Hippocampal replays under the scrutiny of reinforcement learning models

Romain Cazé, Mehdi Khamassi, Lise Aubin, Benoît Girard

Hippocampal replays under the scrutiny of reinforcement learning models

Romain Cazé[a], Mehdi Khamassi[a,b], Lise Aubin, Benoît Girard[b]

Institute of Intelligent Systems and Robotics, Sorbonne Université, CNRS, F-75005 Paris, France.

[a]*Equally contributing authors*
[b]*Correspondence: firstname.lastname@isir.upmc.fr*

Abstract

Multiple *in vivo* studies have shown that place cells from the hippocampus replay previously experienced trajectories. These replays are commonly considered to mainly reflect memory consolidation processes. Some data, however, have highlighted a functional link between replays and reinforcement learning (RL). This theory, extensively used in machine learning, has introduced efficient algorithms and can explain various behavioral and physiological measures from different brain regions. RL algorithms could constitute a mechanistic description of replays, and explain how replays can reduce the number of iterations required to explore the environment during learning. We review here the main findings concerning the different hippocampal replay types and the possible associated RL models (either model-based, model-free or hybrid model types). We conclude by tying these frameworks together. We illustrate the link between data and RL through a series of model simulations. This review, at the frontier between informatics and biology, paves the way for future work on replays.

Keywords: Hippocampus, Place cells, Activity replay, Sleep, Computational modeling, Reinforcement learning, Model-free / Model-based

Introduction

Humans dream but it remains unknown if all animals do. We know, however, that during the night, many species reactivate various brain regions with patterns sometimes mimicking daytime experience (Dave and Margoliash, 2000; Euston et al., 2007; Ji and Wilson, 2007; Lansink et al., 2008; Lee and Wilson, 2002; Ólafsdóttir et al., 2016; Pavlides and Winson, 1989; Wilson and McNaughton, 1994). Hippocampal reactivations in rats provide the highest number of observations. They happen in the CA1 pyramidal cell layer, during network oscillatory events called "sharp wave ripples" (SWR). These last around 50 to 120 milliseconds, during which the local field potential strongly oscillates around 200Hz (Buzsáki et al., 1992; Buzsáki, 2015). During sleep, they appear in the slow-wave sleep periods, but they also exist in wakeful rest. The rat hippocampus hosts place-cells (PC) that encode the position of an animal within the environment (O'Keefe and Dostrovsky, 1971). PCs reactivate during SWRs (Wilson and McNaughton, 1994), and some of the reactivations activate in order, as if they were following credible trajectories in the previously experienced environments (Lee and Wilson, 2002). Such reactivations also happen during quiet awake states (Foster and Wilson, 2006; Karlsson and Frank, 2009) and can then exhibit more complex patterns (Gupta et al., 2010). What could be the role of these reactivations?

Following the proposed role of SWRs in the two-stage memory model (Buzsáki, 1989), researchers have mostly interpreted replays as a sign of a consolidation of memory (Chen and Wilson, 2017; Diekelmann and Born, 2010; Ólafsdóttir et al., 2018; Walker and Stickgold, 2006): they would copy volatile memories for long-term storage into the cortex (Ji and Wilson, 2007; McClelland et al., 1995; Peyrache et al., 2009). The causal role of hippocampo-cortical interactions during sleep in memory consolidation has been recently demonstrated (Maingret et al., 2016). It has also been experimentally shown for a long time that sleep improves learning (Margoliash and Brawn, 2012), in many different domains (sensorimotor learning, perceptual learning, spatial navigation, etc.). Recent results (de Lavilléon et al., 2015; Girardeau et al., 2009) have demonstrated that hippocampal reactivations also play a causal role in reinforcement learning (RL), *i.e.,* the learning processes that seek to maximize expected future rewards through trial-and-error interaction with an environment. Given the fruitful parallels drawn in the past between RL algorithms from the Machine Learning field (Sutton and Barto, 1998) and Neuroscience studies of RL (Schultz et al., 1997), we propose to explore here which RL algorithms are candidates to explain some of the experimentally observed hippocampal reactivations.

In the two main families of RL algorithms (model-based and model-free), off-line (*i.e.,* when the agent is not moving) activations of state representations (position representations in the context of navigation tasks) can essentially be used to accelerate the convergence of learning or to perform trajectory planning in order to guide immediate behavior. Can we disentangle which types of algorithms, in which phases of their operations, are the most suited to explain the various types of observed hippocampal reactivations during sleep or wakefulness?

To address this question, we first rapidly review the experimental observations concerning hippocampal reactivations. Next, we introduce the RL framework and discuss how it exploits reactivations. Finally, we merge experimental and theoretical frameworks to demonstrate how they can (or cannot) fit together, and illustrate this with model simulation results. This review demonstrates that the RL framework can indeed explain many observations and also leads to new predictions.

1. Experimental observations demonstrating hippocampal reactivations

1.1. The hippocampal map

Tolman's seminal work proposes that the brain may sometimes simulate possible outcomes of an action within a mental map of its environment (Tolman, 1948). Tolman made two observations supporting this idea: *vicarious trial-and-error* (VTE) and *latent learning*. The former corresponds to situations where a rat may "hesitate" between two alternative actions (Redish, 2016), one yielding a higher probability of reward, for instance between turning left or right at an intersection in a maze. Tolman proposed that during this type of trials the animal may mentally simulate the possible outcomes of different actions using a cognitive map, in order to evaluate which one is the best action. In contrast, *latent learning* corresponds to situations where an animal is able, after extensive exploration of an environment devoid of rewards, to immediately find the shortest path to a newly introduced reward. He proposed that this ability results from the learning of a cognitive map of the environment during exploration, which can be used to plan the optimal path once the location of rewards is known.

Multiple laboratories have since observed the different cell types underpinning this cognitive map (Hafting et al., 2005; O'Keefe and Dostrovsky, 1971; Taube et al., 1990a,b), among which the place cells of the hippocampus play a central role. John O'Keefe, May-Britt and Edvard Moser conjointly won the nobel prize in 2014 for their work on spatial navigation (Hafting et al., 2005; O'Keefe and Dostrovsky, 1971). They obtained this prize for respectively discovering place and grid cells. Place cells activate when the animal remains in a given location. Grid cells pave the environment by regularly activating at multiple locations forming a lattice. The latter neuron type are in the Medial Entorhinal Cortex whereas the former are in the Hippocampus. Head direction cells display another type of spatial sensitivity (Taube et al., 1990c). These cells fire when the animal faces one direction whatever its position. Co-jointly with grid and place cells, they take part in the animal's spatial representation. But in contrast to the grid and place cells (Ólafsdóttir et al., 2016), we fail to observe replays of these cells' activity.

Wilson and McNaughton (1994) were to our knowledge the first who unraveled the possibility of an activity replay of hippocampal place cells during sleep. They recorded the activity of CA1 pyramidal cells in the hippocampus after rats had explored various apparatus: either a square box or a two-room maze separated by a corridor. Up to 100 neurons were recorded simultaneously during the behavioral exploration (TASK) as well as during slow-wave sleep before (PRE) and after (POST) the task. This enabled the analysis of correlated activity between pairs of neurons. They found that the pairs with correlated activity during TASK showed an increase in correlation during subsequent sleep (POST). These correlations were mostly absent from PRE session sleep. The match between task correlations and those in the POST session sleep gradually decreased session after session, possibly indicating an habituation process or at least a progressive decrease in the need to consolidate memory during sleep following repetitive daily exploration behavior.

This increase in correlation fits with the two-stage model of memory consolidation (Marr, 1971). The first stage would be the exploration of the environment; the second "non-aroused" stage would enable the storage of the information via synaptic potentiation. Following this model, the hippocampus replays waking activity during sleep to consolidate what the rat learned during the day.

1.2. The variety of replays recorded during sleep and awake periods

While the first demonstration of hippocampal experience-dependent reactivation during sleep is due to Wilson and McNaughton (1994), Lee and colleagues were the first to observe fully-fledged replays during sleep (Lee and Wilson, 2002). They recorded simultaneously multiple hippocampal cells during sleep and awake periods. Three rats were trained to run on a linear track during the awake period. In the following sleep period, Lee and Wilson used a decoding algorithm to study the sequential activation of multiple neurons (i.e., about 10). These cells reactivated in the same order as in the awake period but in a much shorter time period. The sequences were played twenty times faster (120ms) than during the awake period (2.4s). This observation of forward replay in the hippocampus during sleep has been replicated by other labs (Roumis and Frank, 2015), and extended to the prefrontal cortex (Euston et al., 2007). There have also been observations of hippocampal replays in the awake state: during immobility periods, when the animal is consuming a food reward (Gupta et al., 2010), when it is waiting at the start of the maze during an inter-trial interval (Diba and Buzsáki, 2007), or when it is preparing a movement towards its starting point or 'home' location within the maze (Pfeiffer and Foster, 2013).

These awake hippocampus (HPC) reactivations seem to be important for decisions that require past experience to be taken into account. For instance, it has been found that disruption of awake HPC reactivations during SWRs in the W-shaped maze only impairs outbound trials – where memory of the previously visited arm is required to know where to go next – but not inbound trials – which just consist of returning to the central arm (Jadhav et al., 2012). Finally, forward sequential activations of hippocampal place-cells have also been observed outside SWRs, when the animal is performing VTE at an intersection (Johnson and Redish, 2007). Interestingly, awake hippocampal *forward replays*, even when occurring during two different types of oscillations such as SWRs and theta oscillations, seem to systematically represent spatial trajectories from the subject's current location to a memorized goal location, and to be at least partly predictive of the animal's future movements. This suggests that awake forward HPC reactivations may possibly reflect a planning mechanism to guide future behavior (Johnson et al., 2007; Pfeiffer and Foster, 2013). Nevertheless, not all awake replays should be seen as directly preparatory or planning future behavior, since they can also happen to reflect past experience in a first environment while the animal is performing another task in a second environment (Karlsson and Frank, 2009).

Replays can also occur in the opposite direction to the one performed by the animal in the environment: Foster and Wilson (2006) observed *backward replays* (cells firing in reverse order to that observed during behavior), a finding that has also been replicated since (Diba and Buzsáki, 2007; Gupta et al., 2010; Karlsson and Frank, 2009). These reverse replays were initially observed during quiet wakefulness, but such replays have since been recorded during sleep (Ólafsdóttir et al., 2016; Wikenheiser and Redish, 2013), albeit less frequently. As with forward replays, backward replays are executed in "compressed time": they can be up to ten times faster than the experience of the same sequence during real exploration (Euston et al., 2007).

Is the hippocampus also able to virtually explore new possibilities? Gupta et al. (2010) have observed the reactivation of place-cells corresponding to novel sequences in a multiple-T-maze: some sequences were decoded which corresponded to a movement from the right reward location to the left one. This movement is physically possible, but was never performed by the animal during the task, as it was allowed to go from the decision point to either the left or the right reward location. These relatively rare events suggest that the hippocampus can also simulate never-experienced trajectories, which we will call *imaginary replays* hereafter.

A recent work from Papale et al. (2016) highlights non-trivial interactions between awake hippocampal replays and behavioral performance. In this work they showed an inverse correlation between the amount of VTE and of SWRs when the animal is at the reward site. Furthermore, they show that the disruption of SWRs increases the number of VTE events. This might suggest that SWRs consolidate memory in a way that can reduce uncertainty for the next plans of actions at the decision point. On top of the off-line role of SWRs, this suggests an influence on the ongoing behavior. Their data support a synthetic hypothesis: SWRs may play a role in the exploration of the cognitive map for decision-making and to sustain the representation of this map.

In summary, multiple studies show that the hippocampus plays a role in learning and using a mental model of the environment to guide future decisions. It can thus be used to explore the environment both on-line and off-line, and to help the animal orient itself. Nevertheless, the experimental results reviewed so far do not tell us whether replays serve to explore this mental map in order to either maintain it or to plan decisions ahead. Reviewing the interaction of these replays with other brain areas can help address this issue.

1.3. Dialogs between brain areas

Peyrache and colleagues recorded joint hippocampus-prefrontal cortex reactivations during sleep in rats before and after a binary decision-making task in a Y-maze (Peyrache et al., 2009). The task involved a series of unpredictable changes in the task rule (i.e., reward on the left arm; reward on the lit arm; reward on the right arm; etc.), similar to a Wisconsin Card Sorting Test, so that animals constantly had to relearn the task rule and never develops habits. They recorded the local field potential in the hippocampus and prefrontal cortex as well as single unit activities in the prefrontal cortex. During slow-wave sleep after the task (POST), Peyrache and colleagues found reactivations of prefrontal cortex single-units in conjunction with SWRs in the hippocampus, which were not present during sleep before the task (PRE). This suggests the formation of cell assemblies during performance of the task, which are then reactivated during sleep for consolidation. These increases in prefrontal cortex replays from PRE to POST were specific to sessions where the animal learned the correct task rule. No significant increase in replay was found in sessions without relevant behavioral events, nor in sessions where the animals did try new behavioral strategies but did not find the correct rule of the task. This suggests that performance monitoring processes (here the detection of increases in obtained reward rate) may play an important role in tagging cell assemblies which are relevant for the task and should thus later be replayed during sleep for memory consolidation.

Closed loop experiments, where features of ongoing activity are used to trigger stimulation, often help efficiently highlight links between brain areas. Michael Zugaro's team extensively employed this approach to study interaction between brain areas (Girardeau et al., 2009; Maingret et al., 2016). Specifically, they disrupted sharp wave ripples during sleep by stimulating the ventral hippocampal commissure (Girardeau et al., 2009): this procedure impaired spatial learning in the animal and was interpreted as an impairment of memory consolidation from HPC to PFC (but it could as well have resulted from the impairment of off-line RL processes). In further work, the same group stimulated the PFC simultaneously with an HPC sharp wave ripple during sleep. This enhanced the performance of animals in a difficult recognition task (Maingret et al., 2016). These results show that HPC and PFC dialog to reinforce the memory acquired during the day by the hippocampus. This argues in favor of the two stage model of memory. An alternative explanation could be that the hippocampus would be the model of the world describing the state and the cortex would propose the actions to be taken. Replay during sleep would bind the two together.

A similar closed loop system approach employed by the group of Karim Benchenane has causally demonstrated the role of place cell reactivations in learning by coupling these reactivations to ventral tegmental area (VTA) dopamine-based reward signals during sleep (de Lavilléon et al., 2015). They recorded from the HPC and stimulated the main bundle of axons from the VTA each time a targeted HPC place cell was reactivated during sleep. This stimulation created a place preference for the place field of the cell they used as trigger: Mice were four to five-fold more likely to stay in this place field during the following awake period. This also demonstrates that the VTA can exert a strong influence during sleep on memories of place-reward associations. Such an influence may not necessarily be directly from VTA to HPC, since less coordinated VTA activity with HPC has been reported during sleep than during awake rest (Gomperts et al., 2015). But it could well be through the ventral striatum (VS), which receives both reward signals from the VTA (Lammel et al., 2011) and place information from the HPC (Albertin et al., 2000).

Along these lines, Lansink and colleagues also observed a coupling between HPC and VS (Lansink et al., 2009). Neurons pairs from HPC and VS can reactivate during awake fast forward replay. This is particularly true in pairs for which the HPC neuron was a place cell and the VS neuron was tuned to reward. The HPC place cell fired preferentially before the VS reward-related neuron. This observation provides a mechanism for consolidating place-reward associations by showing that HPC starts the reactivation in a projection area. Khamassi and Humphries (2012) suggested that these experimental results provide striking examples of neural activity that could underly the learning of the so-called "reward function" in the reinforcement learning theory (Sutton and Barto, 1998), that is a memory of which (state,action) couples are statistically associated with reward within the environment, and which constitutes part of the *internal model* learned by model-based methods (see Section 2.2).

In summary, the main experimental results reviewed here suggest (1) a key role for hippocampal reactivations in memory consolidation and learning, and (2) tight interactions during these reactivations between HPC, PFC, VS and VTA.

