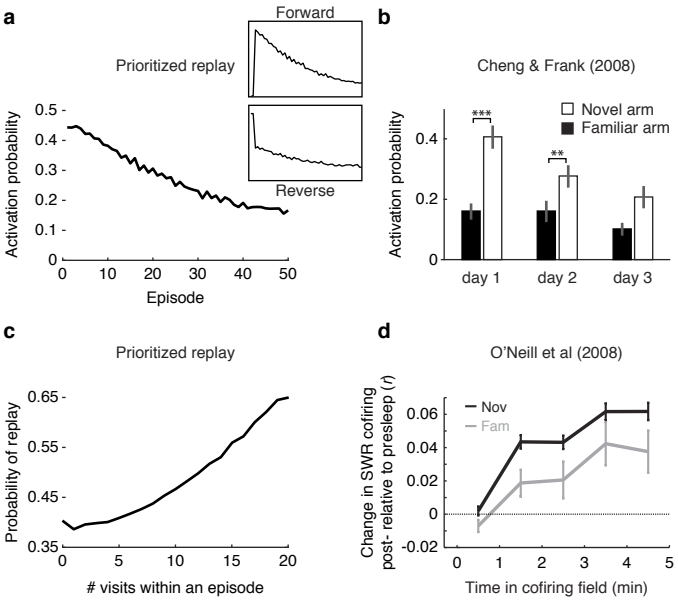

Figure 5: **Replay frequency decays with familiarity and increases with experience.** *(a)* In the linear track, the number of significant replay events decays across episodes, peaking when the environment is novel. Insets show that the number of both forward (top) and reverse (bottom) replay events decay with experience. *(b)* Data from Cheng Frank (2008) showing the activation probability per high-frequency event. Error bars represent standard errors and symbols indicate results of rank-sum test ($\star\star$ $p < 0.01$, $\star\star\star$ $p < 0.001$). *(c)* Probability that significant replay events include a state in the linear track as a function of the number of visits in an episode. Analogously to the effect reported in Fig. 1e, driven by the need term, the probability of a state being replayed increases with experience in that state. *(d)* Data from O'Neil et al (2008) showing that that the more cell pairs that fire together during exploration (time in cofiring field), the larger is the increase in probability that these cell pairs fire together during sleep SWRs.

2.5 Effect of replay on choice behavior

The preceding simulations demonstrate that a wide range of properties of place cell replay can be predicted from first principles, under the hypothesis that the common goal of replay is to drive reinforcement learning and planning involving the reactivated locations. This hypothesis also makes a complementary set of predictions about choice behavior, i.e. that replay will be causally involved in learning which actions to take at the replayed states. Such behavioral effects are most characteristically expected for acquiring tasks (like revaluation and shortcut tasks) that require agents to integrate associations learned separately, or to infer the value of novel actions. This is because this class of tasks cannot be solved by alternative learning mechanisms in the brain (such as model-free TD learning, associated with dopamine and striatum) and exercise the more unique ability of nonlocal replay to compose novel trajectories from separate experiences (Shohamy and Daw, 2015).

Indeed, hippocampal replay can follow novel paths or shortcuts never traversed by an animal (Gupta et al., 2010), and in one report (Ólafsdóttir et al., 2015), activation of a path not yet explored (because it was initially observed behind glass) was followed by rats subsequently being able to choose that path, correctly, over another, consistent with the planning hypothesis. In the open field, forward hippocampal replay predicts future paths, importantly even when the goal location is novel (Pfeiffer and Foster, 2013). Finally, causally blocking sharp wave ripples has a selective effect on learning and performance of a spatial working memory task; although this task does not specifically exercise integrative planning, it does require associating events over space and time (Jadhav et al., 2012). Overall, though, the place cell literature has tended not to focus on connecting hippocampal activity to learning and decision behavior, and a major area for future research suggested by our theory is to combine tasks more specifically diagnostic of model-based planning and reinforcement learning with the monitoring and manipulation of nonlocal replay.

