Replay in minds and machines

Lennart Wittkuhn[1,2,*], Samson Chien[1,2], Sam Hall-McMaster[1,2] & Nicolas W. Schuck[1,2,*]

[1]Max Planck Research Group NeuroCode, Max Planck Institute for Human Development

[2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research

Lentzeallee 94, D-14195 Berlin, Germany

[*]Corresponding authors: wittkuhn@mpib-berlin.mpg.de, schuck@mpib-berlin.mpg.de

Abstract

Experience-related brain activity patterns have been found to reactivate during sleep, wakeful rest, and brief pauses from active behavior. In parallel, machine learning research has found that experience replay can lead to substantial performance improvements in artificial agents. Together, these lines of research have significantly expanded our understanding of the potential computational benefits replay may provide to biological and artificial agents alike. We provide an overview of findings in replay research from neuroscience and machine learning and summarize the computational benefits an agent can gain from replay that cannot be achieved through direct interactions with the world itself. These benefits include faster learning and data efficiency, less forgetting, prioritizing important experiences, as well as improved planning and generalization. In addition to the benefits of replay for improving an agent's decision-making policy, we highlight the less-well studied aspect of replay in representation learning, wherein replay could provide a mechanism to learn the structure and relevant aspects of the environment. Thus, replay might help the agent to build task-appropriate state representations.

Keywords: replay, reinforcement learning, representation learning, decision-making, hippocampus

The power of the mind goes far beyond observing and reacting to our surroundings. Yet, a precise understanding of how the brain can leave the present and enter the realms of memory, planning and imagination is still lacking. Among the most studied aspects is the idea that the hippocampus is central in this process, due to its involvement in retrieving memories when we remember but also when we predict the future (e.g., Squire, 1992; Buckner, 2010). Early indications that the hippocampus gives rise to memory functioning came from studies of lesion patients, such as patient H.M. (Scoville & Milner, 1957), and studies of rodent spatial navigation (O'Keefe & Nadel, 1978). Of particular importance were recordings from hippocampal pyramidal neurons in rodents that demonstrated spatial firing selectivity when the animals navigated in a spatial environment like a maze or linear track (O'Keefe & Dostrovsky, 1971; O'Keefe & Conway, 1978). These spatially modulated neurons – later also found in humans (Ekstrom et al., 2003) – became known as *place cells* and have since been regarded as a core neural substrate for a *cognitive map* of physical space that supports spatial navigation (O'Keefe & Nadel, 1974, 1978; Moser et al., 2008), as well as memory (Cohen & Eichenbaum, 1993; Redish & Touretzky, 1998). How exactly does the brain use this hippocampal machinery to invoke previous events in the service of remembering the past and predicting the future?

1   Replay in the brain

Earlier theoretical and computational work, dating back to Marr (1971), proposed that the brain requires a mechanism mediating the consolidation of recent memory traces into an aggregated memory storage via *reactivation*. The idea that reactivation may help memory formation and should occur particularly when the brain would be undisturbed by external input, e.g., during immobility or sleep, has subsequently found strong theoretical and empirical support (Buzsáki, 1989; Pavlides & Winson, 1989; McClelland et al., 1995). A central development was that technological advances in multi-unit recordings in rodents led to the discovery of *replay*. During periods of rest and sleep, hippocampal cells reactivated sequentially in fast bursts, as if retracing paths the animal had taken during wakefulness (Wilson & McNaughton, 1994; Skaggs & McNaughton, 1996; Kudrimoti et al., 1999; Nádasdy et al., 1999; Gerrard et al., 2001; Lee & Wilson, 2002; Louie & Wilson, 2001; for reviews of these earlier findings, see Redish, 1999; Sutherland & McNaughton, 2000). This

observation was followed by a wealth of findings that established the now classic view of replay: the sequential reactivation of hippocampal place cells that occurs during sleep or rest, reflects previous experience in spatial navigation tasks, and happens on a temporally compressed timescale (for review, see e.g., Foster, 2017).

Over the following decades, much more became known about the biological aspects of replay, and today many studies point to an integral role in memory formation as envisioned by Marr (1971). Replay is commonly detected during brief, high-frequency oscillations of approximately 150 to 250 Hz that last about 50 to 200 milliseconds, called sharp wave-ripples (SWRs) (for review, see e.g., Buzsáki, 2015; Joo & Frank, 2018). SWRs have also been found in human medial temporal lobe (MTL) (Bragin et al., 1999; Staba et al., 2002) and have been linked to memory consolidation during sleep and rest (Axmacher et al., 2008; Staresina et al., 2015; Zhang et al., 2018; Helfrich et al., 2019), as well as awake episodic memory retrieval (Norman et al., 2019; Vaz et al., 2019, 2020). Replay is much faster than wakeful experience, and this temporal compression is believed to induce the conditions that drive learning and the strengthening of memory traces through synaptic plasticity (Bliss & Collingridge, 1993; Magee, 1997; King et al., 1999). A link to memory is also supported by findings showing that the selective disruption of SWRs in rodents during post-task rest slows learning in hippocampus-dependent spatial memory tasks (Girardeau et al., 2009; Ego-Stengel & Wilson, 2010; Jadhav et al., 2012). Finally, interactions between the hippocampus and prefrontal cortex (PFC) during replay events support the idea of consolidating reactivated memories in the brain (for reviews, see e.g., Zielinski et al., 2018; Tang & Jadhav, 2019)

The study of replay in humans has proven to be more difficult compared to rodents because the non-invasive detection of fast and anatomically localized sequential neural events is challenging; but existing findings point to similar conclusions. Memory benefits of non-sequential reactivation during rest or sleep are well documented in humans (Staresina et al., 2013; Deuker et al., 2013; Tambini & Davachi, 2013; Tambini et al., 2010; Gruber et al., 2016). Memory consolidation can also be biased by presenting learning-associated sensory cues during replay-associated sleep phases in humans (Oudiette & Paller, 2013; Lewis & Bendor, 2019), as in rodents (Bendor & Wilson, 2012; Rothschild et al., 2016). Recent progress in neuroimaging analyses have also been able to capture the sequentiality of fast replay events using

magnetoencephalography (MEG) (Kurth-Nelson et al., 2016; Liu et al., 2020) and functional magnetic resonance imaging (fMRI) (Schuck & Niv, 2019; Wittkuhn & Schuck, 2021). In combination, these findings have demonstrated that replay exists in a variety of species and support the idea that it reflects a consolidation process that strengthens memory associations (for reviews, see Sutherland & McNaughton, 2000; Rasch & Born, 2007; O'Neill et al., 2010; Diekelmann & Born, 2010; Carr et al., 2011; Zhang et al., 2017; Tambini & Davachi, 2019).