2. Activity replays in reinforcement learning

In the context of artificial intelligence, reinforcement learning is the problem of learning the policy maximizing the sum of future rewards, using reward and punishment signals (Kaelbling et al., 1996; Sutton and Barto, 1998). This requires the learning system to learn by trial and error: it is distinct from unsupervised learning, where statistical regularities in the inputs are learned without a reward signal, and from supervised learning, where a precise error signal is provided to evaluate each result. Solving RL problems requires efficiently trading off exploration and exploitation: as the relations between actions and subsequent rewards are not known a priori, but have to be discovered by effectively interacting with the environment, exploration has to be performed, especially when beginning to learn a new task. However, when the contingencies become well-known, it is preferable to exploit the acquired knowledge, i.e. to favor choosing the actions estimated as maximally rewarding, and to stop wasting time exploring. Finally, a common situation in many RL problems is that the reward/punishment signals are scarce: many choices have no associated feedback. After a long sequence of actions without feedback, it is therefore challenging to correctly distribute the merits of a reward or punishment feedback to the actions in the sequence that effectively contributed to the results. This is known as the *credit assignment problem*.

Among the diverse algorithms proposed to solve reinforcement learning problems (Sutton and Barto, 1998), two main families have had a strong influence on the neuroscience of decision-making: model-based reinforcement learning (MB-RL) and model-free reinforcement learning (MF-RL). The distinction between the two mainly relies on considering whether decisions are made through the use of an internal model of the task or not. Using a model to simulate alternative action sequences before deciding allows faster convergence and extended adaptability, at the cost of larger computational costs (Chavarriaga et al., 2005; Daw et al., 2005; Dollé et al., 2010; Dollé et al., 2018) and decision time (Keramati et al., 2011; Viejo et al., 2015). Conversely, model-free decisions rely on the slow accumulation of feedback through trial and error. Specifically, it consists in progressively updating action values through reward prediction error signals, a process called temporal-difference (TD) learning, which might explain dopaminergic activity (Schultz et al., 1997). Many experimental results obtained in the study of navigation can be interpreted in the light of this distinction (Khamassi and Humphries, 2012). It is particularly relevant here because, as we will argue, some hippocampal offline reactivations could be interpreted as model-based while some others could not. Therefore, we describe hereafter how the two families of RL algorithms work before drawing possible links with hippocampal replays.

The formalization of reinforcement learning is straightforward: an agent interacts with an environment by executing actions chosen in a set $\mathcal{A} =$

$\{a_0, \ldots a_n\}$, and receives two kinds of signals, observations about the state of the world $o(t)$ (which may describe the environment's state only partially) and reward/punishment $r(t)$ (which usually takes negative values for punishments). The goal of the agent is to learn the function $a = \pi(o)$, also known as the *policy* function, which allows it to choose, for all possible observations, the action that maximizes the utility V. The utility is usually defined as a discounted sum of future rewards (the discount factor γ takes values in $[0, 1[$):

$$V(t) = \sum_{k=0}^{+\infty} \gamma^k r(t+k) \tag{1}$$

The reward signal is commonly sparse in time: a long sequence of actions can be responsible for a single outcome provided at its very end. Identifying which actions are thus responsible or not for this outcome is a non-negligible part of the problem. This is why from a normative point of view, an agent which has just performed an action at time $t-1$ should not only consider the immediate reward r at time t – which corresponds to the first term in this equation, obtained with $k = 0$ –, but also rewards that may occur after a delay ($k > 0$). Nevertheless, the value of an action which is followed by multiple outomes – *e.g.,* a negative reward at time $k = 0$ and a positive reward at time $k = 5$ – should depend more on immediate outcomes than on delayed ones. This is the role that the discount factor γ plays in this equation, giving more weight to an outcome r_1 occurring at k_1 than to an outcome r_2 occurring at k_2 if $k_1 < k_2$.

2.1. Model-free reinforcement learning

To tackle the problem of maximizing V, the model-free family of RL algorithms builds on the observation that the definition of V is recursive: Equation 1 can indeed be rewritten as $V(t) = r(t) + \gamma V(t+1)$. These algorithms aim at predicting the value of V at each timestep t so as to always be able to choose the action that predicts the largest accumulated reward in the future. Should the learning of these predictions $\hat{V}$ converge (*i.e.,* stabilize after learning), we should then have $0 = r(t) + \gamma\hat{V}(t+1) - \hat{V}(t)$ for all t, after moving $\hat{V(t)}$ to the rightside of the equation. This defines a *temporal difference* between two consecutive estimations of value at times t and $t+1$, also known as the *reward prediction error* $\delta = r(t) + \gamma\hat{V}(t+1) - \hat{V}(t)$. This is a key quantity which should be null after learning and should drive the direction of value updates depending on its sign and magnitude during learning: if δ is positive – which corresponds to a positive reward prediction error, or positive “surprise” –, the value estimation of the considered observation should be increased; if negative, it should be decreased; if null, no update in value should be done as it corresponds to a situation where the outcome is as expected.

If we consider the specific Q-learning algorithm (Watkins, 1989), which is one among many similar ways of implementing this idea, the estimation of the future return is computed with a function $Q(o, a)$, which means that we consider that each (observation,action) couple has a specific value that should be learned by the agent through trial-and-error. Each time the agent observes o and tries action a, it will receive a reward signal r (most often equal to zero except at the reward site) and a new observation o'. It will thus be able to update the previous estimation of Q at the *learning* phase as follows:

$$Q(o, a) \leftarrow Q(o, a) + \alpha \times (r + \gamma Q(o, a) - max_{i \in \mathcal{A}} Q(o', i)) \tag{2}$$

Where α is the learning rate. This corresponds to the *learning* phase of the algorithm, where the experience of an interaction with the environment affects the internal representations of the agent.

In addition to the *learning* phase, we will distinguish here two other phases: *inference* and *action selection*. The distinction between the *inference* phase and the *action selection* phase can be seen as equivalent to the distinction between *valuation* and *decision-making* in the field of Neuroscience of decision-making (Padoa-Schioppa and Assad, 2006; Lebreton et al., 2009; Lopez-Persem, 2016). In MF-RL, the *inference* phase of the algorithm is minimal: it simply consists in retrieving the previously learned Q values, for the observation o at hand and for all the possible actions in $\mathcal{A}$. This could involve some computations, for example if the Q values are represented by a multiple-layer neural network (see section 2.4 below), but the computation time of this process still remains quite limited compared to the one required for the tree search process (Daw et al., 2005) in model-based (MB) RL (see next subsection). This is because in most decision-making tasks in Neuroscience, the *inference* phase of MF-RL simply consists in reading from a table the Q values of a small finite set of actions, so as to compare them and make a decision (O’Doherty et al., 2004; Pessiglione et al., 2006; Palminteri et al., 2015; Bavard et al., 2018). In contrast, in MB-RL, retrieving the Q values requires some iterative tree search process where one looks into the future through the model in order to estimate the possible long-term consequences of immediate actions (Daw et al., 2005; Keramati et al., 2011; Lesaint et al., 2014). If this tree search process indicates that one of the immediate actions can be the beginning of an action sequence leading to reward, then its Q value will be high compared to alternative immediate actions. And then a decision can be taken. If the task is multi-step, the larger the number of steps in the action sequence until reward, the higher the *inference* time in MB-RL. Thus, because of this important difference in the number of computations that have to be done to retrieve Q values in MB-RL compared to MF-RL, in the rest of this manuscript we will consider for simplicity that the *inference* phase of MF-RL is negligible compared to that of MB-RL.

From these retrieved Q values, a last *action selection* phase has to be carried out. This requires balancing between two necessities: sampling all the possible actions (exploring), so as to be able to evaluate their real value, and choosing the one with the largest value (exploiting), so as to maximize the utility. A commonly used method is to draw the next action from a probability distribution computed with the softmax function:

$$P(a|o) = \frac{e^{\beta Q(o,a)}}{\sum_{i \in \mathcal{A}} e^{\beta Q(o,i)}} \tag{3}$$

With β the parameter that regulates the compromise between exploration and exploitation: the closer to zero, the more differences between the Q values will be attenuated, and the more the selection will thus be uniform (hence exploratory); conversely, large values (that can go up to infinity) will enhance the contrast between the Q values and will thus favor exploitation of the largest one.

The computations that have to be carried out at each time step are fully defined by these two simple equations, meaning that Q-learning (and TD-learning in general) is quite cheap, from a computational point of view. The counterpart of this is its relative slowness to converge, and to re-adapt in the case of non-stationary contingencies. This is because it is not making full use of the information provided by the interaction with the environment. With this regards, the model-based RL algorithms that will be presented in the following section are much more information-efficient. However, a simple way to improve the MF-RL algorithms is to introduce *experience replay* (Lin, 1992), which is reminiscent of the offline hippocampal activations. It consists in storing quadruplets (o, a, o', r) called *experiences*, containing an experienced observation o, the action taken a, the resulting observation o'

and reward signal r, while interacting with the environment, and replaying them off-line in order to accelerate learning. Also note that, while in the most basic implementations of MF-RL *experience replay*, the experiences to be replayed are chosen randomly, the back propagation of an outcome in long sequences of actions is sped-up when sequences of experiences are replayed backwards, starting from the rewarded transitions (Lin, 1992). In Section 3.1, we will present simulations of different ways to do model-free offline reactivations, in order to analyze their properties and facilitate their comparison with experimentally observed hippocampal reactivations. We will group these under the term *MF-RL replays* because they reactivate some elements of past experience during the learning phase in order to bootstrap learning. In contrast, the offline reactivations of model-based methods presented in the next section will be called *MB-RL inferences*, because they do not replay past experience but rather generate mental trajectories with their model during the inference phase in order to plan a sequence of actions which maximizes reward while minimizing the number of moves.

2.2. Model-based reinforcement learning

In Model-based RL algorithms, the *learning* process aims at building a world model, i.e., a model of how the world changes when actions are taken. This model is usually decomposed into a transition function and a reward function. The transition function $T(o, a, o')$ represents the probability of observing o' next, if action a is taken while observing o. In a discrete case, it can for example be built by storing the number of times each (o, a, o') triplet was encountered. The reward function $R(o, a, o')$ represents the average reward signal experienced when effectively performing the (o, a, o') transition.

In the *inference* phase of MB-RL algorithms, the rewards from the reward function are propagated in the graph defined by the transitions, so as to be able to compute the Q values for any observation (including the current one). A decision can then be made, for example with the same softmax function used in MF-RL (Eqn. 3). This MB-RL *inference* phase can be performed in many different ways, one of the simplest being called *Value Iteration*: it consists in repeatedly updating the Q values of all possible (o, a) combinations by computing a one-step-ahead value prediction:

$$Q(o, a) \leftarrow R(o, a) + \gamma \sum_{o'} T(o, a, o') max_{k \in \mathcal{A}} Q(o', k) \qquad (4)$$

until convergence is obtained.

These updates can be unordered, though it is more efficient to start from rewarding (o, a) combinations, and progressively propagate their value to their predecessors first. This leads to the more general idea of *prioritized sweeping*: update first the observations whose value has changed recently, with a priority given to those o that were associated with the highest absolute Q-value update Δ (after applying Eqn. 4), and to their predecessors. Because the predecessors of a given state o can be difficult to determine in a stochastic world, (Moore and Atkeson, 1993) propose to consider as predecessors all the states o' which have, at least once in the history of the system, performed a one-step transition $o' \rightarrow o$. The priority associated to a predecessor o' can thus be $T(o', a, o)\Delta$.

The opposite optimization can also be used: rather than trying to update values for all observations, most of which are not going to be visited, concentrate on the current situation by updating the values starting from the current observation (*i.e.,* the current estimated position of the animal within the environment) and considering its successors (a strategy called *trajectory sampling* (Sutton and Barto, 1998)).

Finally, inspiration can be even taken from traditional planning techniques (Pohl, 1971), by applying these two strategies simultaneously (*bidirectional search*) with the hope that the forward and backward explorations connect before the whole space has been mentally covered. If the inference phase is to be executed with a limited budget (i.e., a limited number of Q value updates), rather than up to convergence, the *prioritized sweeping* and *trajectory sampling* heuristics usually make better use of this budget than an unordered selection of the updates (Sutton and Barto, 1998).

Nevertheless, it is important to keep in mind that the repeated propagations of value, which are the core of the MB-RL methods, become extremely costly as the number of possible actions and observations grows. These computations are formally equivalent to computing the shortest paths in graphs or to plan – which can be called *tree search* (Daw et al., 2005). Heuristics can be developed to improve the speed of computations, but they remain intrinsically costly.

An advantage of the MB-RL algorithms over the MF ones is that they learn the structure of the environment, rather than directly learning the Q-values governing the policy. As such, what is learned can be re-used if the environment changes. Suppose for example that in a given maze the position of the reward changes: the unchanged structure of the maze, stored in the transition function, is still correct and can be re-used; the reward function only has to be updated, which can be done with only a few attempts at unsuccessfully getting reward at the previous site. Symmetrically, if one corridor of the maze is blocked, but nothing else changes, the update of the now unusable transition will be sufficient to correctly update the Q-values in the whole maze in one-shot, which provides a possible explanation for Tolman's observation that animals then most of the time directly shift to unobstructed paths, because the use of a model may have enabled them to directly infer all other obstructed paths (Martinet et al., 2011). As a consequence, adapting to a new task will be much faster than with MF-RL methods, which directly learn the Q-values specific to a given task, and have to be fully re-learned when something changes in the environment.

2.3. Dyna algorithms

A third family of RL algorithms, the Dyna one, is of particular interest as it can be described as a hybrid between model-free and model-based learning strategies, and as it makes use of off-line reactivations. In Dyna algorithms, on-line learning – when the agent acts in the environment – is performed using model-free updates. Moreover, the *inference* and *action selection* phases are the same as in MF-RL: They simply consist of retrieving Q values from a table and comparing them with softmax (Eqn. 3) to make a decision. But an additional off-line learning phase allows for model-free updates applied to data provided by an internal world model, identical to the one used by MB-RL during the inference phase (Sutton, 1990). The idea is the following: if acting in the real environment is costly (because of energy expenditure, time consumption or lethal risks), it becomes advantageous to build a world model from the real experiences, and to use it to simulate agent-environment interactions at a lower cost. While *experience replays* used in pure MF-RL algorithms employ only experiences that were effectively accomplished, Dyna algorithms will simulate virtual experiences generated by their world model.

Of course, the off-line learning phases of the algorithm, which is very similar to a MB-RL, can make use of the same *prioritized sweeping* (as proposed by Moore and Atkeson (1993); Peng and Williams (1993)) and *trajectory sampling* ordering of virtual experiences to try to improve convergence speed.

For a given replay budget, Dyna algorithms are less efficient than experience replay when the task is static (Lin, 1992). This is because a Dyna algorithm has to learn the world model, and as long as this learning has not converged to a good world-model, the virtual experiences generated by this world model may be erroneous. By definition, *experience replay* refers to correct experiences, as they were experienced in the real world at a given moment in the past. In that case, why bothering building a world-model and using a Dyna algorithm? First, the memory requirements necessary to store past experiences may rapidly grow larger than the more compact world-model representation. Second, if the task changes (modification of the reward site, modification of the topology of the maze, etc.), Dyna algorithms will allow for fast adaptation of the behavior, because of their similarity with MB-RL algorithms, while MF-RL with experience replay will suffer from the same kind of slow adaptation as MF-RL ones. Experience replay may even worsen the performance of MF-RL, as the replay of past outdated experiences, corresponding to the previous configuration, will tend to cancel the learning resulting from new experiences.

2.4. RL with neural networks

Most of the aforementioned algorithms have been first developed to operate in discrete, and most often noiseless, simulated worlds (Fig. 1,A,C), for the sake of simplicity as well as for the possibility to mathematically prove their convergence in such contexts. However, real-world applications of these AI techniques, and their use as realistic models of animal learning capabilities, require value functions and world models to be implemented with function approximators. Among these, multiple-layer neural networks (composed of McCulloch and Pitts (1943) computing units) are a popular choice because of their versatility: they have been widely used since the 90s in the machine learning community (Lin, 1992; Sutton, 1996; Tesauro, 1995), and can be considered a sensible choice when dealing with animal data.