A second feature of the current theory is that it emphasizes that a number of different patterns of replay (forward, reverse and offline; Fig. 1a) can all equally be used to solve the sorts of integrative decision tasks that have largely been assumed to reflect forward "model-based" planning at the time of choice. Indeed, forward planning of this sort may be subserved by preplay (Johnson and Redish, 2007), but in the current theory, this is just one case of a more general mechanism, and equivalent computations can also arise from memories activated in different patterns at other times. These also include reverse replay that allows connecting an experienced outcome with potential predecessor actions, and nonlocal replay composing sequences of experiences during rest. Although these possibilities have not been examined in hippocampal

spatial literature, work with humans using non-spatial versions of revaluation tasks (and activity of category-specific regions of visual cortex to index state reinstatement) verifies that not just forward replay (Doll et al., 2015) but also reverse replay (Wimmer and Shohamy, 2012) and nonlocal replay during rest (Momennejad et al., 2017a) all predict the ability of subjects to solve these tasks. The present theory's account of which replay events are prioritized might provide a basis for investigating why different studies and task variants tend to evoke one or the other of these solution strategies.

3 Discussion

In light of so much experience accumulated in a lifetime, which memories should one access and when to allow for the most rewarding future decisions? We offer a rational account for the prioritization of memory access operations framed in terms of action evaluation through Bellman backups. We propose that the various nonlocal place-cell phenomena in the hippocampus reflect different instances of a single evaluation operation, and that differences in the utility of these operations can account for the heterogeneity of circumstances in which they happen. This utility, derived from first principles, amounts to the product of a gain and a need term. Simulations of the model reproduced qualitatively a wide range of results reported in the hippocampal replay literature over the course of the previous decade without the need for any parameter fitting.

This theory draws new, specific connections between research on the hippocampal substrates of spatial navigation, and research on learning and decision making, with implications for both areas. It has long been recognized that place cell activity (including forward and reverse replay) likely supports learning and decision making (Foster and Wilson, 2006; Johnson and Redish, 2007); the present research renders these ideas experimentally testable by offering a specific hypothesis what the brain learns from any particular replay event. This immediately suggests experiments combining trial-by-trial reward learning (of the sort often studied in the decision literature) with place cell monitoring, to test the the predicted relationship between individual replay events and subsequent choices. The strongest test would use tasks (like sensory preconditioning or multi-step sequential decision tasks) where relevant quantities can't be directly learned "model-free" over experienced trajectories but instead exercise the ability of replay to compose novel sequences (Miller et al., 2017; Momennejad et al., 2017a; Shohamy and Daw, 2015). Upstream of this behavioral function, the theory also makes new testable predictions about place cells themselves, by articulating quantitative experience- and circumstance-dependent criteria for which locations will be replayed.

The hippocampal literature has tended to envision that replay serves disjoint functions in different circumstances, including learning (Foster and Wilson, 2006), planning (Diba and Buzsáki, 2007; Johnson and Redish, 2007; Pfeiffer and Foster, 2013; Singer et al., 2013), spatial memory retrieval (Jadhav et al., 2012), and systems consolidation (Carr et al., 2011; McClelland et al., 1995). By focusing on a specific, quantitative operation (long-run value computation), we sharpen these suggestions and expose their deeper relationship to one another. In RL, learning amounts to propagating long-run value information from one state to adjacent ones to perform temporal credit assignment, with forward "planning" as traditionally conceived being one special case. This perspective unifies the proposed role of forward replay in planning with that of reverse replay in learning (linking recently experienced sequences to their outcome (Foster and Wilson, 2006)), and suggests analogous nonlocal computations, e.g. during sleep. Though serving a common goal, these different patterns of replay are most appropriate in different circumstances; this explains observations of differential regulation (such as asymmetric effects of prediction errors on forward vs. reverse replay), which have otherwise been taken as evidence for distinct functions (Ambrose et al., 2016). As for consolidation, the perspective that replay drives estimation of long-run values echoes other work on systems consolidation (Kumaran et al., 2016; McClelland et al., 1995) in viewing consolidation not merely as strengthening existing memories, but more actively computing new summaries from the replayed content. Also as with other systems consolidation theories, the resulting computed quantities (here, action values) are widely believed to be stored elsewhere in the brain (likely cortico-striatal synapses), and the fuller neural processes of replay presumably involve coordinated evoked activity throughout the brain, especially including value prediction and learning in the mesolimbic and nigrostriatal rewards network (Gomperts et al., 2015; Lansink et al., 2009).