While the above findings have established a foundational knowledge of replay, our understanding of this phenomenon has undergone significant and continued change that sometimes challenged the classic picture of replay (for review, see e.g., Foster, 2017). For instance, replay has been found to be significantly more frequent than initially thought, happening not only during sleep or rest but also during brief wakeful pauses from active behavior (Csicsvari et al., 2007; Foster & Wilson, 2006; Diba & Buzsáki, 2007; Davidson et al., 2009; Karlsson & Frank, 2009; for reviews of awake replay, see e.g., Carr et al., 2011; Tambini & Davachi, 2019). In addition, replay-like sequential reactivation patterns occur at various speeds, from highly accelerated to much slower behavioral timescales (Deng et al., 2020; Denovellis et al., 2020; Tang et al., 2021). Recordings outside of the hippocampus have also made it possible to identify replay-like phenomena in a number of other brain areas, including entorhinal (Ólafsdóttir et al., 2016, 2017; O'Neill et al., 2017; Trettel et al., 2019), prefrontal (Euston et al., 2007; Peyrache et al., 2009; Jadhav et al., 2016; Yu et al., 2018; Shin et al., 2019; Kaefer et al., 2020; Tang et al., 2021), visual and auditory sensory cortices (Ji & Wilson, 2006; Rothschild et al., 2016; Wittkuhn & Schuck, 2021), parietal cortex (Qin et al., 1997; Hoffman & McNaughton, 2002; Harvey et al., 2012), motor cortex (Ramanathan et al., 2015; Gulati et al., 2017), and ventral striatum (Lansink et al., 2009, 2008; Pennartz, 2004; Gomperts et al., 2015). Moreover, replay is not necessarily a faithful replication of previous behavioral sequences but can also reverse the order of experiences (Csicsvari et al., 2007; Foster & Wilson, 2006; Diba & Buzsáki, 2007; Davidson et al., 2009; Karlsson & Frank, 2009) or change the order of actual experiences according to a learned task rule (Liu et al., 2019). It can also represent remote, non-local and never-experienced locations (Karlsson & Frank, 2009; Gupta et al., 2010; Ólafsdóttir et al., 2015), reflect non-spatial and partially observable task features (Schuck & Niv, 2019), and occur even after tasks without

explicit memory requirements (Wittkuhn & Schuck, 2021). Collectively, these results suggest that replay is involved in a much broader range of cognitive functions than memory consolidation and spatial navigation alone.

Indeed, our understanding of hippocampal place cells, and the neural architecture underlying memory and spatial navigation more generally, has also evolved considerably. The "places" represented by hippocampal neurons are not exclusively determined by location in physical space, but can also incorporate other task-relevant aspects, such as sounds (Aronov et al., 2017) and time (MacDonald et al., 2011; see also O'Keefe & Krupic, 2021). Next to place cells, many other cells have been described that contribute to the cognitive map (for review, see e.g., Grieves & Jeffery, 2017), for instance grid cells and object vector cells in the medial entorhinal cortex (Hafting et al., 2005; Fyhn et al., 2007; Høydal et al., 2019), and head direction cells in the subiculum (Ranck, 1984; Taube et al., 1990). Other studies have pointed out that the hippocampus may encode representations that are predictive of future locations, so called successor representations (SRs), that can be used for reinforcement learning (RL) (Stachenfeld et al., 2017), and that grid-like patterns in the entorhinal cortex and ventromedial PFC may represent coordinates of a non-spatial space (Constantinescu et al., 2016). Thus, today the cognitive map in the hippocampal-entorhinal system is often thought to represent relationships of locations and events beyond physical space, from conceptual knowledge to social cognition (for reviews and perspectives, see Kaplan et al., 2017; Epstein et al., 2017; Schafer & Schiller, 2018; Behrens et al., 2018; Bellmund et al., 2018; Peer et al., 2020; Bottini & Doeller, 2020; Spiers, 2020). Map-like representations also exist beyond the MTL, most notably in the medial prefrontal and orbitofrontal cortex (OFC) (Wilson et al., 2014; Schuck et al., 2016; Constantinescu et al., 2016, for a review, see e.g., Schuck et al., 2018), suggesting that relational information derived from past events is also stored in neocortical brain areas. As we will detail below, these findings may have important implications for our understanding of the nature of replayed representations.

How can such a diverse set of findings about replay in the hippocampus and the rest of the brain be integrated? We argue that insights into this question can be gained by considering the machine learning (ML) literature, where "experience replay" was introduced in the early 90s (Lin, 1991). More recently, experience replay has become particularly popular after its importance for training deep neural net-

works (DNNs) to play Atari video games became clear (Mnih et al., 2015). This led experience replay to rise to prominence as a crucial ingredient in building human-level intelligence in artificial agents (Kumaran et al., 2016). Despite the conceptual similarity of biological and artificial replay, research on this subject in neuroscience and ML has progressed largely in parallel. Here, we aim to connect insights from both research fields and review computational perspectives taken in ML on the replay phenomenon. Our goal is to highlight the diversity of possible functions that might be served by a replay mechanism and attempt to answer the question of *why* agents would replay in the first place.

2 Computational benefits of replay

If replay is not just a phenomenon observed in the brain, but is increasingly implemented in artificial agents too, this raises a number of questions: Is replay in machines similar to replay in animals? Why do machines need replay? Do animals and machines have the same reasons to employ this process? Here, we will consider these questions, compare the roles of replay in both biological and artificial agents, and distill the most significant benefits of replay.

Before we begin, we would like to point out some of the more significant aspects in which the concept of replay differs between ML and neuroscience. As described above, neuroscience emphasizes the sequential and often accelerated nature of replay (Genzel et al., 2020), but often subsumes reactivation of memories as well as activation of internally generated neural sequences under the same term. ML methods, in contrast, often distinguish between experience replay, which corresponds to sampling from a memory buffer, and model-based methods, often referred to as planning, in which the agent internally generates new experiences from a learned model of the environment. Moreover, several methods in ML focus on replaying sets of individual transitions (akin to only two adjacent place cells), which do not have to form sequences spanning long trajectories through the environment. Finally, the issue of replay speed has not been a major consideration in ML. Bound to physical interaction with the environment and the timescales of biology, however, the speed of replay takes a more important role in understanding animal cognition than it has in artificial systems.