——————— FIGURE 1 ABOUT HERE ———————

However, replacing tables for neural networks has a cost: the update of the Q-value, of the transition function T or of the reward function R is now enforced through back-propagation, and this algorithm requires the successive samples used for training to be uncorrelated. If such training is made online, after each action of the agent/animal, these samples are likely to be highly correlated. This is especially true if we consider a rodent navigating a maze where its movements are restricted, like a t-maze: only a few identical sensorimotor sequences will be repeated over and over. In such a case, the convergence of the learning process is not guaranteed: even if it will still work in many cases (see for example Lin (1992)), it may fail in simple navigation setups. In Figure 2 we illustrate this failure with the learning of the reward function (mapping the current state and a given action to the predicted reward) in a simulated version of the Gupta et al. (2010) task, where three contingencies have to be learned: rewards are always on the right, always on the left, or alternate. Because of the temporal correlations, when trained on-line, the predictions of reward are erroneous (Fig. 2, left). A solution to this problem is to rather operate off-line: to store the experienced successions of observations, actions and rewards, and to use *experience replay* to train the networks on an unordered set of samples. This strategy is one of the core components of recent spectacular achievements by deep RL (Mnih et al., 2015). In our example, the reward function becomes almost perfect (and good enough to allow learning) when trained off-line, in randomized order (Fig. 2, right). Refer to Aubin et al. (2018) for more details.

In this review paper we focus on high level RL algorithm descriptions, and illustrate our arguments with tabular implementations in discrete worlds, because neuro-mimetic versions of all the considered models do not exist. However, as a general warning, we illustrate here that replacing abstract tables in RL algorithms with approximations of neurons – even crude ones (McCulloch and Pitts, 1943) – forces us to consider additional sources of reactivations. Therefore, improving even further the biological realism of RL replay models may reveal new properties or constraints that do not appear with more abstract models. This highlights the importance of alternating between different modeling levels to gain a more complete understanding of a biological phenomenon. In this specific case, it stresses even further the potential functional role of unordered reactivations. We come back to this issue later.

——————— FIGURE 2 ABOUT HERE ———————

2.5. The neural substrate of reinforcement learning

It has been proposed in the mid 90s, that the Pavlovian and instrumental learning capabilities of mammals could be explained by model-free reinforcement learning algorithms (Barto, 1995; Houk et al., 1995; Schultz et al., 1997). This proposal is rooted in the similarity between dopamine signals recorded during Pavlovian conditioning and the expected variations of the reward prediction error signal, δ, used in MF-RL algorithms (Schultz et al., 1997; Lesaint et al., 2014; Lee et al., 2018). In that scheme, dopamine would be the neuro-modulator carrying this essential teaching signal; under the modulatory control of dopamine, the input synapses of the medium spiny neurons in the striatum would learn and store the values V; and the rest of the basal ganglia would be in charge of selecting actions based on these values. MF-RL models have later been successfully applied to an extended corpus of experimental data, including instrumental conditioning in rodents (Morris et al., 2006; Roesch et al., 2007), but also instrumental learning tasks in humans, through the use of model-based fMRI data analysis approaches (O'Doherty et al., 2004), showing reward prediction error correlates in the human basal ganglia (Pessiglione et al., 2006; Palminteri et al., 2015; Bavard et al., 2018). All these successes contributed in strengthening the popularity of this theory in the Neuroscience field.

Analyzing the vertebrate reinforcement learning capabilities from the sole MF-RL point of view would probably be too behavioristic to explain the phenomena that led Tolman to propose the concept of cognitive maps (Tolman, 1948). Indeed, it has been proposed that MF-RL algorithms are more suitable to explain habitual behaviors, while more flexible behaviors such *goal-directed behaviors* would result from mechanisms akin to MB-RL algorithms (Daw et al., 2005). Surprisingly, the neural substrate of these computations seem to be quite similar to the one of MF-RL, as it would simply involve other cortico-basal loops with the same anatomo-functional organization (Yin and Knowlton, 2006). In the MB-RL context, the possible role of dopamine, and the precise processes that may underly value inference, are still unclear and debated (Daw et al., 2011; Khamassi and Humphries, 2012; Takahashi et al., 2011). Note that the Dyna algorithms presented above, possess a model of the world, but update their estimated values using the same computations as MF-RL, showing a possible implementation of a MB-RL scheme using a MF-RL dedicated substrate.

Should the reactivations of the hippocampus be used to update value estimations in some basal ganglia loops, what could be their communication pathways? A first possibility is to consider the direct connections from the hippocampus to the ventral parts of the basal ganglia, through the nucleus accumbens (Humphries and Prescott, 2010; Thierry et al., 2000; Voorn et al., 2004), which may be important for rapid learning (Bast et al., 2009). A possible indirect pathway, through the ventral medial prefrontal cortex or the orbito-frontal cortex (Goodroe et al., 2018), also exists and may be more implied in incremental learning (Bast et al., 2009). For more details, refer to Khamassi and Humphries (2012) which summarizes these place-encoding-to-behavior-expression pathways, with a specific emphasis on MF-RL and MB-RL.

3. Drawing parallels: Which model for which replay?

When presenting the main categories of RL algorithms (MF-RL, MB-RL and Dyna-RL), we highlighted three main phases in their computations: learning, inference and action selection. Given the way action selection is usually formalized, it does not make use of any type of observation reactivations. We will thus concentrate, for each of these RL categories, on the potential use of reactivations in the learning and the inference processes. Once again for clarity, we will talk about algorithms that generate *MF-RL replays* when they reactivate some elements of past experience during the learning phase in order to bootstrap learning. In contrast, we will talk about algorithms that generate *MB-RL inferences* when they produce offline reactivations which do not replay past experience but rather generate mental trajectories with their internal model during the *inference* phase, in order to plan a sequence of actions that maximizes reward while minimizing the number of moves.

We will try to identify which process of which algorithm can make use of reactivations reminiscent of the hippocampal replays. To do so, we will simply hypothesize that the observations o_i that will be used by the RL algorithms are readouts of the hippocampal activity. We will illustrate our conclusions with simple simulations, where the environment is represented by a set of discrete states (the different positions on a maze). In which case, an observation corresponds to the estimated current position of the agent within the environment.

All illustration simulations[1] have been performed in a discrete version of the multiple T-maze task of Gupta et al. (2010) (Fig. 3). In each simulation, the agent is allowed to perform 50 or 100 trials with the reward located on

[1] Code accessible from https://github.com/MehdiKhamassi/RLwithReplay

the left arm of the maze, followed by a non-signaled task change and another 50 or 100 trials where the reward is always located on the right. MF-RL replays and MB-RL inferences are allowed during each inter-trial interval (ITI) while the animal is waiting in the central arm (unless otherwise mentioned). They are organized into *cycles*, where each cycle consists in replaying the full buffer of observations for MF-RL; or generating an equivalent number of observations with the model for MB-RL. These cycles are repeated over and over again during the ITI until one of the two following criteria is met: either (1) convergence, the cumulated changes (in absolute value) of the Q-values during the cycle do not exceed a certain threshold $\epsilon = 0.01$; or (2) budget used, the number of replay cycles performed during the ITI reaches a certain limit (*e.g.,* 20 replay cycles). Hereafter the budget is infinite, meaning that we continue to do replay cycles during the ITI until convergence of the Q-values, unless otherwise mentioned, in which cases we impose a finite budget.

——————— FIGURE 3 ABOUT HERE ———————

3.1. MF-RL models

3.1.1. Learning process

The learning of MF-RL can be improved by *experience replay* (see section 2.1 and Fig. 4). For a given memorized quadruplet of past experience containing the observation o, the chosen action a, the resulting reward r and the resulting observation o', learning consists of re-computing the reward prediction error δ and re-updating the corresponding value $Q(o, a)$. As such, it does not require that the replay comprises sequences longer than the two observations o and o': the basic implementation of *experience replay* does not require the replay of full trajectories, and could thus be supported by apparently unordered hippocampal reactivations, which from a computational point of view could have the advantage of propagating reward values to all parts of the environment. As explained in section 2.4, unordered experience replay can even be necessary for some learning architectures, like those using neural networks.

———————— FIGURE 4 ABOUT HERE ————————

Nevertheless, as noted by Lin (1992), replaying trajectories backward, starting from the rewarding location, can sometimes speed up learning, this is in line with the initial interpretation of backward replays (Foster and Wilson, 2006), and with the recent observations of Ambrose et al. (2016). However, in our discrete simulations with infinite replay budget, MF-RL with unordered replays produced a similar performance curve to that obtained by MF-RL with backward replays (Fig. 5). The same performance was also obtained with variants of these algorithms which constitute other MF-RL replay methods which to our knowledge have not yet been proposed, and that we tested for the sake of completeness: MF-RL with forward replays (which simply replays the buffer of past observations in the correct order rather than replaying it backwards as proposed by Lin (1992)); MF-RL with prioritized replays (which corresponds to a model-free version of the algorithms proposed by Moore and Atkeson (1993); Peng and Williams (1993) where the buffer of past observations is ordered depending on the absolute value of reward prediction errors – i.e., observations associated with the highest surprise are replayed first – without using a model to propagate replays to topologically proximal states of the environment). These last two methods were tested in order to later compare their performance with their MB-RL counterparts, as presented in Section 3.2. These simulations also enable to illustrate that introducing a budget of a limited number of replay cycles per trial leads to a less noisy performance after learning because the Q-values are fine-tuned over a longer series of consecutive trials, independently from the replay method (Fig. 5). Finally, these simulations show that the same learning curve experimentally observed in different animals may still have different underlying replay mechanisms (here different simulated MF-RL replay methods).

———————— FIGURE 5 ABOUT HERE ————————

Interestingly, the absence of performance differences between the tested MF-RL experience replay methods is mainly due to the small number of discretized states of the maze used for these simulations. In the case of continuous state-space simulations with function approximators such as neural networks as those discussed in Section 2.4, interferences can occur between observations which may lead some particular methods to be more efficient than others (Lin, 1992). Nevertheless, these different variants of MF-RL experience replay still produce different types of reactivations that directly derive from the adopted replay method and which may be classified differently if we attempt, like experimentalists, to a posteriori analyze the simulation data and count how many of the generated replay sequences can be classified as *forward*, *backward* and "other" *unordered* or non-categorized replays.

Fig. 6 illustrates the results obtained when we regrouped observations during all replay cycles into chunks of 3 consecutive observations (Fig. 6A) or 5 consecutive observations (Fig. 6B) and then counted which percentage of these chunks can be classified in each category of replays. A first important result is that both MF-RL forward replay and MF-RL backward replay can sometimes generate replay events that an experimenter would classify as unordered, even when we know that there was no noise behind this simulated process. Thus these MF-RL replay methods could account for some of the experimentally observed apparently unordered hippocampal reactivations.

A second interesting result is that even an MF-RL backward replay method can still sometimes generate some reactivations classified as forward. This happens for instance when the simulated agent moved backward within the central arm, so that the reversed chunk of memorized elements now corresponds to a forward movement according to the task. Importantly, none of the MF-RL methods tested here produced imaginary replays (defined following Gupta et al. (2010), as when the replayed trajectory includes the left and the right arm consecutively without returning to the central arm, which the agent has never performed) except a very few times by chance (Fig. 6), as opposed to some of their MB-RL counterparts described in Section 3.2.

——————— FIGURE 6 ABOUT HERE ———————

3.1.2. Inference process

The inference phase of MF-RL does not need to reactivate previous observations, thus may not explain any type of hippocampal replay.

3.2. MB-RL models

3.2.1. Learning process

In MB-RL, the learning phase consists of learning the world model, composed of a transition and a reward model. When implemented in discrete environments, the world model is made of tables used to enumerate the experienced occurrences of transitions and rewards. In which case, there is no reason to make use of observation reactivations for the learning process of this world model, which should rather be based only on transitions and rewards really observed in the environment. However, as mentioned in section 2.4, if the world model relies on neural networks to predict transitions and rewards, then unordered offline replays will be necessary, as shown in Aubin et al. (2018).

Here an interesting consideration can be made about the availability of a pretraining phase in the multiple T-maze experiment of Gupta et al. (2010) that we reproduced in the discrete state-space simulations shown hereafter with MB-RL and DYNA methods. If we allow a pretraining phase, then the learning phase of the world model will be globally over during the task. In practice, we allow the algorithm to endlessly update its world-model (which can be important to enable adaptation of the model in the case where a change in the maze is introduced, such as the appearance of an obstacle). Nevertheless, because the task is deterministic, the transition probabilities between states of the maze will not change anymore during task performance. In terms of the types of off-line inferences that can be produced by the algorithms (i.e., the types of off-line "reactivations"), this pretraining phase enables the animal to test different trajectories that are not allowed afterwards during task performance, e.g., moving along an arm in reverse order until reaching the central arm. This feature is key to produce "imaginary replays" in the simulations presented below, because the algorithm could not otherwise mentally simulate trajectories that it does not consider as physically feasible, *i.e.,* that are associated with null probabilities in the learned transition function because they have never been performed before. Alternatively, one may consider that even without pretraining, a rat can conceive possible trajectories that it has never experienced itself before by seeing another rat performing these trajectories, or simply by using basic intuitive geometry to mentally visualize new trajectories. Nevertheless, in the current state of the tested algorithms without any of these two features, a prediction of our simulations is that "imaginary replays" would happen much less frequently (i.e., only at chance level) without a pretraining phase.

3.2.2. Inference process

The inference process of MB-RL will necessarily make use of reactivations of couples of successive observations (i.e., moving from one state of the maze to another). These can theoretically be completely unordered, but it is usually not efficient (Sutton and Barto, 1998). The use of *prioritized sweeping* (which consists in searching in priority those states within the model that have been associated with the highest amount of surprise, *i.e.,* largest absolute prediction errors, see section 2.2) or of *trajectory sampling* (which consists in searching within the model for possible action sequences without interruption, rather than randomly picking actions at non-contiguous states, see section 2.2) is much more efficient.

In our simulations of the Gupta et al task, *prioritized sweeping* produces a majority of unordered reactivations (Fig. 7, see also Aubin et al. (2018)), but also a non-negligible amount of backward reactivations (a phenomenon that has also been observed in the neural network-based simulations of Aubin et al. (2018), with similar proportions). Strikingly, these backward reactivations were totally absent from the MF-RL counterpart of this method (i.e., “MF-RL prior” in Fig. 6). Thus, we can conclude that prioritizing the buffer of memorized past experience based on “surprise” – i.e., the absolute value of reward prediction errors – (common to the MF-RL and MB-RL versions) is not sufficient to produce backward reactivations. It is the propagation of these reward prediction errors to state predecessors within the world-model (only present in the MB-RL version) which is key here.

——————— FIGURE 7 ABOUT HERE ———————

By contrast, *Trajectory sampling* produces mostly forward reactivations, but also – and in contrast to the MF-RL forward replay method (Fig. 6) – backward reactivations and a few imaginary and unordered ones (Fig. 7).

Unsurprisingly, the combination of these approaches in *bidirectional search* (which simultaneously searches for possible action sequences starting from the current location, and reversed action sequences starting from a known reward location, until those two search processes intersect, meaning that there exist an action sequence from current location until reward, see section 2.2) produces backward and forward reactivations in a more balanced manner (Fig. 7).

These MB-RL inference methods also have the potential of generating imaginary sequences, as the reactivations they produce are not constrained to previously experienced sequences, in contrast to MF-RL replays. Therefore, the experimental observation of hippocampal replays that go beyond the animal's past experience, such as hippocampal sequences suggesting an unexperienced combination of paths within the maze (Gupta et al., 2010), have so far only been explained in terms of MB-RL, rather than MF-RL, and more specifically in terms of cognitive maps (Wikenheiser and Redish, 2015). Our simulations confirm this MB-RL explanation of "imaginary replays". The last tested MB-RL inference method (*MB-RL-unordered* in Fig. 7A) serves as a control that none of these reactivations categories can reasonably be observed in significant proportions if the inference method is totally random.