Relatedly, while we have hypothesized a specific role for replay in computing long-run action values — and although it is striking that this consideration alone suffices to explain so many regularities of place cell replay — we do not view this function as exclusive of other computations over replayed experiences (Kumaran et al., 2016; McClelland et al., 1995). One interesting variant of our theory is that replay can be used to learn a long-run transition model of the particular locations and outcomes one expects to visit following some action — instead of, as in our theory, the long-run reward consequences alone. Such a long-run outcome representation, known as the successor representation (SR), can serve as an intermediate representation for computing action values (Dayan, 1993), a sort of temporally extended cognitive map. The SR has recently been proposed to be learned within hippocampal recurrents (Stachenfeld et al., 2017) and to explain aspects of human choice behavior (Momennejad et al., 2017b). It can also be updated using replayed experience ("SR-DYNA" (Russek et al., 2017)) analogous to how we learn reward values here, connecting learning from replay more directly with a type of cognitive map building (Gupta et al., 2010). Our ideas carry directly over to this case: In fact, our prioritization computations remain unchanged if replay updates an SR instead of action values; this is because an SR

update step (also based on the Bellman equation) has exactly the same utility (under our myopic approximation) as the corresponding Bellman backup for action values.

From the perspective of decision neuroscience, a key driver of recent progress is the recognition that the details how decision variables are computed — specifically, whether an action's consequences are considered — govern what choices are made. Notably, the view that the brain contains separate systems for "model-based" vs. "model-free" value computation (which differ in whether or not they recompute values via planning at decision time) offers an influential computational reframing of issues such as habits and compulsion. Yet a realistic "model-based" system must necessarily be selective as to which of many branches are considered (Cushman and Morris, 2015; Keramati et al., 2016), and dysfunction in such selection may extend the reach of these mechanisms to explain symptoms involving biased (e.g., abnormal salience or attention in both compulsive and mood disorders; craving) and abnormal patterns of thought (e.g., rumination, hallucination). The current theory goes beyond just prioritizing planning about the immediate future, to also consider value computation at nonlocal states not immediately implicated in decision, e.g. "offline" replay during sleep or rest (Johnson and Redish, 2005; Ludvig et al., 2017; Sutton, 1990). This systematizes several instances by which tasks typically thought to index "model-based" planning at choice time are apparently solved by computations occurring earlier (Gershman et al., 2014; Momennejad et al., 2017a; Wimmer and Shohamy, 2012) and suggests links between these phenomena and different patterns of replay. Finally, by recasting planning as learning from remembered experience, the theory envisions that the value learning stage of it might be subserved via the same dopaminergic error-driven learning operation long thought to support model-free learning from direct experience. This more convergent picture of the substrates of these two sorts of learning would explain results (puzzling on a separate systems view) that dopaminergic activity is both informed by (Daw et al., 2011; Sadacca et al., 2016) and supports (Deserno et al., 2015; Doll et al., 2016; Sharpe et al., 2017) model-based evaluation.

The AI literature suggests one other candidate hypothesis for the prioritization of backups, known as Prioritized Sweeping (PS) (Momennejad et al., 2017a; Moore and Atkeson, 1993). The idea is that large prediction errors (whether negative or positive) should drive backup to propagate the information to predecessor states. Our approach adds the need term (focusing backups on states likely to be visited again), and also, in the gain term, considers the effect of a backup on an agent's policy, such that propagating information with no behavioral consequences has no value. Data support both of these features of our model over PS: Gain (unlike PS) predicts asymmetric effects of positive and negative prediction errors (Ambrose et al., 2016) (Fig. 3c-f). Because of the need term, our model can also produce searches forward from the current state, in addition to PS's largely backward propagation of error. The need term has a second effect, which is to channel sequential activity along recently or frequently observed trajectories. This may help to explain why nonlocal place cell activity follow extended sequences even though a straightforward error propagation is often more breadth-first (Moore and Atkeson, 1993; Peng and Williams, 1993).

The need term also bears close resemblance to the concept of *need-probability* from rational models of human memory (Anderson and Milson, 1989) — the probability that an item needs to be retrieved from memory because of its relevance to the current situation. Indeed, although we have framed our theory in terms of memory access, our use of a determinisitic, static task moots important distinctions between different sorts of memory, such as episodic and semantic. In particular, replay-based methods can learn equivalently either from remembered experiences (e.g., episodic memories of particular trajectories), or from simulated experiences (e.g., trajectories composed by first learning a semantic map or model of the task, then generating experiences from it, as in model-based RL), blurring the distinction between model-based learning and model-free with stored samples (Vanseijen and Sutton, 2015). In the current setting, prioritizing over experiences, locations, and maps all amount to the same thing, since due to the nature of the task, any episode of going north from a particular location is identical to any other. An important goal of future work will be to tease apart the role of episodic vs. semantic knowledge in computing action values, and understand their relative prioritization (Gershman and Daw, 2017; Lengyel and Dayan, 2008).