Similar to previous work (e.g., Foster & Knierim, 2012; Cazé et al., 2018; Mo-

mennejad, 2020), our review is guided by the RL framework (Sutton & Barto, 2018), which allows parallels to be drawn between reactivation of neural patterns in biological agents and replay of task states in artificial agents. The RL framework considers agents that learn from interactions with their environment and thereby gather experiences one at a time. RL techniques are designed to learn from experience gradually, through trial-and-error, using every new experience immediately to adjust the agent's knowledge about the task. This has the benefit of accruing knowledge without delay, while integrating information over all experiences gained so far, rather than using just the most recent experience to make decisions. Typically, small adjustments are made to the agent's knowledge with each new experience because large updates risk overwriting the effects of earlier learning and can limit generalization. Box 1 describes the fundamental aspects of RL.

Which benefits can an agent obtain from using replay? In the next sections, we will discuss five potential computational functions of replay: increasing the speed and data efficiency of learning, reducing forgetting, reorganizing experiences, planning, and generalization. We do not consider these functions to be entirely separable. We distinguish them because they each offer a unique perspective on what an intelligent agent, biological or artificial, stands to gain from replaying past experiences. This perspective also sheds light on why replay can have different properties in different study contexts, which have found replay to be sometimes backward and sometimes forward, or in some cases to occur immediately and in other cases long after the experience was acquired. In doing so, we try to emphasize how the research fields of neuroscience and ML can learn from each other, identifying commonalities that point to overarching computational principles and delineating differences that could cross-fertilize novel ideas in both fields (for similar perspectives, see e.g., Hassabis et al., 2017; Botvinick et al., 2020).

Finally, we will consider an underexplored computational aspect of replay. Most research to date has studied how replay may be used to optimize goal-directed behavior *given* an already established cognitive map. Here, we explore whether replay may also be involved in shaping the cognitive map itself. We hypothesize that the content and function of replay is determined by its interplay with the current representation of the task and the representational demands of the task at hand, a notion which has recently received some computational (Russek et al., 2017; Caselles-Dupré et al.,

2019; Momennejad, 2020) as well as empirical support (Schuck & Niv, 2019). In this view, replay can be understood not only as a phenomenon that retrieves relational information stored in a cognitive map, but also as a process that changes relational information and internal state representations of an agent (see Box 1 for a definition of state representations).

Box 1   What is reinforcement learning?

Reinforcement learning (RL) theory provides a formal framework to describe how agents learn to optimize their behavior through interactions with the environment that yield rewards or punishments and can result in transitions from one situation (or "state") to the next (Sutton & Barto, 2018). Agents learn which actions to take in which state based on the long-term reward they get. Actions that lead to a net positive sum of future outcomes will be performed more often, whereas actions that lead to a net negative sum of outcomes will be performed less often. At the heart of RL lies the assumption that the agent-environment interaction can be modelled as a Markov decision process (MDP). The MDP consists of the following components:

- An **environment** that is represented by a discrete set of states S (e.g., locations in a maze or positions in a board game). The agent is assumed to occupy one of those states at each point in time. Note that the states refer to the agent's internal representation of the environment, which may not be equivalent to the objective, "true" state of the environment. In some environments it can be sufficient to form an internal state representation based on incoming sensory signals alone, but in many cases this approach is not feasible and instead requires to supplement the state definition with non-observable information (e.g. Wilson et al., 2014; Schuck et al., 2018). The question of how an agent can know how to internally represent the state is therefore a topic of great interest in RL research, and will be discussed later in this review. Importantly, by describing the environment as a state space of an MDP, the states can represent any type of task, whether spatial locations in a maze or task states indicated by objects in a non-spatial decision-making task. Note that MDPs can also have continuous state or action spaces, but for simplicity, we only consider finite and discrete state and action spaces here.

- A discrete set of **actions** A the agent can perform in each state (e.g., movement into a particular direction in a maze).

- A **state transition function** $T(s_t, a_t, s_{t+1})$ describing the probabilities of moving from the current state $s_t \in S$ to the next state $s_{t+1} \in S$ after taking an action $a_t \in A$. The state transition function can be understood as a cognitive map that entails the connections between states (e.g., route from location A to location B).

- A **reward function** $R(s_t, a_t, s_{t+1})$ maps each state-action pair to a scalar reinforcement signal r, received after taking an action a_t in state s_t. Generally, rewards are controlled by the environment, and therefore lie outside of the agent. Still, the value of particular outcomes are inferred by the agent, and learning therefore depends on an internal representation of the reward (e.g., Juechems & Summerfield, 2019). This internal representation can be subject to further transformations such as reward normalization, which in turn may depend on the agent's state.

In an MDP, the agent-environment interaction is assumed to be Markovian with respect to reward (and state), which means that the state and reward at the next time point $t + 1$ depend only on the state and action of the current time point t, but not on any states or actions before. The current state s_t therefore contains all information from the previous history. Formally, the Markov property can be stated as:

$$P(s_{t+1}, r_t | s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1}, r_t | s_t, a_t) \tag{1}$$

In brief, this means that we can think of learning from trial-and-error as the following process: the agent represents the current state of the environment, s_t and then performs an action a_t. The action will affect the environment, changing the agent's state from s_t to s_{t+1} (as described by the transition function T), and potentially yield a reward, as described by the reward function R. The agent's goal is to always perform the actions that maximize the sum of total rewards obtained over the course of its interaction with the environment.

How can the agent know which actions to perform that will achieve this goal? A first step is to formalize the agent's goal. The goal of the agent is expressed as the expectation over the discounted sum of its future rewards: $E[\sum_{i=0}^{\infty} \gamma^i r_i]$, where i indexes all future time points, and $\gamma \in [0, 1]$ is known as the discount factor which

attenuates the influence of distal rewards and ensures that the sum of rewards has a finite value. In RL, the agent is assumed to gather knowledge about this discounted sum of rewards and store it as a *value function*. The agents should therefore hope to eventually have a very accurate value function, which reflects the true sum of discounted rewards. Future rewards will depend on the agent's behavior which is reflected in the agents' value function, defined as the expected reward after taking action a_t from state s_t, then following policy π thereafter:

$$Q^\pi(s_t, a_t) = E_\pi \left[\sum_{i=0}^{\infty} \gamma^i r_i | s_0 = s_t, a_0 = a_t \right] \qquad (2)$$

which can be written recursively as:

$$Q_\pi(s_t, a_t) = \sum_{s_{t+1}} T(s_t, a_t, s_{t+1})[R(s_t, a_t, s_{t+1}) + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))] \qquad (3)$$