——————————— FIGURE 8 ABOUT HERE ———————————

When considering the original versions of the MB-RL algorithms, one could expect that these reactivations would be restricted to awake reactivations, because the algorithms need to infer – and thus generate observation sequences – only when the animal needs to make decisions. A parallel can thus be drawn between the forward mental simulations of a MB-RL with *trajectory sampling*, and the experimentally recorded hippocampal population activity in rats during vicarious trial and error (Johnson and Redish, 2007) (VTE): while animals remain immobile at the decision point of a T-maze, the estimated spatial position decoded from the replayed hippocampal activity suggest mental simulations from the current position along each arm successively (Fig. 7C), which has been interpreted as a consideration of each possible trajectory before deciding (Redish, 2016). Such an idea has, for example, been used in the Pezzulo et al. (2013) rodent navigation model, where the MB-RL component of the agent, which uses *trajectory sampling*, performs forward activations at decision points, akin to those observed during VTE. Along these lines, one might be tempted to allow VTE to only occur at the decision-point of the maze – considering that it is a point of high uncertainty – as we did for the simulations shown in Fig. 7C. If instead we allow the model to reactivate during VTE-like events in any state of the environment, our simulations predict that reactivations will not only occur at the decision-point, but also at different points of the central arm and at the reward locations (Fig. 9). Specifically, the model decided to stop to perform off-line reactivations when the Q-values were found to be variable, and these reactivations were prolongated until Q-values varied less then a certain threshold. Such a variability in Q-values can be seen as an indirect measure of uncertainty. Strikingly, this produces transient increases in the number of reactivations in response to changes in task condition (Fig. 8A), consistent with the increase in VTE in animals after task rule changes (Redish, 2016). Another interesting point about MB-RL forward reactivations is that they do not produce sweeps that become selective to left or right depending on the task condition (Fig. 8B), in contrast to the MF-RL trajectory replay method (Fig. 4) and as observed by Johnson and Redish (2007). Thus, the present simulations illustrate that a model-free explanation of these experimental data should still be considered as open.

——————— FIGURE 9 ABOUT HERE ———————

One can moreover design MB-RL variations operating under budget constraints, for which sleep activations would be beneficial. Indeed, if the number of steps available for inference at a decision point is limited – a relatively natural limit when you consider that in many contexts, reactions cannot be delayed arbitrarily – the Q-values used to make a decision may not have converged, and the decision will be suboptimal. It is then advantageous to use inactivity periods, like sleep, to perform inference up to convergence, to store the resulting Q-values in memory, and to use them as a bootstrap for the inferences of the next awake period. Even if we are not aware of a computational study using such a model, it is quite straightforward to implement, and has to be considered as a possible explanation of sleep sequential replays. Note that such MB-RL variations are formally very close to the off-line updates carried-out in Dyna-RL algorithms, and will thus predict the same kind of replays.

3.3. Dyna models

3.3.1. Learning process

In Dyna algorithms, the on-line learning of the value function is done using MF-RL learning rules, thus Dyna-RL has, on this point, the same properties as MF-RL: it does not require replays. The use of neural networks to store the value function will however require unordered replays to avoid the correlated training samples problem (section 2.4).

The off-line learning phase of the value function uses of a world-model and can benefit from the same improvements as MB-RL on-line inference: *prioritized sweeping*, *trajectory sampling* and *bidirectional search* can be used to improve the performance. These respectively predict unordered and backward reactivations, forward reactivations and a mix of forward and backward ones (Fig. 10). While computationally, these three approaches are acceptable, it is interesting to note that reward magnitude changes affect backward replays only (Ambrose et al., 2016), suggesting that rodent brain may be using *prioritized sweeping* (Foster, 2017). As previously noted for MB-RL, imaginary reactivations can be observed for *trajectory sampling* and, though marginally, for *bidirectional search*. Interestingly, while the proportions of different types of off-line reactivations are not different in the tested Dyna and MB-RL methods, more reactivations are needed in the former before convergence in order to reach the same reward rate in the task, leading to prolonged periods of reactivations in Dyna compared to MB-RL (Fig. 11). This is because the learning rule used in Dyna during off-line reactivations is model-free, while the learning rule used in MB-RL is model-based.

——————— FIGURE 10 ABOUT HERE ———————

Finally, similarly to MB-RL algorithms, if the learning of the world model of Dyna-RL algorithms is based on neural networks, it requires additional unordered replays.

——————— FIGURE 11 ABOUT HERE ———————

3.3.2. Inference process

The Dyna inference phase is done as in MF-RL, by directly comparing the stored values, and thus does not require reactivations.

4. Discussion

In this paper, we have reviewed a series of experimental results about hippocampal reactivations (so called *replays*) in rodents during reward-based maze learning tasks. These replays can either occur during sleep or during awake rest periods. They can be more or less noisy, modulated by reward magnitude, and of particular importance here, they can occur in forward sequences, backward sequences, in imaginary orders (defined following Gupta et al. (2010)), or with an apparent lack of order. Finally, the awake ones occur mostly at specific locations within the environment, such as the decision-point, the reward location or the starting position during inter-trial intervals.

We have then reviewed theoretical work employing the reinforcement learning framework. We have described different families of methods (model-free, model-based, and Dyna) and illustrated with a series of numerical simulations how they could account for different types of hippocampal replays experimentally observed.

Based on these considerations, we summarized in Tab. 1 which of the reviewed RL algorithms are candidate explanations for the currently documented hippocampal reactivations. Our machine learning-based analysis leads us to suggest that the replay phenomenon may not be unified, but rather be composed of various types of replays, subserving various mnemonic and learning functions. This is in line with recent experimental results (Ambrose et al., 2016; Ólafsdóttir et al., 2017). Even if the situation is currently not clear-cut, in the sense that multiple algorithmic explanations can be proposed for some of the experimentally observed reactivation phenomena, and that testable predictions allowing to disentangle them still have to be devised, we can extract a few notable points.

The first important point to discuss is that the use of the term *replay* might sometimes be misleading, as it somehow suggests that a sequence of place cell activations observed in an immobile animal is the reactivation of a past experience. Experimental studies have shown that it is not always the case (Gupta et al., 2010; Pfeiffer and Foster, 2013; Wu and Foster, 2014), as suggested in the reviews of Pfeiffer (2017) and Foster (2017). We have here highlighted that, while MF-RL may indeed benefit from *experience replay* mechanisms, MB-RL and Dyna-RL can use a world model to *generate simulated sequences* that do not correspond to specific sequences experienced in the past. Despite being sometimes similar to replayed sequences, especially in many experimental mazes where movements are highly constrained by corridors, these simulated sequences do not result from the same mechanisms. Could it be that reactivations observed in a single brain region, the HPC, but during different states (awake or asleep) and different types of oscillations (theta, SWRs), rely on different mechanisms (*e.g.,* model-based or model-free), sometimes replaying past experience and sometimes mentally generating new ones, and this even with different ordering or priorization processes? Hereafter, we review what simulations may tell us about these different types of HPC reactivations.

The second important point is: Why does it matter if some hippocampal reactivations observed experimentally can be best modeled as model-free RL, while other reported hippocampal reactivations may be best modeled as model-based RL? First, it can give a better clue about the possible information content of a hippocampal reactivation event. If it is more likely model-free, then this means that the information might be past-oriented and could actually be a replay of previously experienced sequences. If it is more likely model-based, then this means that it might rather be future-oriented, reflecting a prospective mental simulation of possible future actions. Second, because MF-RL and MB-RL involve different types of computations, this can give an indication about which other types of activity one could search for in the brain simultaneously to the hippocampal reactivation, and what communication between areas might be involved at this precise moment. For instance, if the model-free interpretation is more likely, this means that dopaminergic phasic reinforcement signals are likely to occur simultaneously (Gomperts et al., 2015). In contrast, in the case of likely model-based reactivations, such phasic dopaminergic signals are not necessary for learning (Khamassi and Humphries, 2012; Lesaint et al., 2014; Lee et al., 2018).

From the present simulations, it seems that the most reasonable RL explanation of awake forward reactivations is that they result from a process using *trajectory sampling* or *bidirectional search* with an internal world model, which could be either the inference phase of a MB-RL algorithm, in line with the interpretation of the findings of Pfeiffer and Foster (2013); Wu et al. (2017), or the learning phase of a Dyna algorithm. From a computational point of view, these reactivations will not necessarily correspond to the upcoming behavior (a question raised in Pfeiffer (2017); Ólafsdóttir et al.

(2018)). Indeed, the forward reactivations would be used to evaluate the different action options. Even in the context of a limited computational budget, the most promising ones should be explored as a priority, and those options revealed *a posteriori* as less worthy than the others would have a lower probability of being selected. Even further, the presence of such reactivated-but-not-executed forward sequences could be a marker of such inference processes, because reactivating only the actions that will be executed would mean that the best course of action was known before inference, and there was thus no reason to plan in the first place. This is illustrated in our simulations by the reduction of the number of replays to its minimum once the performance has reached a plateau and the uncertainty about which action to perform is reduced – which is indirectly measured by the convergence of action values in our simulation (Fig. 8, Fig. 11). Besides, it is worthy of note that these forward reactivations may well be accompanied by model-free updates as occurs with our simulations of the Dyna version of trajectory sampling. Future experiments are required to disentangle these alternative possibilities.

Importantly, MB-RL algorithms making *inferences* structured by *trajectory sampling* and allowed to make pauses in any state of a maze to perform these reactivations – a common strategy in the machine learning field – will spend most of their inferences not only at the decision-point, but also around the reward locations (Fig. 9). In contrast, when the MB-RL inference method uses *prioritized sweeping*, reactivations should be mainly observed around the reward locations of the maze where the largest prediction errors occur at each condition change (Fig. 12). This first suggests that MB-RL *prioritized sweeping* cannot be a reasonable explanation for experimental results showing numerous HPC reactivations at the decision-point. Secondly, this predicts that if MB-RL *trajectory sampling* is a candidate for explaining HPC awake reactivations during VTE at the decision-point (Johnson and Redish, 2007), experimentalists should also find some HPC forward reactivations at reward locations.

——————— FIGURE 12 ABOUT HERE ———————

Another conclusion which arises from the present simulations is that the most reasonable RL explanation of “imaginary” reactivations (including trajectories or combinations of trajectories that have never been performed by the animal during task performance, *e.g.,* Gupta et al. (2010)) is that they result from an MB-RL or Dyna-RL algorithm which can use its internal world-model to generate novel trajectories along *state transitions* that are judged feasible by the model. Importantly, to enable these novel "imaginary" trajectories to emerge in the model, we had to make the trajectory sampling mechanism more exploratory than the decisions taken during task performance, so that mentally generated trajectories do not stick to the shortest path to reward. If from a computational point of view, we assume the same level of exploration during the MB-RL *trajectory sampling* simulations accounting for awake HPC forward reactivations during VTE (Johnson and Redish, 2007), then simulations do not generate a prevalence of replayed trajectories along the rewarded side of the maze (Fig. 8-middle). In contrast, an MF-RL *forward* replay algorithm can produce such effect due to the episodic memory buffer containing more observations from the rewarded side when the agent reaches a good performance (Fig. 4-middle). Nevertheless, one could hypothesize a less exploratory MB-RL *trajectory sampling* process during awake HPC reactivations than during asleep HPC replays. This would be consistent with the observation that the latter are on average more noisy and less accurate than the former (Roumis and Frank, 2015). Note that even if these distinctions about awake and asleep reactivations, suggested by RL considerations, have not been directly tested yet, Tang et al. (2017) convincingly showed that awake and asleep reactivations probably play different roles.

Our analyses revealed that, in a reinforcement learning context, unordered offline reactivations may be as important as ordered ones. Even if structured reactivations of hippocampal place cells are of particular interest, the reinforcement learning theory tells us that apparently unordered reactivations (where no clear previously experienced sequences can be found) may also have a functional role in reinforcement learning: almost all models have a use for them. In some cases (learning with neural networks) they may even become essential to break the temporal structure of the data, which could otherwise prevent the convergence of learning. Specifically, if one makes the (relatively strong) hypothesis that learning RL-related functions (value, transition or reward functions) with backpropagation-trained neural networks is a good approximation of the real learning processes occurring in the brain, then one is led to draw the following conclusion: all learning replays are safer done unordered. Consequently, the observed HPC sequential reactivations could result from model-based inference only (because these are not detrimental, even if implemented with neural networks). Strikingkly, the need for unordered offline reactivations to learn the internal model was true only in our neural networks implementations – which deals with a continuous state space and thus with interferences between relatively similar observations in different parts of the environment –, but not in our discrete state space simulations. This highlights the importance of alternating between different levels of implementation of the same computational processes to gain further insights about constraints that appear at some levels while remaining hidden at others.

Only a few models seem to have a use for the forward HPC reactivations observed during sleep. These types of reactivations could be accounted for by an MB-RL or a Dyna-RL algorithm with *trajectory sampling* for its inference or learning phase, respectively, as well as by an MF-RL *forward replay* for its learning phase. Nevertheless, several theoretical arguments should be considered which reduce the plausibility of the MB-RL and MF-RL models here. MB-RL inference is classically seen as a way to plan upcoming behavior. There is thus in theory no reason to perform MB-RL inference during sleep, which would imply storing the resulting action plan and model-based action values for performance during subsequent wakefulness, except with the very specific limited budget version of MB-RL we proposed in section 2.2 which could justify the need for additional inference during sleep (hence the gray cells in Tab. 1). In contrast, Dyna-RL algorithms are in principle meant to use model-based inference in order to store updated model-free action values for long-term use. It thus perfectly makes sense to conceive of Dyna-RL reactivations during sleep. Similarly, MF-RL replay methods are plausible candidates for forward reactivations during sleep, because they also consist of updating action values for long-term use. Nevertheless, it should be noted that the MF-RL *forward replay* does not benefit from the same theoretical support as other MF-RL replay methods such as *backward replay* or *unordered replay*. We have simulated the former for completion, so as to analyze its replay dynamics, and found that it performed as well as the latters in our simplified discrete maze simulations. Nevertheless, in more complex tasks such as continuous state spaces, the latters are more efficient (Lin, 1992). The only advantage of the former we could think of would be its parsimony in the case where we assume that the HPC stores ordered sequences of past events so that MF-RL *forward replay* simply preserves this order during reactivation. In contrast, MF-RL *backward and unordered replay* methods require changing the order of the memory buffer so as to reverse it or shuffle it respectively, which might represent an additional computational cost. Further work would be required to assess whether some HPC reactivations might still be consistent with MF-RL *forward replay*.

The present model simulations also inform the possible computational interpretations of HPC backward reactivations. First, even if this has not yet been observed, from a RL point of view, decision-point reactivations (VTE) could as well be backward, or a mix of backward and forward. This constitutes an interesting prediction to test experimentally. Second, HPC backward reactivations have initially been mainly thought as consistent with model-free learning (Foster and Wilson, 2006), because reversing the sequence order is righfully an efficient way to more rapidly propagate value from reward location to preceding state than the forward order (Lin, 1992). Nevertheless, it is interesting to note that the present model simulations still suggest that HPC backward reactivations may also be reproduced by model-based inference and Dyna learning, with either of the following methods: *prioritized sweeping*, *trajectory sampling*, *bidirectional search* (Figure 7, Figure 10). Theoretical considerations can nevertheless disentangle plausible candidates to explain HPC asleep backward reactivations: As mentioned above, asleep reactivations are not followed by immediate action, and thus are less likely to rely on model-based inference. An interesting resulting prediction is that asleep HPC backward replays should be accompanied by model-free action value updates in the striatum (because both MF-RL and Dyna perform model-free learning processes). In contrast, awake HPC backward replays could still be compatible with all three families of models: MB-RL, MF-RL and Dyna.

Ólafsdóttir et al. (2018) have very recently proposed an in-depth review of the hippocampal replay phenomenon, and already stressed the possible multiple roles of replays: they distinguished *memory consolidation* replays from *planning* replays. The latter clearly corresponds to the *MB-RL inference* category (section 3.2.2), and the data they reviewed specifically point at the *trajectory sampling* flavor of inference. Interestingly, our grid of analysis, based on the reinforcement learning algorithm families (model-free, model-based or Dyna), and their components (learning, inference and selection), suggests that their *memory consolidation* category could correspond to multiple distinct mechanisms. Indeed, many of the processes they describe as *memory consolidation* correspond to the *learning* processes of reinforcement learning algorithms, and as presented in sections 3.1.1, 3.2.1 and 3.3.1, they are of two main types. First, MB-RL and Dyna require learning the world model, i.e. the transitions between states (S-R-S associations) and the rewarded states, information that could be stored in the hippocampus and the prefrontal cortex. Second, MF-RL and Dyna require learning the values of states (or action-state), an information expected to be stored in the input synapses of the striatum (see section 2.5). This last point suggests an extension of their conclusions: when studying replays, the modification of memory traces should not be examined in the cortex and hippocampus only, but also in the striatum (Lansink et al., 2009).