There are a number of other limitations to the model, many of which are also opportunities for future work. Though we have used a spatial framing due to the links with hippocampal replay, our theory is formalized generally over states and actions and can be applied beyond navigation to other sequential tasks. However, we omitted many model features to construct the simplest instantiation that most clearly exposes the key intuition behind the theory: the interplay between gain and need and their respective roles driving reverse and forward replay. For instance, we restricted our simulations to two simple environments (a linear track and an open field), and assumed a stationary and deterministic environment that can be learned by the agent without uncertainty — and accordingly omitted these features from the model also. Yet, a full account of prioritized deliberation must surely account for uncertainty about the action values and its sources in stochasticity and nonstationarity. This will require, in future, re-introducing these features from previous accounts of online deliberation (Daw et al., 2005; Keramati et al., 2011); with these features restored, the current theory will inherit its predecessors' successful account of phenomena of habits, such as how they arise with overtraining.

This also relates to perhaps the most important limitation of our work: to investigate the decision theoretic considerations governing replay, we define define priority in the abstract, and do not offer a mechanism or process-level recipe for how the brain would realistically compute it. Although the need term is straightforward (it corresponds to the SR (Dayan, 1993), which the brain has already been proposed to track for other reasons (Momennejad et al., 2017b; Stachenfeld et

al., 2017)), the calculation of gain, as we define it, requires that the agent knows the effect of the backup on its policy prior to deciding whether to perform the backup. We use this admittedly unrealistic decision rule to investigate the characteristics of efficient backup, but a process-level model will require heuristics or approximations to the gain term; here again previous work on deliberation under uncertainty suggests a candidate approximation, called the myopic value of perfect information (Keramati et al., 2011).

To highlight the role of sequencing computations, we have also constructed the theory at a single spatial and temporal scale, focusing on a single Bellman backup as the elementary unit of computation. We build both forward and reverse replay trajectories recursively, step by step, with value information propagating along the entire trajectory. Of course, research in both hippocampus and decision making (separately) stresses the multi-scale nature of task representations; a fuller account of learning, planning and prediction would include temporally extended actions ("options") (Botvinick et al., 2009; Cushman and Morris, 2015; Dezfouli et al., 2014) or similar long-scale state predictions (Dayan, 1993; Sutton, 1995).

4 Methods

4.1 Model description

The framework of reinforcement learning (Sutton and Barto, 1998) formalizes how an agent interacting with an environment through a sequence of states should select its actions so as to maximize some notion of cumulative reward. The agent's policy π assigns a probability $\pi(a|s)$ to each action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. Upon executing an action A_t at time t, the agent transitions from state S_t to state S_{t+1} and receives a reward R_t. The goal of the agent is to learn a policy that maximizes the discounted return G_t following time t defined as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{1}$$

where $\gamma \in (0, 1]$ is the *discount factor* that determines the present value of future rewards.

The expected return obtained by performing action a in state s and subsequently following policy π is denoted $q_\pi(s, a)$ and is given by:

$$q_\pi(s, a) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right] \tag{2}$$

The policy that maximizes the expected return is the *optimal policy* and denoted q_*. Following Q-learning (Watkins and Dayan, 1992), the agent can learn an action-value function Q that approximates q_* through iteratively performing Bellman backups:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_t + \gamma \max_{a \in \mathcal{A}} Q(S_{t+1}, a) - Q(S_t, A_t)\right], \tag{3}$$

where $\alpha \in [0, 1]$ is a learning rate parameter. As in the DYNA architecture, this operation is performed automatically after each transition in real experience, as well as nonlocally during simulated experience (Sutton, 1990).

The following framework provides a rational account for prioritizing nonlocal Bellman backups according to the improvement in cumulative reward expected to result. Let the agent be in state $S_t = s$ at time t. We denote π_{old} the current policy, prior to executing the backup, and π_{new} the resulting policy after the backup. The utility of accessing experience $e_k = (s_k, a_k, r_k, s_{k+1})$ to update the value of (s_k, a_k) is denoted $EVB(s_k, a_k)$ and is defined as:

$$EVB(s_k, a_k) = \mathop{\mathbb{E}}_{\pi_{new}}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] - \mathop{\mathbb{E}}_{\pi_{old}}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] \tag{4}$$

This quantity can be separated into the product of a gain and a need term, i.e. $EVB(s_k, a_k) = Gain(s_k, a_k) \times Need(s_k)$.