Equation 3 is known as the Bellman equation, and its solution is obtained through an iterative bootstrapping process by first estimating the Q-values, then repeatedly updating the Q-values using the Bellman equation until they converge to the true values. This approach requires knowledge of a model of the environment (i.e., state transition function), and such methods are known as model-based RL (Doya, 1999; Daw et al., 2005). However, in many situations the agent does not have such a model, thus a separate category of algorithms known as model-free RL is used. These model-free methods utilize a combination of sampling and bootstrapping to compute the value functions directly without the MDP model. The best-known example is the temporal difference (TD) learning algorithm, which adjusts the value function incrementally using a prediction error (PE) capturing the discrepancy between the actual reward and the estimated state value. A widely-used variant of TD learning is the Q-learning algorithm:

$$Q_{k+1}(s_t, a_t) = Q_k(s_t, a_t) + \alpha[r_t + \gamma \max_{a_{t+1}} Q_k(s_{t+1}, a_{t+1}) - Q_k(s_t, a_t)] \qquad (4)$$

The Q-learning algorithm assumes that the agent picks the best action on the next step based on its current knowledge. The resulting Q-value, i.e., $\max_{a_{t+1}} Q_k(s_{t+1}, a_{t+1})$, is then used as the target to update the current Q-value in iteration k, which circumvents the necessity to maintain a transition function as in model-based approaches. The learning rate $\alpha \in [0, 1]$ determines the extent of adjustment in each update.

2.1 Faster learning and data efficiency

The gradual approach to learning in RL has many benefits, but it results in very slow learning that may need thousands or even millions of iterations to achieve the optimal policy. Even worse, the slowness of learning grows exponentially with the number of states in the task environment, a phenomenon known as the "curse of dimensionality" (Bellman, 1957). To be a feasible approach to learning in complex and changing environments, gradual methods therefore must be complemented by mechanisms that will speed up learning without sacrificing the benefits of immediate knowledge acquisition and stable long-term memory. In this light, the idea of recapitulating previous experiences seems particularly appealing for machines, because it is easy and cheap for artificial agents to relearn from past experience stored in a memory buffer.

Our brain arguably faces a similar challenge. We, and other animals, often have to learn directly from the outcomes of our decisions. Yet, repeating errors can pose actual risks to us, which limits the usefulness of exclusively relying on a slow, trial-and-error-based learning mechanism. More generally, the number of experiences we acquire with a particular situation in a lifetime is quite limited in relation to the complexity of our environments and our brain, which contains approximately 10^{14} synapses (Tang et al., 2001). In order to make thousands of gradual adjustments to each of these synapses, the ability to reuse experience efficiently is paramount.

In the RL literature, "experience replay" was initially introduced to address these issues of slow learning and data inefficiency (Lin, 1991, 1992, 1993). In his seminal paper, Lin (1992) wrote that "[...] Q-learning algorithms [...] are inefficient in that experiences obtained by trial-and-error are utilized to adjust the networks only once and then thrown away. [...] Experiences should be reused in an effective way." (Lin, 1992, p. 299). Lin (1992) proposed that experiences can be used to update knowledge in a dual fashion; (1) immediately when experiences are acquired, and (2) at later time points, after experience itself may have long passed. Specifically, Lin (1992) proposed replaying full sequences of experiences, starting from an initial state to a final state in backward order, and learning from these experiences, as if they were real. Lin (1992) then showed that this is a more efficient use of data that accelerates learning of an RL agent. In line with these ideas, many others have since emphasized the computational benefit of replay for maximizing data efficiency and the speed of

learning (for reviews, see e.g., Kumaran et al., 2016; Hassabis et al., 2017). Studies in rodents reported increases in SWR-associated reactivation following initial learning in novel environments (Cheng & Frank, 2008; Eschenko et al., 2008; O'Neill et al., 2008; van de Ven et al., 2016; Tang et al., 2017), which suggests an important role of replay in speeding up learning in biological agents too. Several studies reported that it requires only a few experiences in a novel environment for replay to occur, and that it can be detected already during the awake state immediately after behavior (Foster & Wilson, 2006), but see Jackson et al. (2006). Disrupting replay-related SWRs during awake rest in rodents slows learning in a spatial navigation task (Jadhav et al., 2012).

There are several reasons why replay can help learning. The core idea of RL is that an agent learns which actions to take through positive and negative outcomes. In the real world, outcomes are often only obtained after a long sequence of events and actions but agents still need to know how to behave at the start of the sequence, as for instance, in a chess game. This problem is known in RL as the *temporal credit assignment problem* (Minsky, 1961) and replay may help to solve it. The early work by Lin (1991, 1992, 1993) pointed out that replay could help an agent to remember the sequence of previous states and actions that led to a given outcome, and assign credit for the reward to the sequence of states and actions that preceded it. This also explains why sequential replay may sometimes proceed in backward order (Lin, 1992). In the brain, awake backward replay has indeed been frequently observed, where rewarded spatial trajectories of an animal are replayed in reverse order (Diba & Buzsáki, 2007; Foster & Wilson, 2006; Singer & Frank, 2009), and the frequency of awake backward (but not forward) replay is modulated by the change in reward magnitude (Ambrose et al., 2016). Another aspect is that as the agent's knowledge of the rewards becomes better with time, outcomes in the past should be re-evaluated in light of this updated knowledge (van Seijen & Sutton, 2015). By comparing past rewards against current value estimates, replay can serve this function.

In line with the idea that backward replay reflects learning through temporal credit assignment, the rate of backward replay was also observed to be more frequent in novel compared to familiar environments (Foster & Wilson, 2006; Singer & Frank, 2009) and to decrease its bias to reflect previous paths to the goal location as a function of learning (Shin et al., 2019), suggesting that the relevant trajectory has

been learned and does not need to be reinforced through replay anymore (Foster & Knierim, 2012). Computational work shows how backward replay can strengthen forward synaptic pathways through spike timing dependent plasticity (STDP) (Haga & Fukai, 2018) and thus support forward replay during sleep and active behavior (Johnson & Redish, 2007; Pfeiffer & Foster, 2013; Wikenheiser & Redish, 2015b, 2013). Further evidence for the role of replay in assigning credit are findings showing that replay is coordinated with subcortical activation of brain areas related to processing reward (Gomperts et al., 2015; Lansink et al., 2009; Pennartz, 2004; Gomperts et al., 2015), which could convey reward signals to other brain regions like the hippocampus. So far, existing empirical studies support the idea that awake backward replay supports temporal credit assignment by retrieving states that led to the outcome, accelerating learning for cases in which a long delay between rewards and actions must be encoded.