Finally, it is important to note that the computational approach adopted here mostly remains at a relatively high level, describing dynamics of information flows that can be related to animal behavioral adaptation, but not straightforwardly to neuronal dynamics within cell assemblies. On this aspect, complementary computational approaches employing spiking neural networks are required to account for the complex neural dynamics that have been observed during hippocampal replay and which suggest intertwined relations and alternations between engagement of fast and slow synapses, which could subserve the progressive stabilization of attractors (Pfeiffer and Foster, 2015). Further investigations are required to draw a proper link between these different levels of computations and contribute to a better understanding of the role hippocampal replays may play in memory consolidation.

Algorithm	step	flavor	Awake (SWR)				Awake (VTE)	Asleep			
			fwd	bwd	img	uno	fwd	fwd	bwd	img	uno
MF-RL	value function learning	vanilla (i.e., without replay)									
		unordered experience replay				■					■
		backward experience replay		■					■		
		forward experience replay	■ ◇				■ ◇	■ ◇			
		prioritized exp. replay				■ ◇					■ ◇
		NN-based value function				■					■
	inference	vanilla									
MB-RL	world model learning	vanilla									
		NN-based world model				■					■
	inference	vanilla (i.e., unordered)				■					■
		prioritized sweeping		■		■			▒		▒
		trajectory sampling	■	■	■	■	■	▒	▒	▒	▒
		bidirectional search	■	■		■	■	▒	▒		▒
Dyna-RL	value function learning	vanilla (i.e., unordered)				■					
		prioritized sweeping		■		■			■		■
		trajectory sampling	■	■	■	■	■	■	■	■	■
		bidirectional search	■	■		■	■	■	■		■
		NN-based value function				■					
	world model learning	vanilla									
		NN-based world model				■					■
	inference	vanilla									

Table 1: Summary of the possible algorithmic explanations (rows) for documented hippocampal reactivations (columns). Reactivations can correspond to forward (fwd), backward (bwd), imaginary (img) sequences, or be unordered (uno). Note that some replays that could be explained by RL algorithms, but that have not been documented yet (like backward sequences at decision points) are not considered. Black cells: the considered algorithm can explain the observed reactivation; Gray cells: in the case of awake limited inference budget, asleep reactivations of the same nature as awake reactivations are expected (see text); ◇: these variants of the algorithms have to our knowledge not been proposed before but were here tested as principles for the purpose of the demonstration.

While this paper illustrates how different types of hippocampal replays could be interpreted in terms of different families of learning methods, it also leads to some experimental predictions which could lead to some future studies to help better characterize this neural phenomenon.

Acknowledgements

The authors would like to thank Mark Humphries for comments on the manuscript. This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 640891. This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

The authors would like to thank Olivier Sigaud, Nicolas Perrin and Ludovic Denoyer for the discussions that enriched this paper.

Figure legends

Fig. 1. **Discrete vs. continuous implementations of reinforcement learning**: illustration with discrete vs. continuous observations in Q-learning. In a world where observations o are discrete, for example in (A) the agent can access directly to the cell number it occupies, the Q-values can be stored in (and retrieved from) a table (C) where observations o and actions a combinations are exhaustively enumerated. With continuous observations, for example the activity of a population of possibly noisy place cells (B), which is akin to a vector of real components, Q-values have to be computed with a function approximator. For example a neural network (D). Figure by Girard, 2017; available at https://doi.org/10.6084/m9.figshare.5616418.v2 under a CC-BY4.0 license.

Fig. 2. **Comparison of on-line versus off-line learning of the reward function with a 2-layer neural network**, in a navigation task simulation similar to the experimental setup of Gupta et al. (2010) (white arrows in the first panel illustrate movement directions in the maze). The task is composed of three reward contingencies: the reward is always on the left ("Left only" panels), or always on the right ("Right only" panels), or alternates between left and right ("Alternate, left lap" and "right lap" panels). When the training is carried out on-line, the temporal correlations between the successive samples prevents the network from correctly learning the reward function, while the same network, once trained off-line with unordered samples, learns it efficiently. Figure by Girard & Aubin, 2018; available at https://doi.org/10.6084/m9.figshare.5822109 under a CC-BY4.0 license.

Fig. 3. **Illustration of simulations within a discrete representation of the multiple T-maze task of Gupta et al. (2010).** This simulation environment is used for all numerical simulations shown hereafter in the paper. The maze has been decomposed into 54 states. The reward is alternatively located at positions (1,5) and (9,5). Depending on the simulation, replays are either allowed only at reward sites (same locations), at the departure state corresponding to position (5,2) or in all states. Here the figure shows the simulation of 10 consecutive trials where an agent is controlled by a MF-RL algorithm with unordered replays. The black line illustrates the noisy simulated trajectory of the agent. The color of the different states indicate the maximum Q-value learned by the algorithm in each state at the end of these 10 trials. Replays in this simulation occur during the inter-trial interval at the departure state located at position (5,2), while the reward is here always located on the left arm, at position (1,5). Note that thanks to replay, a single error on the non-rewarded right arm was sufficient for the algorithm to then stick to the rewarded left arm.

Fig. 4. **Results obtained for 10 simulations of the MF-RL backward replay algorithm.** Confidence intervals show the standard deviation. (Top) The algorithm, which replays the reverted buffer of past observations, with here an infinite replay budget during the inter-trial interval (ITI) of each trial, shows a number of performed replay cycles which sharply increases at the beginning of the task and after a change in reward location (trial 100), while remaining low the rest of the time. (Middle) The fact that the agent quickly learns to go to the correct side of the maze makes the buffer contain observations on the left arm when it is rewarded or on the right arm when rewarded, so that replay sweeps observed during ITI most of the time concentrate on the rewarded arm of the maze. (Bottom) Performing ITI replays boosts learning performance while MF-RL without replays slowly learns to increase reward rate and learns even slower after a change in reward location (trial 100). ln: natural logarithm; prop: proportion; L: left; R: right.

Fig. 5. **Results obtained for 10 simulations of different MF-RL replay methods.** Same conventions as Fig. 4. The four tested MF-RL experience replay methods (unordered replays, backward replays, forward replays and prioritized replays – *i.e.,* replaying observations with highest absolute reward prediction errors without using a model) show neither difference in performance nor in computation time (i.e., Napierian logarithm of the number of replay steps per trial), no matter if they are tested with an infinite budget (A) or with a limited budget of 20 replay cycles per trial (B). Note the less noisy and more optimal performance in the latter case because the Q-values are fine-tuned over a longer series of replay trials while the number of trials performing the task online is the same between these two conditions.

Fig. 6. **Proportions of different types of replays obtained for 10 simulations of different MF-RL replay methods.** Bars show the standard deviation. MF-RL prior: prioritized replays without model. (A) The whole sequence of replayed observations during the full experiment has been subdivided into groups of 3 consecutive observations, which were then classified as either forward, backward, imaginary (following Gupta et al. (2010), when the replayed trajectory includes the left and the right arm consecutively without returning to the central arm, which the agent has never performed during task performance) or other replays. (B) Same analysis performed on subgroups of 5 consecutive observations, hence reducing the probability of observing ordered replays by chance. (C) Example of a sequence of replayed observations by the MF-RL backward algorithm and categorized as *backward replays* by the analysis. (D) Example of a sequence of replayed observations by the MF-RL unordered algorithm and categorized as *other replays*.

Fig. 7. **Proportions of different types of inferences obtained for 10 simulations of different MB-RL inference methods.** Same convention as Fig. 6 (A) The whole sequence of inferences during the full experiment has been subdivided into groups of 3 consecutive observations, which were then classified as either forward, backward, imaginary (following Gupta et al. (2010), when the replayed trajectory includes the left and the right arm consecutively without returning to the central arm, which the agent has never performed during task performance) or other replays. (B) Same analysis performed on subgroups of 5 consecutive observations, hence reducing the probability of observing ordered inferences by chance. (C) Example of a sequence of inferences by the MB-RL trajectory sampling algorithm sequentially covering the left and right arms of the maze, reproducing the experimental results of Johnson and Redish (2007).

Fig. 8. **Results obtained for 10 simulations of different MB-RL inference methods.** Same convention as Fig. 4. (Top) An MB-RL algorithm with either unordered inference (discrete states are drawn randomly; black curve), trajectory sampling (red curve) or prioritized sweeping (blue curve) with infinite inference budget in central arm at each trial before deciding to either go left or right, performs a large number of inference cycles especially at the beginning of the task and after a change in reward location (trial 100). These could be interpreted as moments where the agent takes more time to make a decision, and correspond well to the moments of the task where *vicarious trial and error (VTE)* are commonly observed experimentally (Redish, 2016). (Middle) The fact that trajectories are drawn randomly during off-line inference in the MB-RL trajectory sampling algorithm makes the left and right arms on average equally represented across the experiment, unlike simulation results with MF-RL methods (Fig. 4) and the experimental results of Johnson and Redish (2007). (Bottom) All three MB-RL methods show transient decreases in performance after a change in reward location (trial 100) and then a quick adaptation to the new task contingency.

Fig. 9. **Normalized distribution of the total duration of inferences performed by the MB-RL trajectory sampling algorithm obtained for 10 simulations.** Most off-line inferences occur around the reward locations and in the central arm.

Fig. 10. **Proportions of different types of inferences obtained for 10 simulations of different DYNA methods.** Same convention as Fig. 6. Because the DYNA algorithms tested here employ the same inference methods than their MB-RL counterparts, they show similar proportions of forward, backward and imaginary inferences. Thus their main difference (i.e., model-free and model-based action value updates, respectively) only predict behavioral differences in terms of reaction times without predicting different profiles of off-line “activity replays”.

Fig. 11. **Results obtained for 10 simulations of DYNA with prioritized sweeping compared with the MB-RL version of prioritized sweeping.** Same convention as Fig. 4. Here both algorithms are allowed to perform off-line inferences in all states of the maze without budget constraints. (Top) The DYNA version is computationally more costly than the MB-RL version in that it performs a larger number of off-line inference steps during a larger number of trials. This is because of the model-free learning mechanism in DYNA during both on-line and off-line performance, the world model being here used only to determine state predecessors for the prioritized sweeping process. (Bottom) Nevertheless, both methods perform equally well in terms of reward rate.

Fig. 12. **Normalized distribution of the total duration of inferences performed by the MB-RL prioritized sweeping algorithm obtained for 10 simulations.** Most inferences occur around the reward locations where the largest prediction errors can be experienced after each condition change.

Bibliography

Albertin, S. V., Mulder, A. B., Tabuchi, E., Zugaro, M. B., and Wiener, S. I. (2000). Lesions of the medial shell of the nucleus accumbens impair rats in finding larger rewards, but spare reward-seeking behavior. *Behavioural brain research*, 117(1-2):173–183.

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.

Aubin, L., Khamassi, M., and Girard, B. (2018). Prioritized sweeping neural DynaQ with multiple predecessors, and hippocampal replays. In *Living Machines*, page TBA.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 215–232. The MIT Press, Cambridge, MA.

Bast, T., Wilson, I. A., Witter, M. P., and Morris, R. G. (2009). From rapid place learning to behavioral performance: a key role for the intermediate hippocampus. *PLoS biology*, 7(4):e1000089.

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., and Palminteri, S. (2018). Reference point and range-adaptation produce both rational and irrational choices in human reinforcement learning. *Nature Communications*. To appear.

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

Buzsáki, G. (2015). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188.

Buzsáki, G., Horvath, Z., Urioste, R., Hetke, J., and Wise, K. (1992). High-frequency network oscillation in the hippocampus. *Science*, 256(5059):1025–1027.

Chavarriaga, R., Strösslin, T., Sheynikhovich, D., and Gerstner, W. (2005). A computational model of parallel navigation systems in rodents. *Neuroinformatics*, 3(3):223–241.

Chen, Z. and Wilson, M. A. (2017). Deciphering neural codes of memory during sleep. *Trends in Neurosciences*.

Dave, A. S. and Margoliash, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning. *Science*, 290(5492):812–816.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704.

de Lavilléon, G., Lacroix, M. M., Rondi-Reig, L., and Benchenane, K. (2015). Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nature neuroscience*, 18(4):493–495.

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

Diekelmann, S. and Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114.

Dollé, L., Chavarriaga, R., Guillot, A., and Khamassi, M. (2018). Interactions of spatial strategies producing generalization gradient and blocking: A computational approach. *PLoS computational biology*, 14(4):e1006092.

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., and Guillot, A. (2010). Path planning versus cue responding: a bio-inspired model of switching between navigation strategies. *Biological cybernetics*, 103(4):299–317.

Euston, D. R., Tatsuno, M., and McNaughton, B. L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *science*, 318(5853):1147–1150.

Foster, D. J. (2017). Replay comes of age. *Annual review of neuroscience*, 40:581–602.

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.

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., and Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature neuroscience*, 12(10):1222–1223.

Gomperts, S. N., Kloosterman, F., and Wilson, M. A. (2015). Vta neurons coordinate with the hippocampal reactivation of spatial experience. *Elife*, 4:e05360.

Goodroe, S. C., Starnes, J. M., and Brown, T. I. (2018). The complex nature of hippocampal-striatal interactions in spatial navigation. *Frontiers in Human Neuroscience*, 12:250.

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

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

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 249–271. The MIT Press, Cambridge, MA.

Humphries, M. D. and Prescott, T. J. (2010). The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progress in neurobiology*, 90(4):385–417.

Jadhav, S. P., Kemere, C., German, P. W., and Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science*, 336(6087):1454–1458.

Ji, D. and Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature neuroscience*, 10(1):100.

Johnson, A. and Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189.

Johnson, A., van der Meer, M. A., and Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Current opinion in neurobiology*, 17(6):692–697.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Karlsson, M. P. and Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature neuroscience*, 12(7):913.

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

Khamassi, M. and Humphries, M. D. (2012). Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in Behavioral Neuroscience*, 6:79.

Lammel, S., Ion, D. I., Roeper, J., and Malenka, R. C. (2011). Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron*, 70(5):855–862.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., Joosten, R. N., McNaughton, B. L., and Pennartz, C. M. (2008). Preferential reactivation of motivationally relevant information in the ventral striatum. *Journal of Neuroscience*, 28(25):6372–6382.

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., and Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, 7(8).

Lebreton, M., Jorge, S., Michel, V., Thirion, B., and Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron*, 64(3):431–439.

Lee, A. K. and Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, 36(6):1183–1194.

Lee, B., Gentry, R., Bissonette, G., Herman, R., Mallon, J., Bryden, D., Calu, D., Schoenbaum, G., Coutureau, E., Marchand, A., Khamassi, M., and Roesch, M. (2018). Manipulating the revision of reward value during the intertrial interval increases sign tracking and dopamine releases. *PLoS Biology*.

Lesaint, F., Sigaud, O., Flagel, S., Robinson, T., and Khamassi, M. (2014). Modelling individual differences observed in pavlovian autoshaping in rats using a dual learning systems approach and factored representations. *PLoS Computational Biology*, 10(2):e1003466.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3/4):69–97.

Lopez-Persem, A. (2016). *The Brain Valuation System and its role in decision-making*. PhD thesis, Université Pierre et Marie Curie. PhD thesis.

Maingret, N., Girardeau, G., Todorova, R., Goutierre, M., and Zugaro, M. (2016). Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature neuroscience*, 19(7):959–964.

Margoliash, D. and Brawn, T. P. (2012). Sleep and learning in birds: Rats! there's more to sleep. In *Sleep and Brain Activity*, pages 109–146. Elsevier.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 262(841):23–81.