4.1.1 Gain term

The gain term quantifies the expected improvement in return accrued at the target state, s_k:

$$
\begin{aligned}
Gain(s_k, a_k) &= EVB(s_k, a_k | S_t = s_k) \\
&= \mathop{\mathbb{E}}_{\pi_{new}} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s_k \right] - \mathop{\mathbb{E}}_{\pi_{old}} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s_k \right]
\end{aligned} \tag{5}
$$

A myopic estimate of this value uses $Q_{\pi_{new}}$ as a stand-in for the true Q-value, and compares the expected cumulative reward under an updated policy (reflecting the Bellman backup) against the previous, un-updated policy (prior to performing the backup). Formally:

$$
Gain(s_k, a_k) = \sum_{a \in \mathcal{A}} Q_{\pi_{new}}(s_k, a) \pi_{new}(a|s_k) - \sum_{a \in \mathcal{A}} Q_{\pi_{new}}(s_k, a) \pi_{old}(a|s_k) \tag{6}
$$

where $\pi_{new}(a|s_k)$ represents the probability of selecting action a in state s_k *after* the Bellman backup, and $\pi_{old}(a|s_k)$ is the same quantity *before* the Bellman backup.

4.1.2 Need term

The need term measures the discounted number of times the agent is expected to visit the target state, a proxy for the current relevance of each state:

$$
Need(s_k) = \mu_\pi(s_k) = \sum_{i=t}^{\infty} \gamma^{i-t} \delta_{s_i, s_k}, \tag{7}
$$

where δ_{s_t, s_k} is the Kronecker delta function. Notice that, for $\gamma = 1$, the need term is the exact count of how many visits to state s_k are expected in the future, starting from state s_t.

The need term can be estimated by the Successor Representation (Dayan, 1993), which can be learned directly by the agent or computed from a model. Here, we assume that the agent learns a state-state transition probability model $\mathcal{T}$ for the purpose of computing the need term. The need term is thus obtained directly from the k-th row of $(\mathcal{I} - \gamma \mathcal{T})^{-1}$.

An alternative option is to use the stationary distribution of the MDP, which estimates the asymptotic fraction of time spent in each state (and is correspondingly easy to estimate from experience). This formulation is particularly useful when the transition probability from the agent's current state is unavailable (e.g., during sleep).

4.1.3 Simulation details

We simulated two "grid-world" environments (Fig. 1b) where an agent could move in any of the four cardinal directions – i.e. $\mathcal{A} = \{\text{up}, \text{down}, \text{right}, \text{left}\}$. At each state, the agent selected an action according to a softmax decision rule over the estimated Q-values, $\pi(a|s) \propto e^{\frac{Q(s,a)}{\tau}}$, where τ is the temperature parameter which sets the balance between exploration versus exploitation. In our simulations, $\tau = 0.2$. Upon selecting action A_t in state S_t, the agent observes a reward R_t and is transported to an adjacent state S_{t+1}. The set of Q-values is then updated according to (3) using $\alpha = 1.0$ and $\gamma = 0.9$.

The first environment was a linear track (Fig. 1b, *left*), which was simulated as two disjoint 1×10 segments. (The motivation for this was for the state space to differentiate both location and direction of travel, as do hippocampal place cells in this sort of environment; this also clearly disambiguates forward from reverse replay.) The agent started in location $(1, 1)$ of the first segment. Upon reaching the state $(1, 10)$, the agent received a unit reward with standard deviation of $\sigma = 0.1$ and was transported to state $(1, 10)$ of the second segment. Upon reaching state $(1, 1)$ in the second segment, the agent received a new unit reward ($\sigma = 0.1$) and was transported back to state $(1, 1)$ of the first segment. Each simulation comprised of 50 episodes (i.e. sequence of steps from starting location to reward).

The second environment was a 6×9 field with obstacles (Fig. 1b, *right*), with a unit reward ($\sigma = 0.1$) placed at coordinate $(1, 9)$. Each simulation comprised of 50 episodes and the start location was randomized at each episode.