Figure 1 provides an illustration of how backward replay of full sequences works in the context of RL, and which effect it has on the speed of learning. We consider an RL agent navigating in a square environment with 20×20 tiles that contains several walls and one goal location with a reward (see Figure 1A). The agent can move into one of the four cardinal directions (up, down, left, right). It receives no reward for moving, a small negative reward for bumping into a wall (-0.1), and a reward of 1 when arriving at the goal location. The goal of any RL agent is to maximize the sum of future rewards. Thus, in this case the best policy is to navigate to the reward with as little steps as possible, and without bumping into the wall. This is a well-known "grid world" problem that can be solved using RL because any agent using the gradual updating procedures described above will eventually learn to navigate directly to the reward – but without replay this learning might be painfully slow. For illustrative purposes, we use the off-policy, model-free Q-learning algorithm described in Equation 4 in Box 1. The learning rate α, temperature τ and discounting factor γ were arbitrarily set to 0.3, 1 and 0.99 for the purposes of this illustration.

In each episode, the agent starts in a random position and navigates until it has found the reward. After the reward is found, the agent is replaced at a new starting location. The starting locations varied randomly, but in order to avoid excessive fluctuations in performance caused by variability in starting locations, their distances
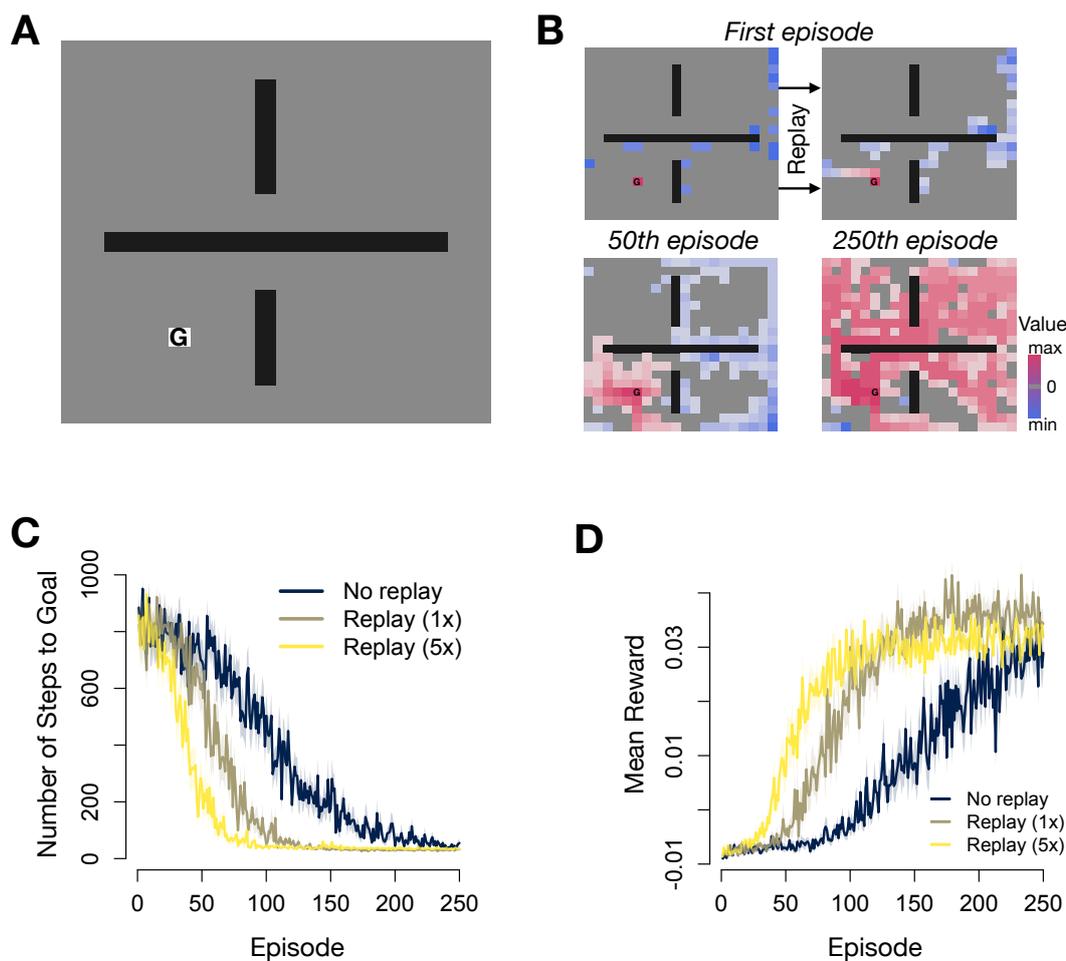
Figure 1: **Replay speeds up learning to navigate to a goal in a grid world.**
(A) Square environment ("grid world") with 20×20 tiles (shown in gray) that
contains several walls (black tiles) and one goal location (white tile labelled "G")
that contains a reward. At the beginning of each episode, an RL agent is placed in
a random location and can move into one of four cardinal directions (up, down, left,
right). The agent receives no reward for moving, a small negative reward for bumping
into a wall (-0.1), and a reward of 1 when arriving at the goal location. An episode
is terminated once the agent reaches the goal location or a maximum of allowed steps
per episode set to 1000. **(B)** Illustration of the learned value function after the first
episode of experience (top left), following replay of the first episode (top right), and
after the 50th and 250th episode (bottom). Colors indicate the values of locations
from smallest (blue) to highest (red). Values are under the best possible policy,
which is assuming that the agent would perform the value-maximizing action in each
location. The increasing prevalence of red tiles after 250 episodes therefore reflects
that after training the agent has learned a policy for most locations that will avoid any
collisions with the wall and reach the goal within the maximum number of allowed
steps. Color mapping is scaled for each plot and values smaller than 0.1 are shown as
gray tiles. **(C)** Number of steps (y-axis) needed by the RL agent in each consecutive
episode (x-axis) to reach the goal location when using no replay (blue line) between
episodes, or when replaying the previous episode in backward order once (brown
line) or five times (yellow line). **(D)** Mean reward (y-axis) achieved by the RL
agent in each consecutive episode (x-axis). Colors as in (c). The computer code for
the simulations is publicly available at https://github.com/nschuck/replaysim
-wittkuhn-etal2021. © Wittkuhn et al., doi: 10.6084/m9.figshare.14261636.v1,
CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