Martinet, L.-E., Sheynikhovich, D., Benchenane, K., and Arleo, A. (2011). Spatial learning and action planning in a prefrontal cortical network model. *PLoS computational biology*, 7(5):e1002045.

McClelland, J. L., McNaughton, B. L., and O'reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., and Others (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, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience*, 9(8):1057.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, 304(5669):452–454.

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.

Ólafsdóttir, H. F., Bush, D., and Barry, C. (2018). The role of hippocampal replay in memory and planning. *Current Biology*, 28(1):R37–R50.

Ólafsdóttir, H. F., Carpenter, F., and Barry, C. (2016). Coordinated grid and place cell replay during rest. *Nature neuroscience*, 19(6):792.

Ólafsdóttir, H. F., Carpenter, F., and Barry, C. (2017). Task demands predict a dynamic switch in the content of awake hippocampal replay. *Neuron*, 96(4):925–935.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223.

Palminteri, S., Khamassi, M., Joffily, M., and Coricelli, G. (2015). Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Nature Communications*, 6:8096.

Papale, A. E., Zielinski, M. C., Frank, L. M., Jadhav, S. P., and Redish, A. D. (2016). Interplay between Hippocampal Sharp-Wave-Ripple Events and Vicarious Trial and Error Behaviors in Decision Making. *Neuron*, 92(5):1–8.

Pavlides, C. and Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 9(8):2907–2918.

Peng, J. and Williams, R. J. (1993). Efficient learning and planning within the Dyna framework. *Adaptive Behavior*, 1(4):437–454.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., and Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106):1042.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, 12(7):919–926.

Pezzulo, G., Rigoli, F., and Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, 4.

Pfeiffer, B. E. (2017). The content of hippocampal "replay". *Hippocampus.*

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

Pfeiffer, B. E. and Foster, D. J. (2015). Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science (New York, N.Y.)*, 349(6244):180–183.

Pohl, I. (1971). Bi-directional search. *Machine intelligence*, 6(127-140):10.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3):147–159.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, 10(12):1615.

Roumis, D. K. and Frank, L. M. (2015). Hippocampal sharp-wave ripples in waking and sleeping states. *Current opinion in neurobiology*, 35:6–12.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

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*, pages 216–224.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in neural information processing systems*, pages 1038–1044.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

Takahashi, Y. K., Roesch, M. R., Wilson, R. C., Toreson, K., O'donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nature neuroscience*, 14(12):1590.

Tang, W., Shin, J. D., Frank, L. M., and Jadhav, S. P. (2017). Hippocampal-prefrontal reactivation during learning is stronger in awake as compared to sleep states. *Journal of Neuroscience*, pages 2217–2291.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990a). Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990b). Head-direction cells recorded from the postsubiculum in freely moving rats. ii. effects of environmental manipulations. *Journal of Neuroscience*, 10(2):436–447.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990c). Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 10(2):436–47.

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68.

Thierry, A.-M., Gioanni, Y., Dégénétais, E., and Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: Anatomical and electrophysiological characteristics. *Hippocampus*, 10(4):411–419.

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

Viejo, G., Khamassi, M., Brovelli, A., and Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in behavioral neuroscience*, 9.

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., and Pennartz, C. M. (2004). Putting a spin on the dorsal–ventral divide of the striatum. *Trends in neurosciences*, 27(8):468–474.

Walker, M. P. and Stickgold, R. (2006). Sleep, memory, and plasticity. *Annu. Rev. Psychol.*, 57:139–166.

Watkins, C. (1989). *Learning from delayed rewards.* PhD thesis.

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.

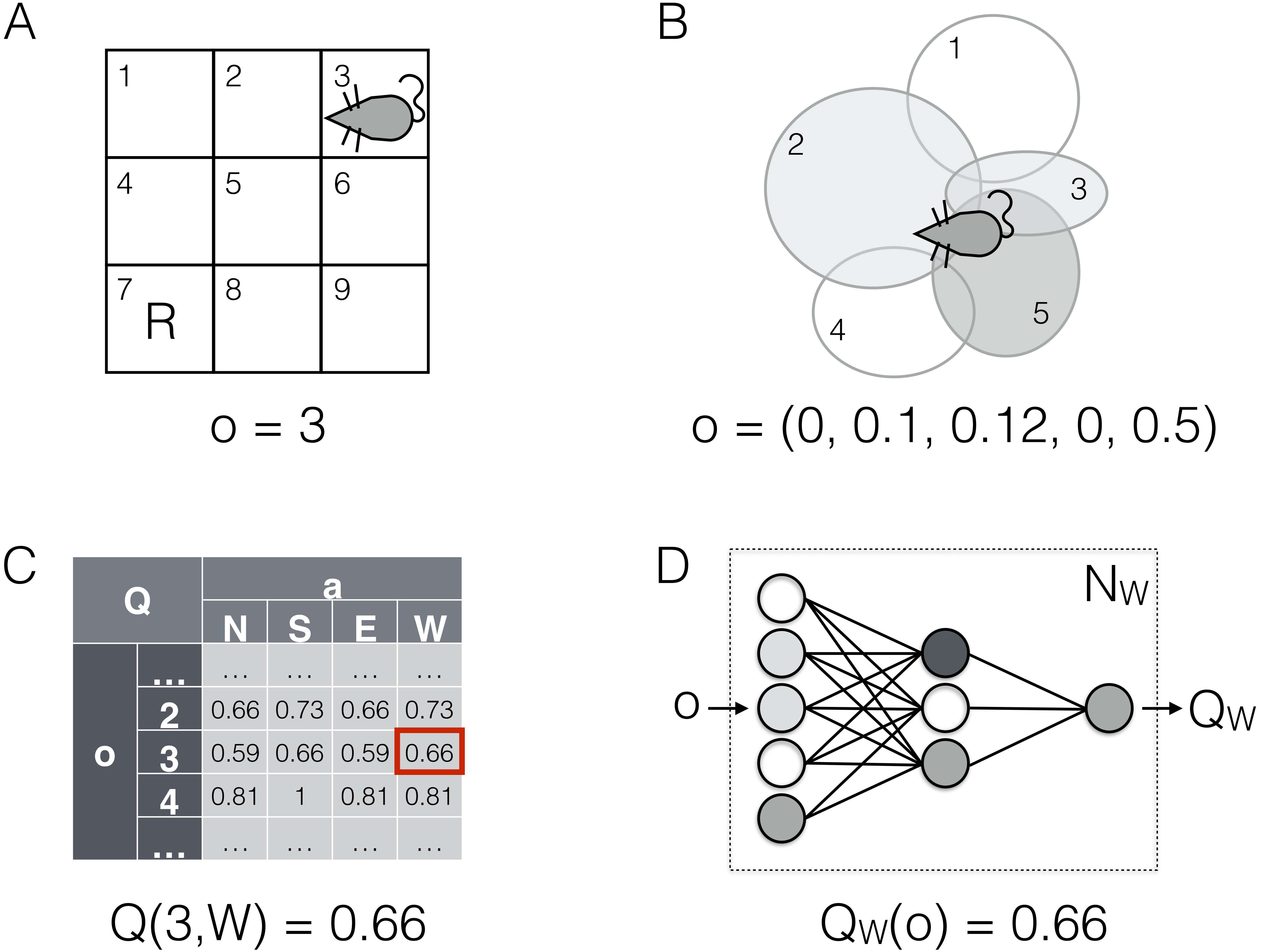
Wikenheiser, A. M. and Redish, A. D. (2015). Hippocampal Sequences and the Cognitive Map. In *Analysis and Modeling of Coordinated Multi-neuronal Activity*, pages 105–129. Springer.

Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science (New York, N.Y.)*, 265(5172):676–679.

Wu, C.-T., Haggerty, D., Kemere, C., and Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nature neuroscience*, 20(4):571.

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.

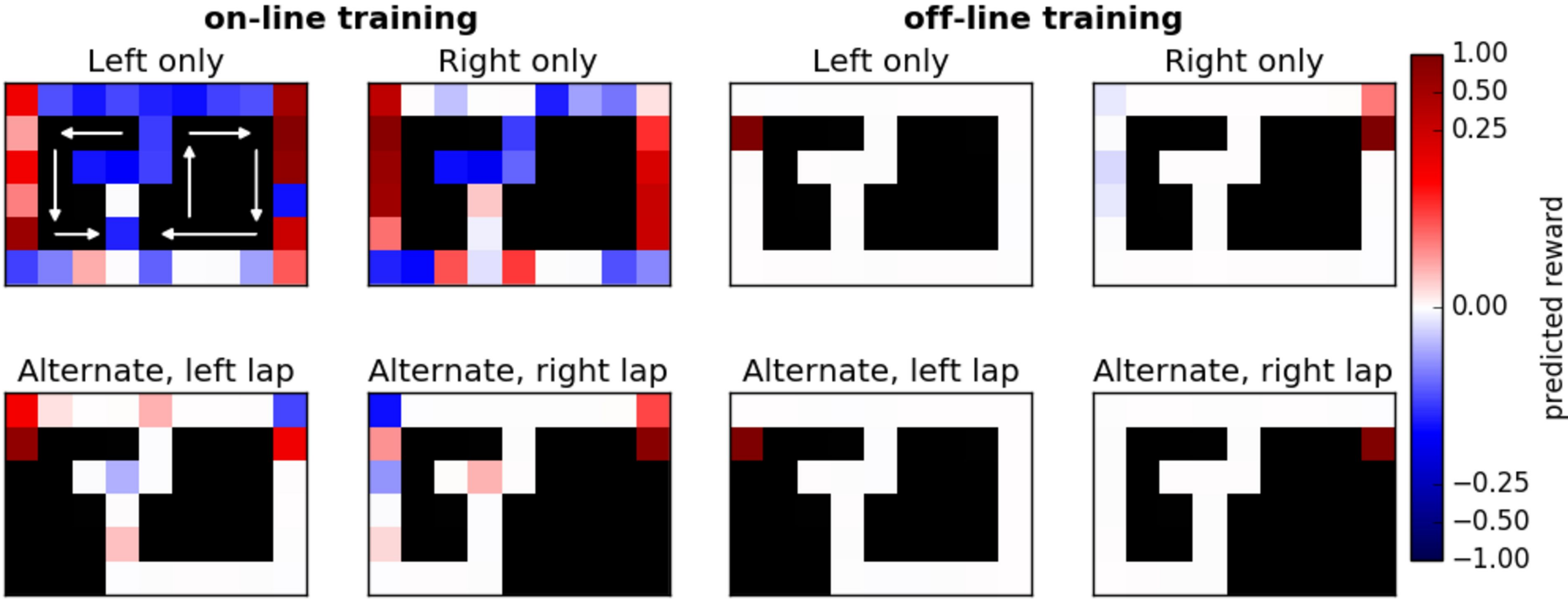
Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464.



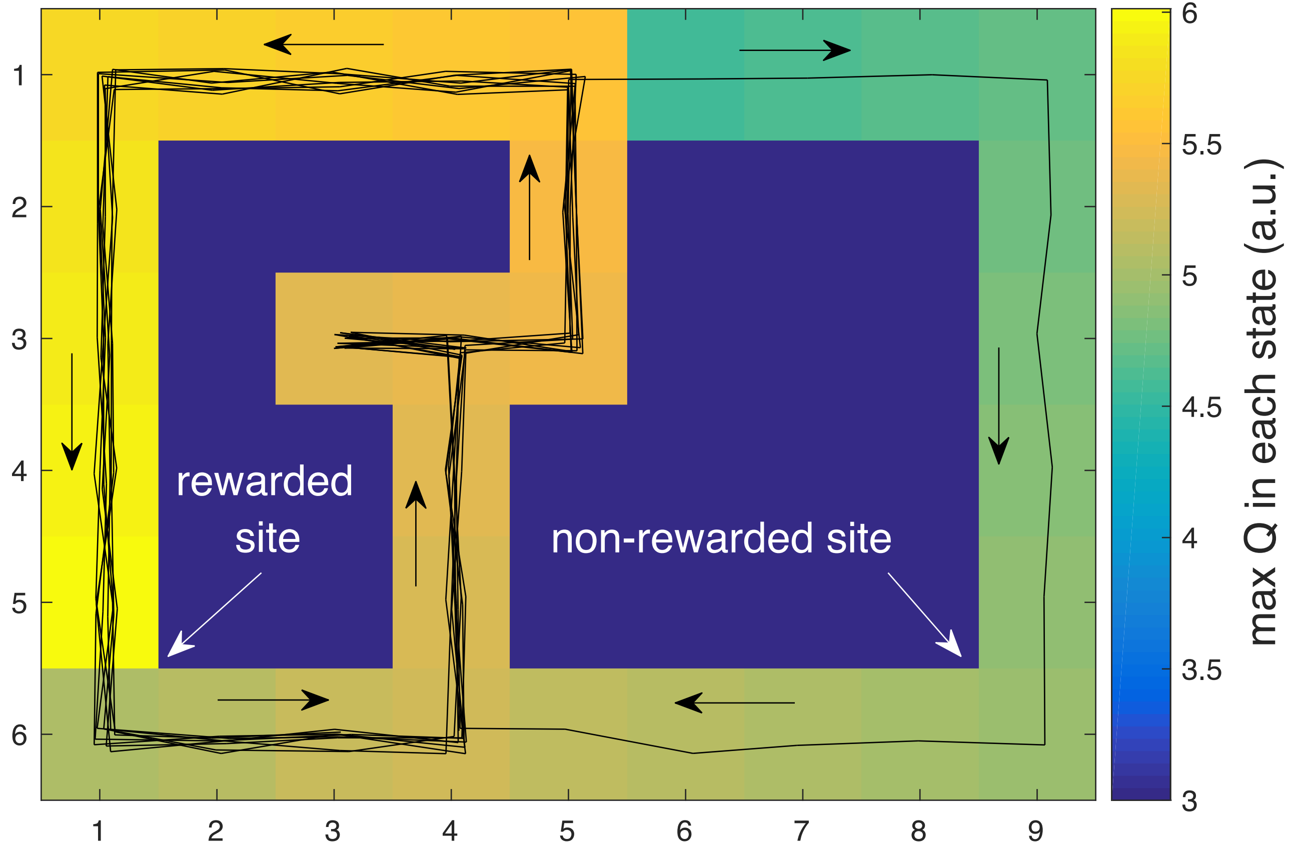
Q		a			
		N	S	E	W
o	…	…	…	…	…
	2	0.66	0.73	0.66	0.73
	3	0.59	0.66	0.59	0.66
	4	0.81	1	0.81	0.81
	…	…	…	…	…