The agent was allowed 20 planning steps at the beginning and end of each episode. In each planning step, the agent selected the experience with highest utility EVB. Reactivation of experience $e_k = (s_k, a_k, r_k, s'_k)$ propagates the one-step reward r_k and the value of s'_k to the state-action pair (s_k, a_k) according to (3) (similarly using $\alpha = 1.0$ and $\gamma = 0.9$). Because the gain term is a function of the current set of Q-values, the utilities EVB were recalculated for all experiences after each planning step. In order to ensure that all 20 planning steps were used, a minimum gain of 10^{-10} was used for all experiences.

Successive backups often involved adjacent states, despite the fact that EVB was calculated independently for each individual experience. Because a Bellman backup propagates value directionally (from successor to predecessor state),

successive backups in the reverse direction propagate value information along extended trajectories (i.e. from the destination state to each of a series of predecessors, built up recursively). To make the model symmetric with respect to information propagation over a forward trajectory, we also allowed multi-step forward backups to be built recursively. Specifically, at each step, in addition to each individual experience $e_t = (s_t, a_t, r_t, s_{t+1})$, EVB was also calculated for an expanded experience 1-step longer than the previously executed backup. The candidate experience appended in such cases corresponded to the action sampled from $\pi(a_{t+1}|s_{t+1})$ — i.e., a_{t+1} is an action that would be selected in the resulting state of the previous backup (s_{t+1}), following the same decision rule as in behavior. Formally, appending experiences recursively allowed for n-step forward rollouts, which propagate value information along trajectories in the forward direction (from an end state back to all predecessors), equivalent to the reverse case. EVB for n-step backups was divided by the length n of the trajectory, to account for the fact that n Q-values were updated simultaneously, although all results were equivalent without this division.

Prior to the first episode, the agent was initialized with a full set of experiences corresponding to executing every action in every state (equivalent to a full state-action-state transition model, which in sparse environments like these can be inferred directly from visual inspection when the agent first encounters the maze), including transitions from goal states to starting states. The state-state transition probability model $\mathcal{T}$ (for the need term) was initialized from this model under a random action selection policy, and thereafter updated after each transition using a delta rule with learning rate $\alpha_{\mathcal{T}} = 0.9$. In all simulations in the online setting, the need term was then estimated from the SR matrix, $(\mathcal{I} - \gamma\mathcal{T})^{-1}$. In the only simulation of sleep replay (Fig. 4i,j), where the agent is not located in the environment where need is computed, we estimated the need term as the stationary distribution of the MDP, i.e., the vector μ such that $\mu\mathcal{T} = \mu$.

Acknowledgements

We thank Máté Lengyel, Daphna Shohamy, and Daniel Acosta-Kane for many helpful discussions, and Dylan Rich for his comments on an earlier draft of the manuscript. We acknowledge support from NIDA through grant R01DA038891, part of the CRCNS program, and Google DeepMind. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

References

Ambrose, R. E. et al. (2016). "Reverse replay of hippocampal place cells is uniquely modulated by changing reward". In: *Neuron* 91.5, pp. 1124–1136.

Anderson, J. R. and R. Milson (1989). "Human memory: An adaptive perspective." In: *Psychological Review* 96.4, p. 703.

Botvinick, M. M. et al. (2009). "Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective". In: *Cognition* 113.3, pp. 262–280.

Buhry, L. et al. (2011). "Reactivation, replay, and preplay: how it might all fit together". In: *Neural plasticity* 2011.

Carr, M. F. et al. (2011). "Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval". In: *Nature neuroscience* 14.2, pp. 147–153.

Cheng, S. and L. M. Frank (2008). "New experiences enhance coordinated neural activity in the hippocampus". In: *Neuron* 57.2, pp. 303–313.

Cushman, F. and A. Morris (2015). "Habitual control of goal selection in humans". In: *Proceedings of the National Academy of Sciences* 112.45, pp. 13817–13822.

Davidson, T. J. et al. (2009). "Hippocampal replay of extended experience". In: *Neuron* 63.4, pp. 497–507.

Daw, N. D. and P. Dayan (2014). "The algorithmic anatomy of model-based evaluation". In: *Phil. Trans. R. Soc. B* 369.1655, p. 20130478.

Daw, N. D. et al. (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12, pp. 1704–1711.

Daw, N. D. et al. (2011). "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6, pp. 1204–1215.

Dayan, P. (1993). "Improving generalization for temporal difference learning: The successor representation". In: *Neural Computation* 5.4, pp. 613–624.