to the reward location were constrained to lie at least 10 tiles away from the goal location (in euclidean distance, i.e., considering direct flight paths). The blue lines in Figure 1C–D show the number of steps the agent needs to navigate to the goal location. As can be seen, the agent needs a considerable amount of experience in order to learn to quickly find the goal location, about 250 episodes in which the agent could explore the states between the start and goal location. Next, we investigate the speed of learning when we supply the agent with a simple replay mechanism. In particular, we used backward replay of entire episodes at the end of each successful episode, similar to the work of Lin (1992), and in line with the considerations about temporal credit assignment discussed above that backward replay can propagate the value back to preceding states. This meant that upon encountering a reward, the agent would not only update the Q-value of the last state. Instead, the agent internally traversed backwards through the sequence of states, actions and rewards until the beginning of the episode $(s_t, a_t, r_t, \ldots, s_1, a_1, r_1)$, updating its Q-value at each step just as if it was experiencing the full trajectory episode again. As can be seen in Figure 1C–D (brown and yellow lines), adding replay dramatically speeds up learning of the agent, reducing the number of interactions needed to achieve ceiling performance to less than half of what was observed without replay. Note that the choice to replay the full sequence of states, actions and rewards between the start location and the goal location is not without consequence, and a variety of different definitions of what constitutes an episode are common in RL and neuroscience. We consider them in Box 2 and will discuss their implications further below.

One notable aspect of our example is that replayed experiences are used to update Q-values, which slowly integrate knowledge over many experiences to control behavior. This is different from relying directly on specific single episodes for behavioral control, an approach often used in the context of *episodic RL* (for recent reviews, see e.g., Gershman & Daw, 2017; Botvinick et al., 2019). During episodic RL, specific single episodes are stored in memory and retrieved to directly determine behavior when the same or a similar situation is encountered again (Lengyel & Dayan, 2007; Gershman & Daw, 2017; Botvinick et al., 2019). Episodic control also has benefits for the speed of learning under particular circumstances. For example, repeating memorized sequences of actions that led to desirable outcomes in the past is a reasonable strategy for an agent that prioritizes exploitation over exploration and has

only limited experience with a complex and noisy environment that it might rarely encounter again in the future (Lengyel & Dayan, 2007). In humans, the retrieval of single experiences in decision-making is associated with the hippocampus (Lee et al., 2015; Bornstein & Norman, 2017; Wimmer & Büchel, 2020) and reinstatement of information from past choice trials at decision-time biases present choices towards decisions made previously in the reinstated context (Duncan & Shohamy, 2016; Bornstein et al., 2017; Bornstein & Norman, 2017). To what extent these retrievals of single episodes are supported by sequential replay remains an open question. While this direct use of experiences can be very efficient, the brain must strike a balance between such efficiency (i.e., taking the same route to a goal location that was used before) and the flexible updating of behavior in cases when the most efficient behavioral strategy cannot be used (i.e., the most direct path to a goal location is blocked). The distinction between reusing experiences for episodic RL or updating Q-values highlights the fact that there are multiple ways to instantiate replay in an RL agent, and the illustration in Figure 1 only serves as a basic introduction to computational replay. Below we will describe the many different aspects that determine the usefulness of replay for different problems in more detail.

Box 2   What is replayed?

Replay is generally thought to represent previous experience. In artificial agents, an experience e_t is commonly defined as a quadruple consisting of the current state s_t, the taken action a_t, the reward r_t received after taking action a_t in state s_t, and the next state s_{t+1}, together $e_t = (s_t, a_t, r_t, s_{t+1})$, effectively describing a single transition between two states as the atomic unit of an artificial replay event. In earlier RL implementations of experience replay (Lin, 1991) and our own demonstration in Figure 1, replay involves entire episodes (i.e., sequences of chained experiences from the start to the goal location). In contrast, deep neural networks (DNNs) use a single experience as a replayed unit to create independent and identically distributed (i.i.d.) samples required for stochastic gradient descent (SGD). For example, in the DNN by Mnih et al. (2015) the states S provided to the Q-function consist of pre-processed versions of Atari pixel frames. In biological agents, replay is commonly defined as *sequential* reactivation of neural activity patterns relating to previously

encoded experience (Genzel et al., 2020). What constitutes a replayed experience is a much more difficult question to answer for biological agents. In non-human animals, this typically involves recordings of hippocampal place cells representing physical locations in a spatial environment, thus equating states to locations and episodes of experiences to trajectories of locations. However, hippocampal cells appear to be quite flexible in encoding task-relevant information other than physical space, for instance sounds (Aronov et al., 2017), trial history (Wood et al., 2000) or run laps (Sun et al., 2020). This notion is even more important in humans, where, depending on the experimental design, replayed memory representations can be multi-modal and thus it is not immediately apparent which elements of an experience should be stored and where in the brain those might be represented. It therefore seems appropriate to consider that replay not only reflects physical location in spatial environments but rather a state of task-relevant information, which, in typical rodent tasks might be majorly, but not exclusively, defined by physical location.

Which and how many experiences are stored? How long are they stored? In DNNs using replay, the newest experiences are stored at each time step in a memory buffer $D_t = e_1, \ldots, e_t$ with finite size N (e.g., Mnih et al., 2013; Mnih et al., 2015; Zhang & Sutton, 2017). Since the success of the DNN by Mnih et al. (2015), the memory buffer is typically set to a size of $N = 10^6$ newest experiences, which continuously replace the oldest experiences (Fedus et al., 2020; Zhang & Sutton, 2017). Recently, Fedus et al. (2020) investigated the relationship between the number and age of experiences stored in the memory buffer. First, they found that increased memory capacity improved learning performance, likely due to a larger coverage of state-action pairs (Fedus et al., 2020). Second, decreasing the age of the oldest experience in the memory buffer also improved performance, likely because of older experiences that resulted from policies that are inconsistent with the current on-policy decision strategy, and is in line with earlier findings noting that experience replay is only beneficial if it is consistent with the current decision policy (Lin, 1991, 1993). An exception to this are Atari games that are characterized by sparse rewards and require high levels of exploration. In such tasks sampling from older off-policy experiences is still beneficial (Fedus et al., 2020). These considerations about the size of the memory buffer in artificial agents point to an intriguing trade-off between old and new memories: On the one hand, a youthful memory buffer storing only recent

experiences can effectively drive the current decision policy and quickly abandon old and potentially inefficient behavior. On the other hand, keeping older experiences and integrating them with recent ones may foster generalization and prevent an agent from becoming stuck in a decision policy that is suboptimal.