$Q(3,W) = 0.66$

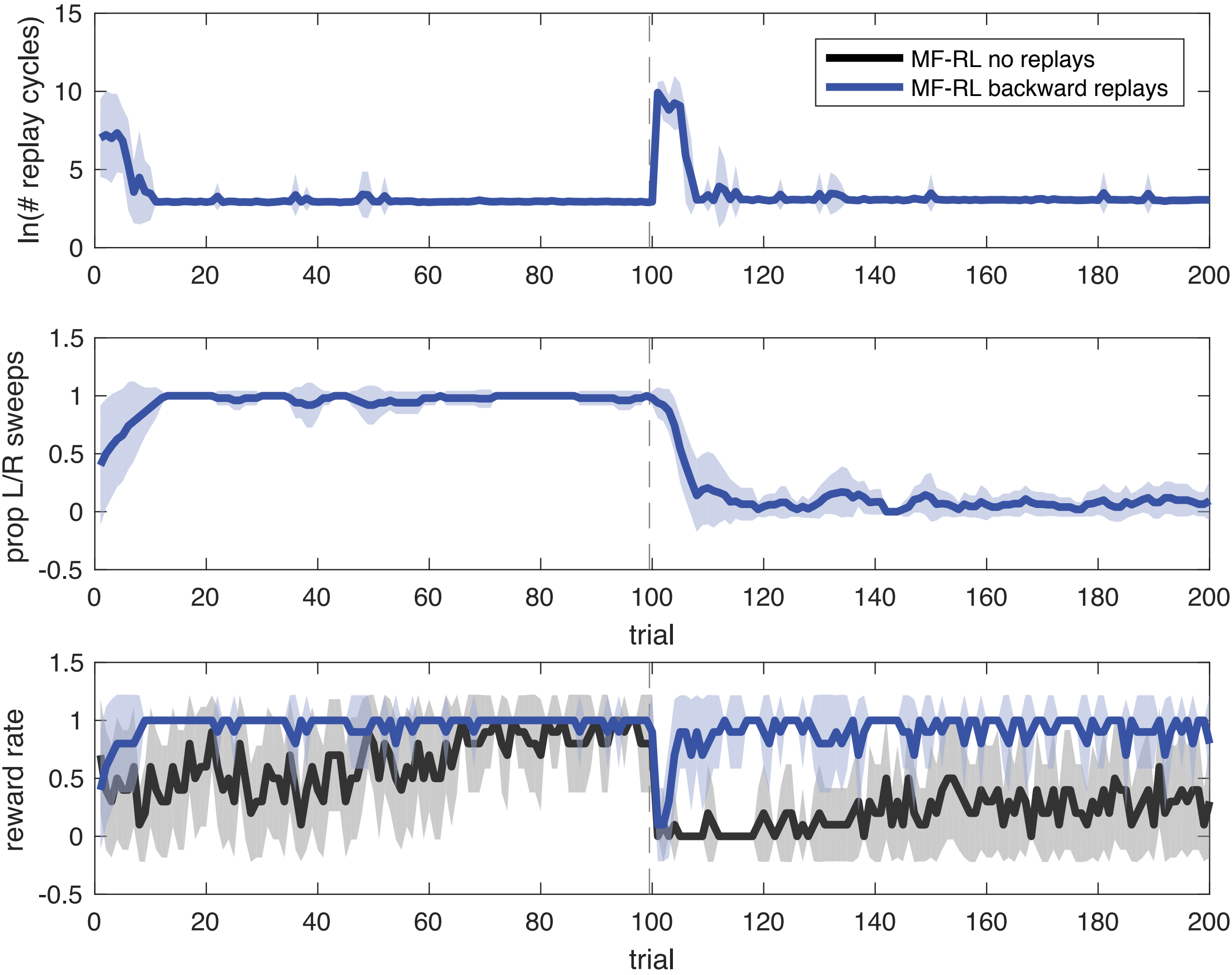
$Q_W(o) = 0.66$

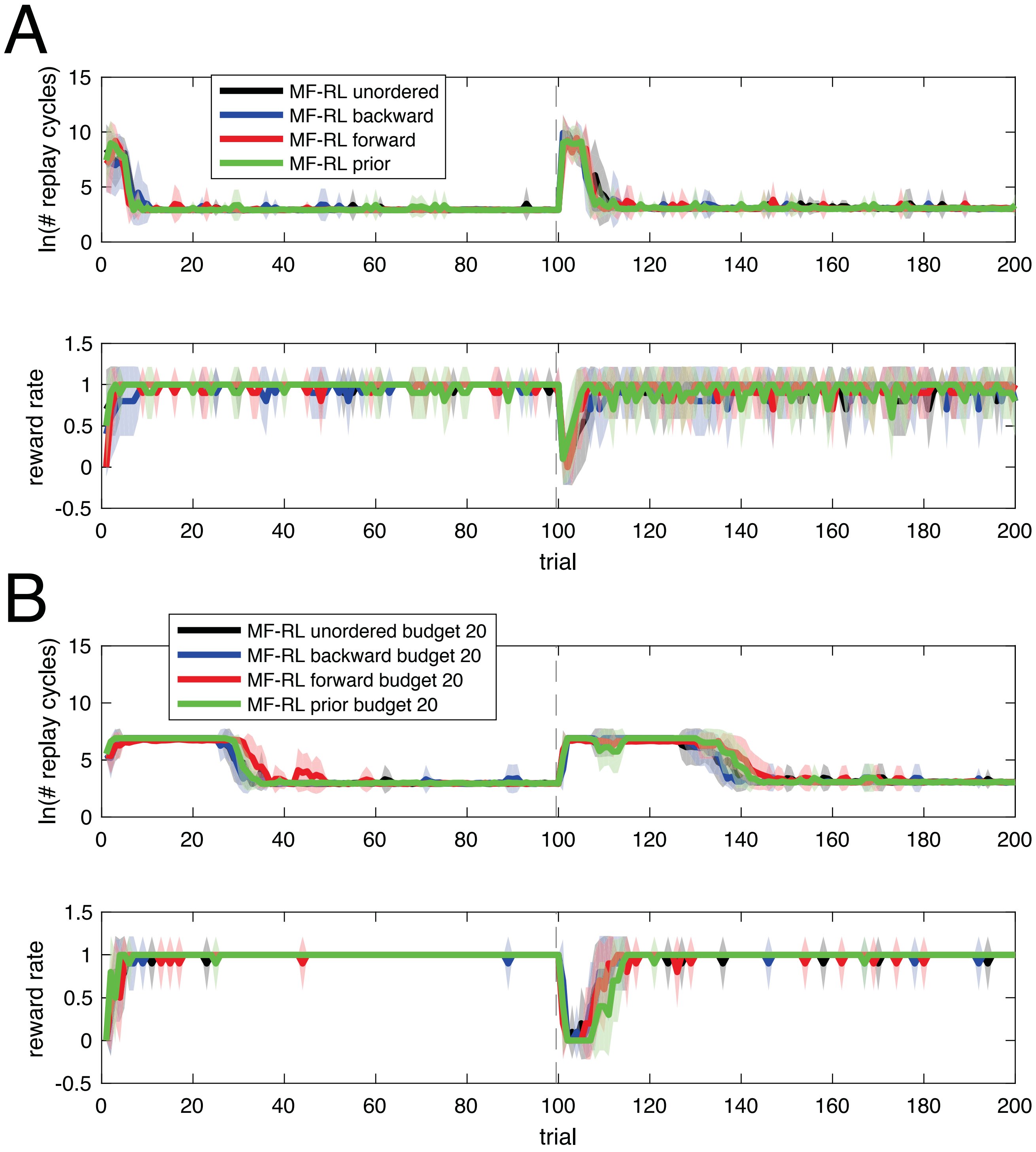

on-line training
off-line training
Left only
Right only
Left only
Right only
Alternate, left lap
Alternate, right lap
Alternate, left lap
Alternate, right lap
1.00
0.50
0.25
0.00
−0.25
−0.50
−1.00
predicted reward

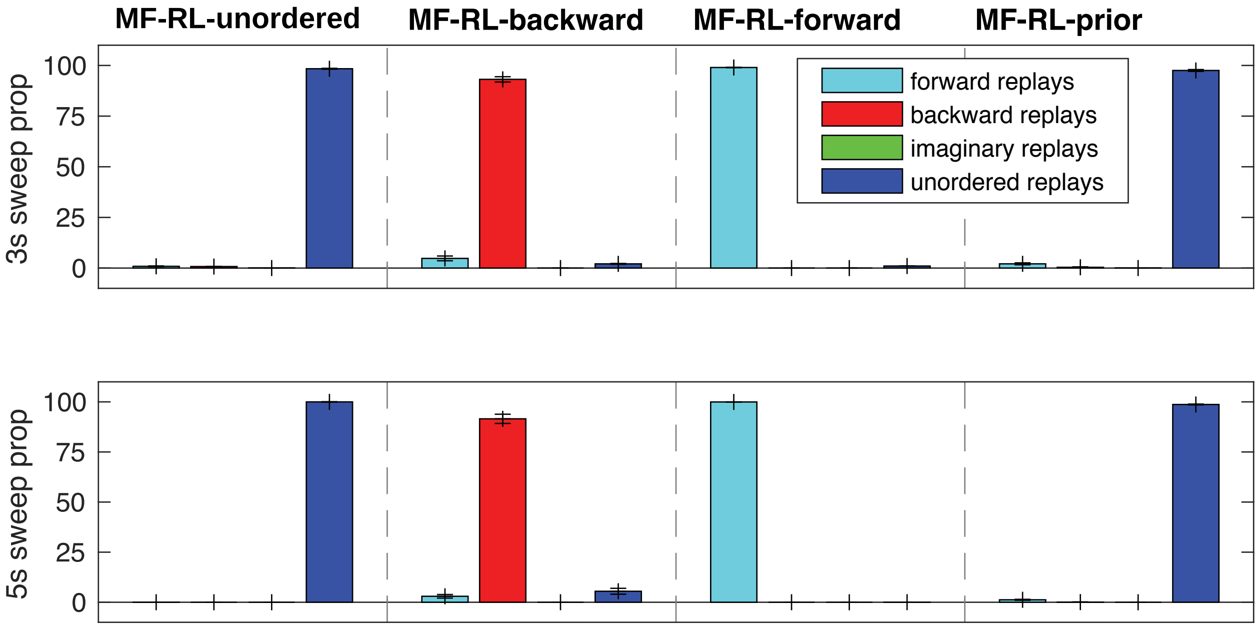
Left only

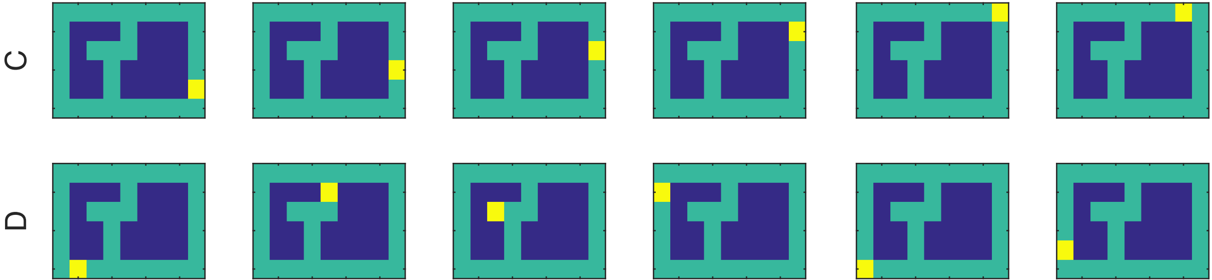


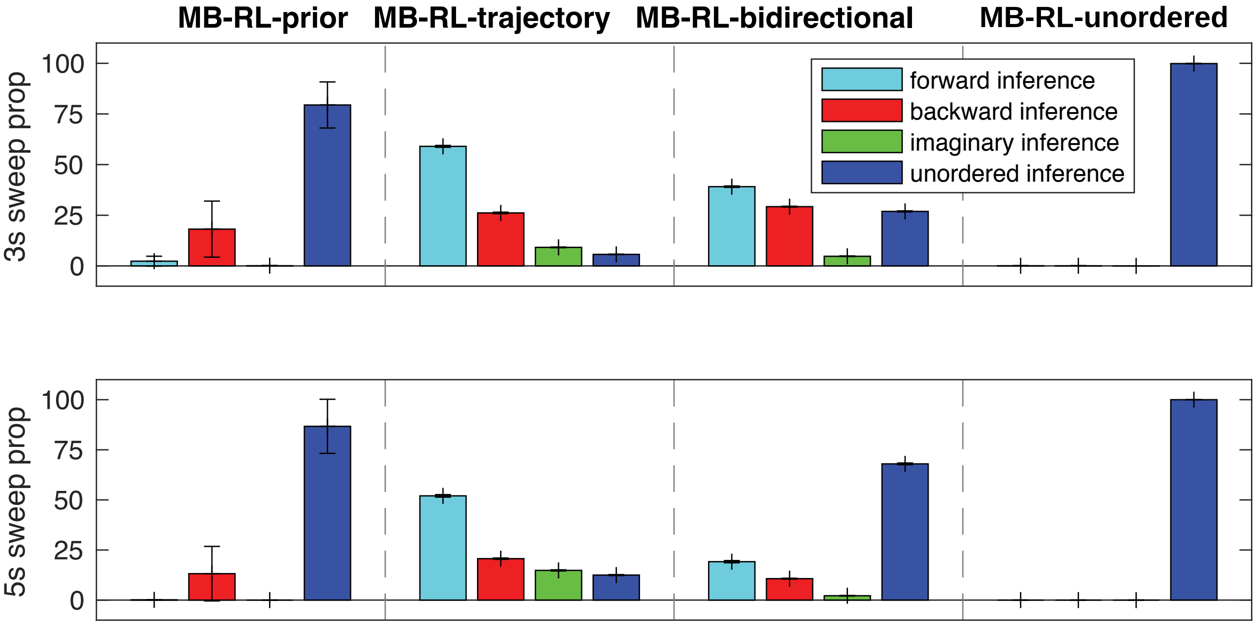
rewarded
site
non-rewarded site
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
6
5.5
5
4.5
4
3.5
3
max Q in each state (a.u.)

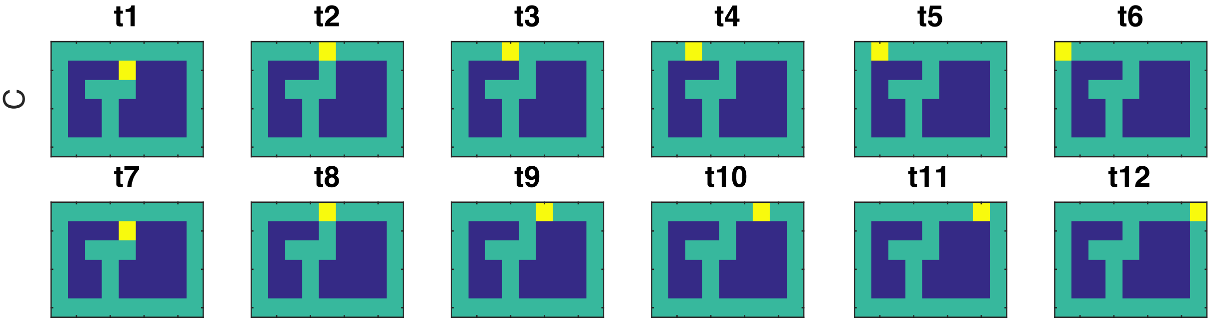

MF-RL no replays
MF-RL backward replays
ln(# replay cycles)
15
10
5
0
prop L/R sweeps
1.5
1
0.5
0
-0.5
reward rate
1.5
1
0.5
0
-0.5
0
20
40
60
80
100
120
140
160
180
200
trial
trial