Deserno, L. et al. (2015). "Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making". In: *Proceedings of the National Academy of Sciences* 112.5, pp. 1595–1600.

Dezfouli, A. et al. (2014). "Habits as action sequences: hierarchical action control and changes in outcome value". In: *Phil. Trans. R. Soc. B* 369.1655, p. 20130482.

Diba, K. and G. Buzsáki (2007). "Forward and reverse hippocampal place-cell sequences during ripples". In: *Nature neuroscience* 10.10, p. 1241.

Doll, B. B. et al. (2015). "Model-based choices involve prospective neural activity". In: *Nature neuroscience* 18.5, pp. 767–772.

Doll, B. B. et al. (2016). "Variability in dopamine genes dissociates model-based and model-free reinforcement learning". In: *Journal of Neuroscience* 36.4, pp. 1211–1222.
Dupret, D. et al. (2010). "The reorganization and reactivation of hippocampal maps predict spatial memory performance". In: *Nature neuroscience* 13.8, pp. 995–1002.
Foster, D. J. and M. A. Wilson (2006). "Reverse replay of behavioural sequences in hippocampal place cells during the awake state". In: *Nature* 440.7084, pp. 680–683.
Foster, D. et al. (2000). "A model of hippocampally dependent navigation, using the temporal difference learning rule". In: *Hippocampus* 10.1, pp. 1–16.
Gershman, S. J. and N. D. Daw (2017). "Reinforcement learning and episodic memory in humans and animals: An integrative framework". In: *Annual review of psychology* 68, pp. 101–128.
Gershman, S. J. et al. (2014). "Retrospective revaluation in sequential decision making: A tale of two systems." In: *Journal of Experimental Psychology: General* 143.1, p. 182.
Gillan, C. M. et al. (2016). "Characterizing a psychiatric symptom dimension related to deficits in goal-directed control". In: *Elife* 5, e11305.
Gomperts, S. N. et al. (2015). "VTA neurons coordinate with the hippocampal reactivation of spatial experience". In: *Elife* 4, e05360.
Gupta, A. S. et al. (2010). "Hippocampal replay is not a simple function of experience". In: *Neuron* 65.5, pp. 695–705.
Huys, Q. J. et al. (2015a). "Depression: a decision-theoretic analysis". In: *Annual review of neuroscience* 38, pp. 1–23.
Huys, Q. J. et al. (2015b). "Interplay of approximate planning strategies". In: *Proceedings of the National Academy of Sciences* 112.10, pp. 3098–3103.
Jackson, J. C. et al. (2006). "Hippocampal sharp waves and reactivation during awake states depend on repeated sequential experience". In: *Journal of Neuroscience* 26.48, pp. 12415–12426.
Jadhav, S. P. et al. (2012). "Awake hippocampal sharp-wave ripples support spatial memory". In: *Science* 336.6087, pp. 1454–1458.
Johnson, A. and A. D. Redish (2005). "Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model". In: *Neural Networks* 18.9, pp. 1163–1171.
– (2007). "Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point". In: *Journal of Neuroscience* 27.45, pp. 12176–12189.
Karlsson, M. P. and L. M. Frank (2009). "Awake replay of remote experiences in the hippocampus". In: *Nature neuroscience* 12.7, pp. 913–918.
Keramati, M. et al. (2011). "Speed/accuracy trade-off between the habitual and the goal-directed processes". In: *PLoS Comput Biol* 7.5, e1002055.
Keramati, M. et al. (2016). "Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum". In: *Proceedings of the National Academy of Sciences* 113.45, pp. 12868–12873.
Kool, W. et al. (2016). "When does model-based control pay off?" In: *PLoS computational biology* 12.8, e1005090.
Kumaran, D. et al. (2016). "What learning systems do intelligent agents need? Complementary learning systems theory updated". In: *Trends in cognitive sciences* 20.7, pp. 512–534.
Kurth-Nelson, Z. et al. (2016). "Fast Sequences of Non-spatial State Representations in Humans". In: *Neuron* 91.1, pp. 194–204.
Lansink, C. S. et al. (2009). "Hippocampus leads ventral striatum in replay of place-reward information". In: *PLoS biology* 7.8, e1000173.
Lee, A. K. and M. A. Wilson (2002). "Memory of sequential experience in the hippocampus during slow wave sleep". In: *Neuron* 36.6, pp. 1183–1194.