While the size and content of a memory buffer in artificial agents can be crafted by ML researchers, determining number and nature of memories in brains is topic of ongoing debate for neuroscientists. Humans and other animals are faced with a continuous stream of experiences and, given the gigantic amount of experiences a biological agent acquires in a lifetime, putting a number on the memory size in brains appears virtually impossible. Even if biological agents had an unlimited memory storage, selecting memories for replay from that storage would become challenging with a large amount of experiences. In fact, forgetting irrelevant experiences might be an adaptive biological process (Hardt et al., 2013) and resonates with the idea that the brain should form compact memory representations that have the highest utility for behavior. Tackling the question of what counts as an event or experience, a common theme in neuroscience has been to posit that the brain segments continuous experience into representations of distinct neural states that transition at boundaries between events or shifts in context (for reviews, see e.g., Richmond & Zacks, 2017; Brunec et al., 2018; Shin & DuBrow, 2020; Bird, 2020; Maurer & Nadel, 2021), a process that might also happen retroactively, *after* experiences have been obtained (Clewett et al., 2019). This formation of segmented memory traces is thought to be driven by various factors, including inferred changes in the environment (DuBrow et al., 2017), prediction error signals elicited by reward outcomes (Rouhani et al., 2020) or discontinuities in the statistical structure of the environment (Gershman et al., 2014). A practical approach for human research therefore seems to be to define events as "meaningful" units of experience (Bird, 2020) within the current experimental paradigm, and to potentially formalize them as states in an MDP. Finally, in understanding memory as a constructive process, it is important to note that neural task representations may change from perception to reactivation (Favila et al., 2020). This aspect is particularly crucial for the study of replay in humans, because activity patterns that are expected to reactivate are commonly determined based on simple localizer tasks that do not involve mnemonic task components. The brain might have already transformed its input data to a representation that is

different from what the researcher was hoping to see re-merge from replayed activity patterns.

2.2 Less forgetting

Increasing the speed of learning is an important computational benefit of replay, but not the only one. Many statistical learning mechanisms were built under the assumption that the agent encounters its environment entirely at random, and therefore can learn from examples that are independent and identically distributed (i.i.d.). This is often not true for the type of agent-environment interactions encountered in real life, and has been a particular problem for neural networks which use stochastic gradient descent (SGD) for learning, for which i.i.d. samples are one key requirement. Using replay to learn from experience that is not i.i.d. has been an important determinant of the renewed success of neural networks in dealing with RL-type problems (e.g., Mnih et al., 2015). Moreover, as we outline below, although RL models acknowledge that successive events are dependent on each other (see Box 1), they still benefit from replay, especially if an agent cannot freely sample all possible events in a given environment.

Temporal auto-correlation in the transition structure of real-world scenarios is an important property of the environment that can be hardly changed. Why does this cause problems for gradual learning algorithms? While gradual learning mechanisms are able to integrate experiences over longer periods, they still emphasize the most recent experience, which can cause the agent to forget about important past experiences. This is particularly apparent in a setting in which an agent engages in two tasks in a *blocked* manner. If a DNN is first trained to perform a task A and subsequently trained with another task B, performance on task A drops dramatically, as if the network forgot how to solve A or that learning task B interferes with what was learned about task A. This problem is known as catastrophic forgetting, or catastrophic interference, and has long been recognized as a major problem in the ML field (McCloskey & Cohen, 1989; Ratcliff, 1990; French, 1999; Hassabis & Maguire, 2007; Kumaran et al., 2016; Parisi et al., 2019). Catastrophic forgetting is one of the main reasons why artificial agents can usually learn a single task quite well but subsequent training on a different task results in poor performance on the previously learned task. This prevents the agent from achieving competencies across multiple

tasks, which comes relatively easily to humans. Catastrophic interference can also be understood as an issue threatening the stability of a cognitive map representation (Gupta et al., 2010; McClelland et al., 1995; O'Reilly & McClelland, 1994).

A potential solution to catastrophic interference is *interleaved learning* where new experiences are interleaved with existing knowledge to reconcile competing memory representations (McClelland et al., 1995). This influential idea, rooted in the complementary learning systems (CLS) theory (McClelland et al., 1995; O'Reilly et al., 2014; Schapiro, Turk-Browne, et al., 2017), also suggests that replay may be the mechanism that agents can use to "mentally" interleave past with present experience. While the DNN tunes its connection weights to solve task A, the learning experience is stored as episodes in a memory buffer. During learning of task B, previous experience with task A is integrated during offline periods via a replay-like mechanism, allowing the agent to perform well on both tasks. Shin et al. (2017), for instance, proposed an approach that learns a generative model based on experience with one classification task A. When switching to an independent classification task B, the system is retrained using a combination of new task data and fictitious sequences from the generative model, resulting in rapid generalization to the new task with little performance loss. Similarly, implementing replay in this way in DNNs can help to overcome performance deficits in incremental task learning scenarios and continuous task environments (Mnih et al., 2015; van de Ven & Tolias, 2018; van de Ven et al., 2020).

Humans and other animals do not necessarily seem to suffer from the computational problem of catastrophic interference. They can solve a wide set of tasks throughout their lifetime, despite temporal autocorrelation of experience and even learn well from blocked experience which troubles DNNs (Flesch et al., 2018). The idea that this ability might be related to replay (Antony & Schapiro, 2019) is supported by several studies. Karlsson and Frank (2009) for instance have observed replay of episodes from a remote spatial context. Reactivation studies in humans of previously learned events in the hippocampus that overlap with newly encoded memories leads to better retention (Kuhl et al., 2010).

Another challenge arises when learning must occur in environments where some events happen rarely, but are nevertheless of great significance for the agent's success or well-being. Evidence that replay might be used to mitigate this problem

in animals comes from studies showing that actions which should be avoided will be reactivated, like paths to a shock zone (Wu et al., 2017) or paths to devalued outcomes (Carey et al., 2019). In addition to learning about events that should be avoided, replaying rare events that are only weakly encoded could allow the agent to form a stable representation of the entire environment even if only a smaller subset is experienced frequently. In Gupta et al. (2010) non-local replay was stronger for remote sequences if they were experienced less frequently. Using MEG in humans, Jafarpour et al. (2017) showed that stronger reactivation of one of three previously encoded stimuli was determined by how weakly the stimulus was attended to during encoding. These findings are supported by an fMRI study by Schapiro et al. (2018), who demonstrated that older, less well remembered task stimuli were selectively reactivated during a subsequent rest period resulting in memory improvement, an effect that was particularly strong in participants who slept in the 12-hour interval between test sessions. In another study, the benefits of targeted memory reactivation (TMR) were stronger for weakly learned information (Tambini et al., 2017). Further, replay-associated electroencephalography (EEG) sleep spindles during a nap follow- ing difficult but not easy memory encoding were related to improved subsequent memory performance (Schmidt et al., 2006). Overall, the beneficial role of sleep for weakly encoded information is well documented by several studies (Drosopoulos et al., 2007; Peters et al., 2007; Diekelmann et al., 2010; McDevitt et al., 2015; Sio et al., 2013; Kuriyama et al., 2004; Cairney et al., 2016; Talamini et al., 2008; Schapiro, McDevitt, et al., 2017). Together, replay liberates an agent from needing to consider transitions only in proportion to how many times they were experienced. Instead, replay can flexibly increase or decrease the number of opportunities for learning experiences from single episodes.