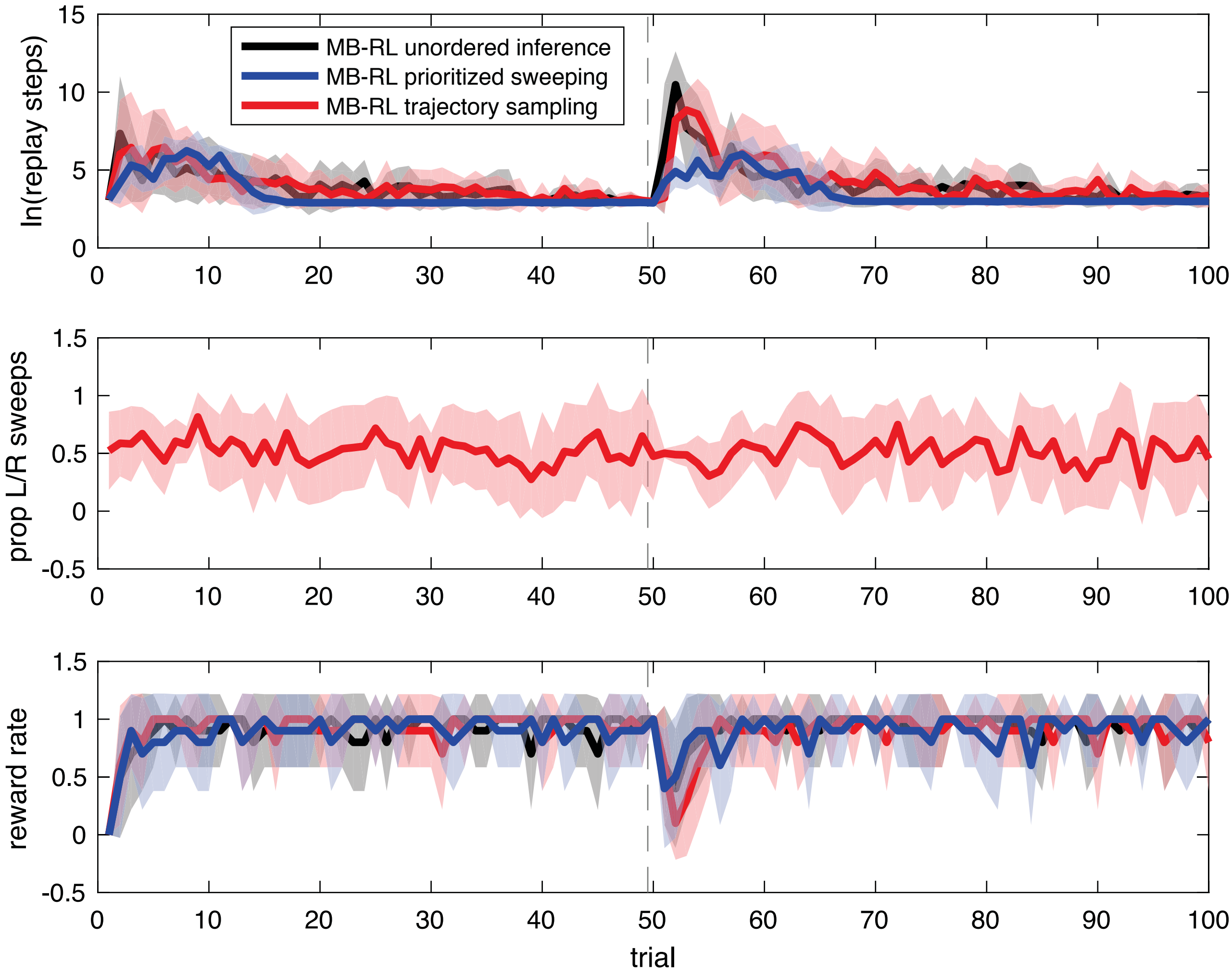

A
ln(# replay cycles)
15
10
5
0
MF-RL unordered
MF-RL backward
MF-RL forward
MF-RL prior
0
20
40
60
80
100
120
140
160
180
200
reward rate
1.5
1
0.5
0
-0.5
0
20
40
60
80
100
120
140
160
180
200
trial
B
ln(# replay cycles)
15
10
5
0
MF-RL unordered budget 20
MF-RL backward budget 20
MF-RL forward budget 20
MF-RL prior budget 20
0
20
40
60
80
100
120
140
160
180
200
reward rate
1.5
1
0.5
0
-0.5
0
20
40
60
80
100
120
140
160
180
200
trial

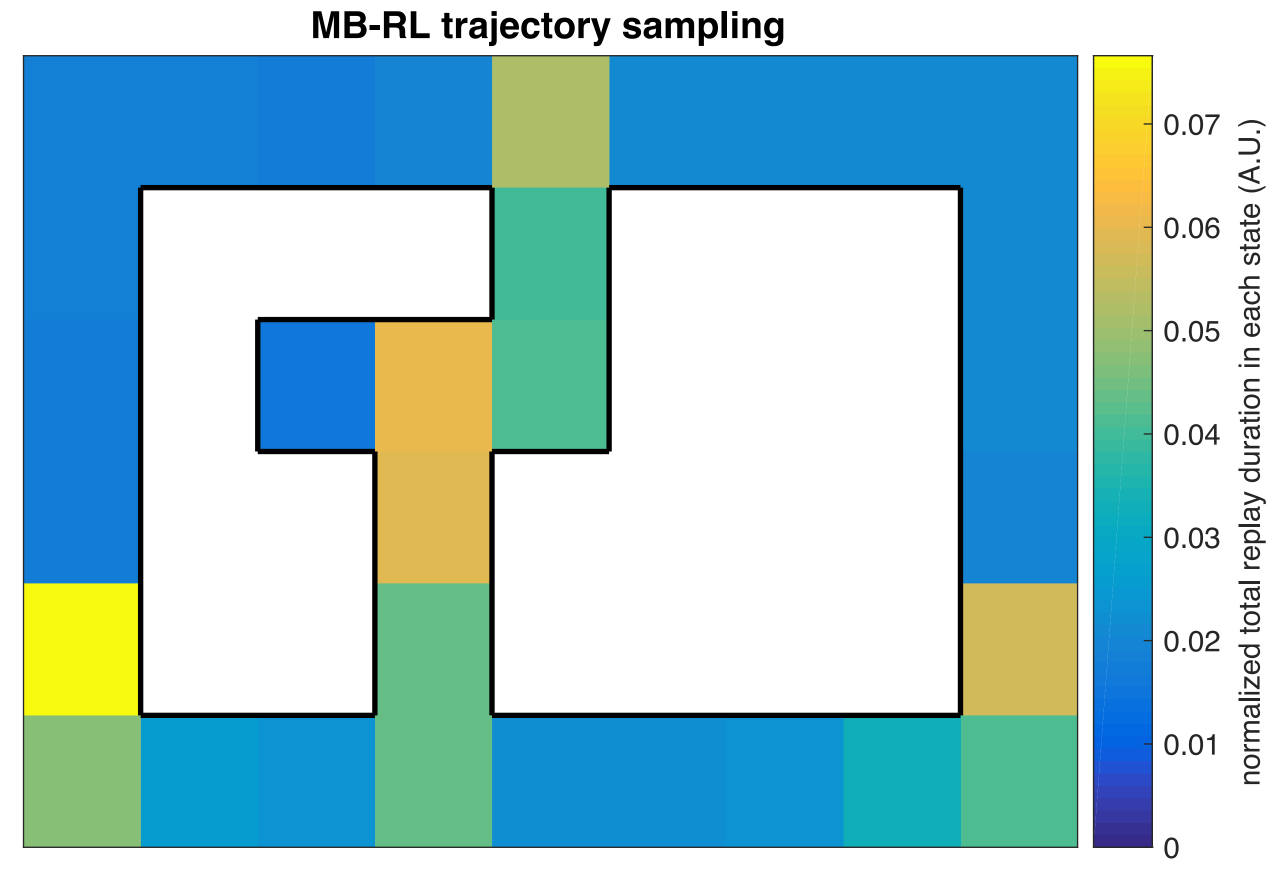

MF-RL-unordered
MF-RL-backward
MF-RL-forward
MF-RL-prior
3s sweep prop
5s sweep prop
100
75
50
25
0
forward replays
backward replays
imaginary replays
unordered replays


C
D

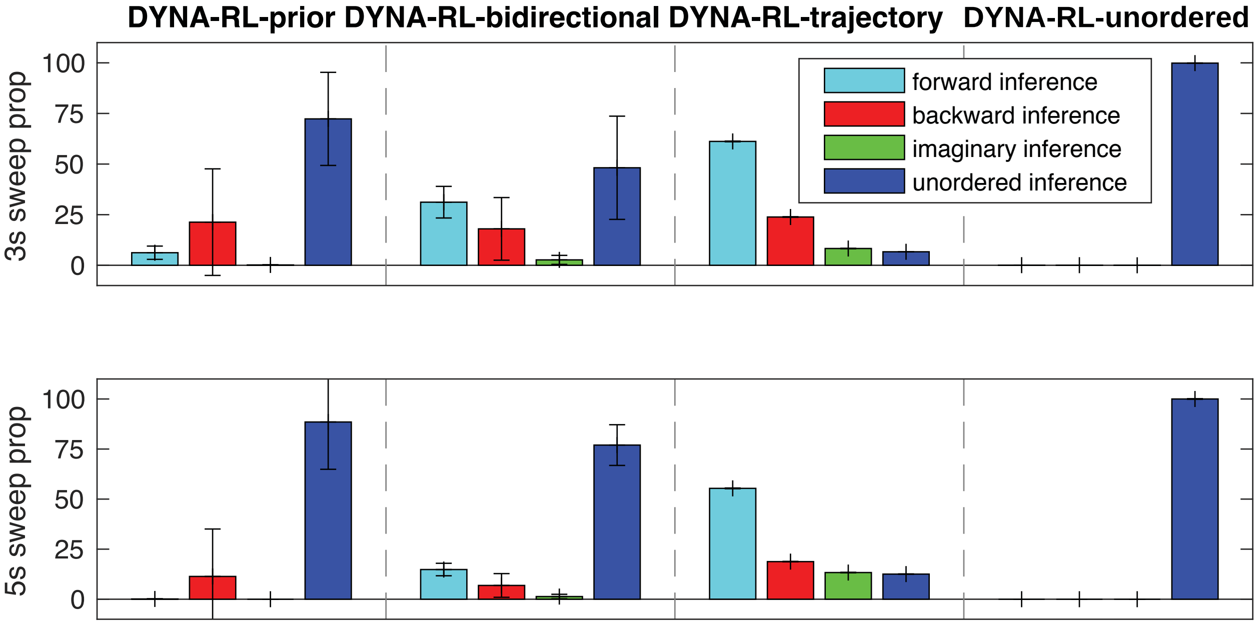

MB-RL-prior
MB-RL-trajectory
MB-RL-bidirectional
MB-RL-unordered
forward inference
backward inference
imaginary inference
unordered inference
3s sweep prop
5s sweep prop
100
75
50
25
0

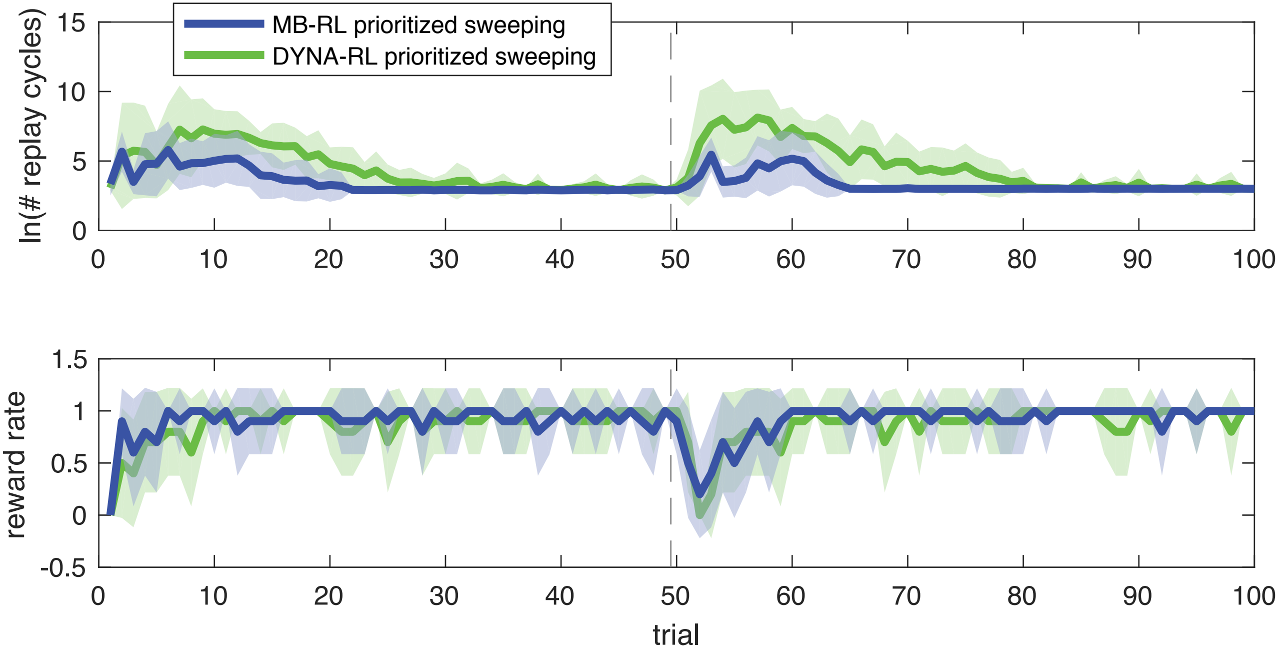

C
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
t11
t12

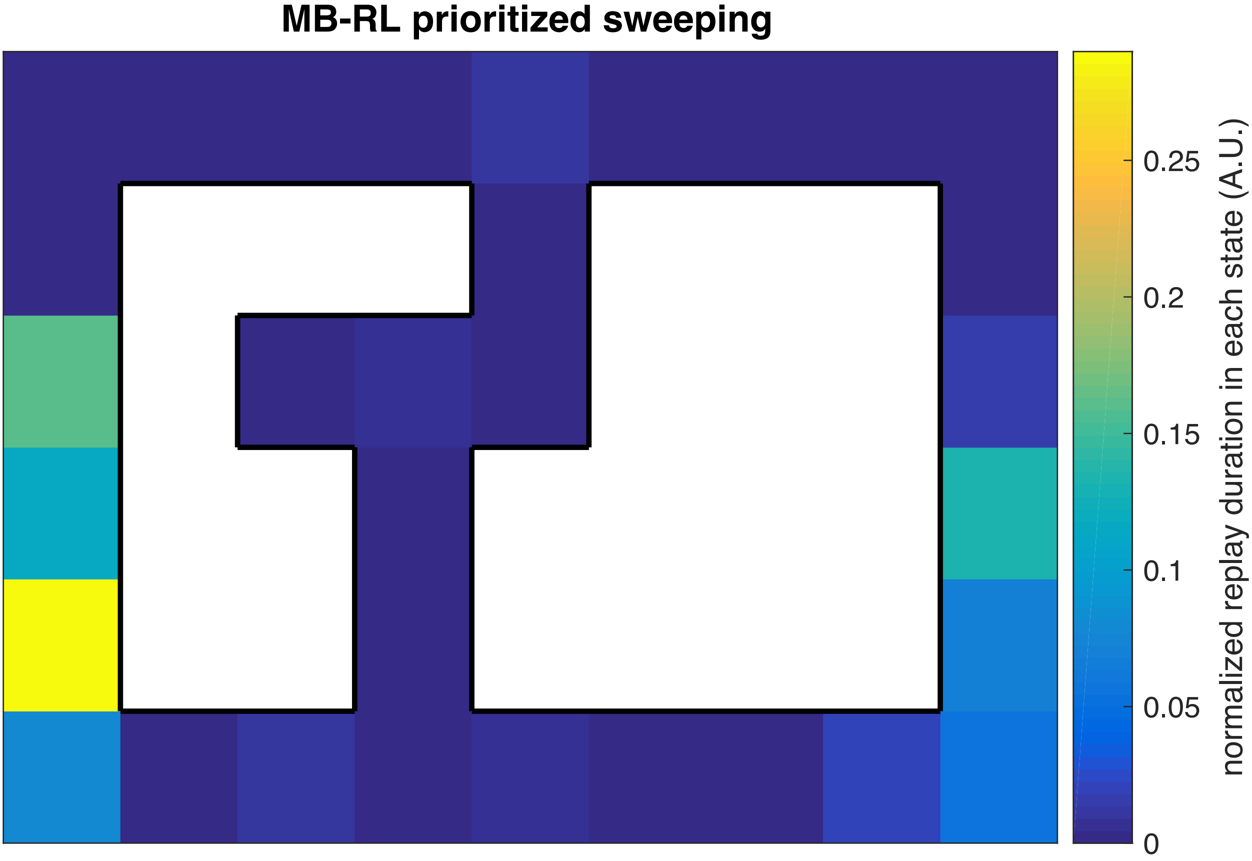

15
10
5
0
ln(replay steps)
MB-RL unordered inference
MB-RL prioritized sweeping
MB-RL trajectory sampling
0
10
20
30
40
50
60
70
80
90
100
1.5
1
0.5
0
-0.5
prop L/R sweeps
0
10
20
30
40
50
60
70
80
90
100
1.5
1
0.5
0
-0.5
reward rate
0
10
20
30
40
50
60
70
80
90
100
trial



MB-RL trajectory sampling
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
normalized total replay duration in each state (A.U.)


DYNA-RL-prior
DYNA-RL-bidirectional
DYNA-RL-trajectory
DYNA-RL-unordered
3s sweep prop
5s sweep prop
100
75
50
25
0
forward inference
backward inference
imaginary inference
unordered inference


MB-RL prioritized sweeping
DYNA-RL prioritized sweeping
ln(# replay cycles)
15
10
5
0
reward rate
1.5
1
0.5
0
-0.5
0
10
20
30
40
50
60
70
80
90
100
trial



MB-RL prioritized sweeping