Lengyel, M. and P. Dayan (2008). "Hippocampal contributions to control: the third way". In: *Advances in neural information processing systems*, pp. 889–896.
Ludvig, E. A. et al. (2017). "Associative learning from replayed experience". In: *bioRxiv*, p. 100800.
McClelland, J. L. et al. (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." In: *Psychological review* 102.3, p. 419.
Meer, M. A. van der and A. D. Redish (2011). "Theta phase precession in rat ventral striatum links place and reward information". In: *Journal of Neuroscience* 31.8, pp. 2843–2854.
Miller, K. J. et al. (2017). "Dorsal hippocampus contributes to model-based planning". In: *Nature Neuroscience* 20.9, pp. 1269–1276.
Momennejad, I. et al. (2017a). "Offline Replay Supports Planning: fMRI Evidence from Reward Revaluation". In: *bioRxiv*, p. 196758.
Momennejad, I. et al. (2017b). "The successor representation in human reinforcement learning". In: *Nature Human Behaviour* 1.9, p. 680.
Moore, A. W. and C. G. Atkeson (1993). "Prioritized sweeping: Reinforcement learning with less data and less time". In: *Machine learning* 13.1, pp. 103–130.
O'keefe, J. and L. Nadel (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Ólafsdóttir, H. F. et al. (2015). "Hippocampal place cells construct reward related sequences through unexplored space". In: *Elife* 4, e06063.
O'Neill, J. et al. (2008). "Reactivation of experience-dependent cell assembly patterns in the hippocampus". In: *Nature neuroscience* 11.2, p. 209.
Peng, J. and R. J. Williams (1993). "Efficient learning and planning within the Dyna framework". In: *Adaptive Behavior* 1.4, pp. 437–454.
Pezzulo, G. et al. (2013). "The mixed instrumental controller: using value of information to combine habitual choice and mental simulation". In: *Frontiers in psychology* 4.
Pfeiffer, B. E. and D. J. Foster (2013). "Hippocampal place-cell sequences depict future paths to remembered goals". In: *Nature* 497.7447, pp. 74–79.
Russek, E. M. et al. (2017). "Predictive representations can link model-based reinforcement learning to model-free mechanisms". In: *bioRxiv*, p. 083857.
Sadacca, B. F. et al. (2016). "Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework". In: *Elife* 5, e13665.
Schaul, T. et al. (2015). "Prioritized experience replay". In: *arXiv preprint arXiv:1511.05952*.
Schultz, W. et al. (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599.
Scoville, W. B. and B. Milner (1957). "Loss of recent memory after bilateral hippocampal lesions". In: *Journal of neurology, neurosurgery, and psychiatry* 20.1, p. 11.
Sharpe, M. J. et al. (2017). "Dopamine transients are sufficient and necessary for acquisition of model-based associations". In: *Nature Neuroscience*.
Shohamy, D. and N. D. Daw (2015). "Integrating memories to guide decisions". In: *Current Opinion in Behavioral Sciences* 5, pp. 85–90.
Singer, A. C. and L. M. Frank (2009). "Rewarded outcomes enhance reactivation of experience in the hippocampus". In: *Neuron* 64.6, pp. 910–921.
Singer, A. C. et al. (2013). "Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task". In: *Neuron* 77.6, pp. 1163–1173.
Smith, A. et al. (2006). "Dopamine, prediction error and associative learning: a model-based account". In: *Network: Computation in Neural Systems* 17.1, pp. 61–84.
Stachenfeld, K. L. et al. (2017). "The hippocampus as a predictive map". In: *bioRxiv*, p. 097170.
Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming". In: *Proceedings of the seventh international conference on machine learning*, pp. 216–224.
– (1995). "TD models: Modeling the world at a mixture of time scales". In: *ICML*. Vol. 12, pp. 531–539.
Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.
Tolman, E. C. (1948). "Cognitive maps in rats and men." In: *Psychological review* 55.4, p. 189.
Vanseijen, H. and R. Sutton (2015). "A deeper look at planning as learning from replay". In: *International Conference on Machine Learning*, pp. 2314–2322.
Watkins, C. J. and P. Dayan (1992). "Q-learning". In: *Machine learning* 8.3-4, pp. 279–292.
Wimmer, G. E. and D. Shohamy (2012). "Preference by association: how memory mechanisms in the hippocampus bias decisions". In: *Science* 338.6104, pp. 270–273.