2.3   Re-inventing the past

In our introductory example (see Figure 1), replayed content was a close reflection of past experience. Replay occurred immediately after an episode was experienced and reflected past trajectories from start to finish, albeit in reverse order. This setup stands in contrast to the ideas discussed in the section on forgetting, which imply that replay must not necessarily respect the structure of experiences, but could, for instance, change the order and frequency of events. Beyond dealing with unevenly

distributed events, replay could in fact be used to arbitrarily alter the distribution of events upon which memory is built.

Such a reorganization of experience also requires a different understanding of what constitutes an episode. In our simulation, we had assumed that the minimal unit of replayed content is one entire sequence of states, actions and rewards that occurred between a random start position and the encounter of a goal. This meant that episodes were often quite long, involving several hundreds of steps particularly early in learning (see Figure 1C), and that the transitions between locations had to be replayed in the order in which they were experienced. In order for replay to be able to reorganize experience, an episode is often divided into a much smaller unit of experience, a simple sequence of just one state, one action, one reward and the next state, known as a (s_t, a_t, r_t, s_{t+1})-tuple. Arguably, replaying such minimal experiences risks losing the benefits of temporal credit assignment, which come with backward replay of full trajectories, because values will not necessarily propagate to starting positions. But it does also have benefits, which we will discuss below. In consequence, the question of what constitutes an atomic unit from the perspective of replay has important implications and is therefore actively debated (see Box 2).

Using minimal transitions, there is a large variety of ways in which replay may alter the structure of experiences that have been discussed in the ML and neuroscience literature. One possibility is to reactivate (s_t, a_t, r_t, s_{t+1})-tuples in a random order, which artificially crafts similar conditions as during supervised learning that allows artificial neural networks trained with SGD to excel (Botvinick et al., 2020). Such uniformly sampled (s_t, a_t, r_t, s_{t+1})-tuples have therefore played an important role in adapting DNNs to RL problems, such as the famous Deep Q-Network (DQN) (Mnih et al., 2015). Interestingly, some animal studies have also found replay of seemingly random trajectories following exploration of a familiar open-field arena (Stella et al., 2019). Note however, that Stella et al. (2019) still observed replay of sequentially organized transitions that reflected the spatial constraints of the environment, whereas random replay used in ML can involve sets of single transitions that do not form sequential trajectories. This highlights the different understanding of replay content in ML and neuroscience. Additionally, most animal studies impose the assumption of sequentiality during data analysis, and would discard fully random activation of transitions as noise. In both ML and neuroscience, however, random replay refers to

sequential reactivation that is unrelated to previously experienced action sequences, and can be seen at one extreme of a continuum describing how closely replay matches actual behavioral sequences (see Swanson et al., 2020, their Figure 2).

Another particularly important idea from the ML literature is to prioritize replay of transitions that led to large surprises, with the idea that some experiences are more informative than others, resulting in more efficient learning (Schaul et al., 2015; Horgan et al., 2018). Such prioritized replay records a prediction error (PE), the difference between the expected and actual reward, for every encountered transition and uses this signal to select experiences for replay later. This method is very similar to, and inspired by, an earlier algorithm in model-based planning known as prioritized sweeping, which selects the state to be updated according to the magnitude of the change in value upon the execution of the update (Moore & Atkeson, 1993; Peng & Williams, 1993; Andre et al., 1998). Based on the success of the prioritized replay approach, more frequent sampling of transitions with a high absolute TD error is now a common approach to train DNNs (Fedus et al., 2020). Using RL models, Mattar and Daw (2018) extended previous approaches by focusing prioritization on behaviorally relevant states that are likely to be encountered again in the future and those transitions where a policy change would yield the largest net increase in discounted future reward. Note that although prioritization algorithms assume selection of replay content on the level of individual transitions, they can, under some circumstances, still lead to sequential replay. This is true, for example, for the model of Mattar and Daw (2018), because expectations about increases in future reward are themselves often auto-correlated.

The core idea that replay should be influenced by reward and surprise is in line with several animal studies. Place cell sequences associated with reward are replayed more often (Ólafsdóttir et al., 2015; Foster & Wilson, 2006; Bhattarai et al., 2019), in particular those with a high PE (Singer & Frank, 2009; Michon et al., 2019; Roscow et al., 2019), and the rate of SWRs is also influenced by reward (Ambrose et al., 2016; Singer & Frank, 2009). These results highlight replay's role in credit assignment, as discussed above. In human neuroimaging studies, hippocampal activity is modulated by reward magnitude (Wolosin et al., 2012; Igloi et al., 2015). This is in line with the link between backward replay and selection of transitions based on changes in value proposed by Mattar and Daw (2018). Replay is also more

likely to contain behaviorally significant locations, such as the current goal (Gupta et al., 2010; Pfeiffer & Foster, 2013; Ólafsdóttir et al., 2015) and is biased by novelty (Cheng & Frank, 2008; Foster & Wilson, 2006). It has also been observed that optogenetic manipulation of dopaminergic input neurons, thought to signal PEs, increase replay during subsequent sleep (McNamara et al., 2014).

Reactivating smaller sequences in a different order from what was experienced can also be used to connect experiences in novel ways or strengthen weakly learned relationships. Under such conditions replay can correspond more closely to sampling from an internal model of the environment, rather than a veridical recapitulation of past experiences (Sutton, 1991). Among the most early ideas about replay, the Dyna architecture (Sutton, 1991) used an internal model to generate experiences that were then used to train a model-free agent. Indeed, replay can be seen as a way to blur the lines between model-free RL, such as the Q-learning method introduced in Box 1, and model-based RL, during which the agent stores an explicit model of the environment and can use it for planning (van Seijen & Sutton, 2015; Russek et al., 2017; Momennejad et al., 2017).

Another dimension of experiential reorganization is to replay experiences that happened long ago, or even refer to imagined experiences in the future. During behavior, replay events can switch between reflecting immediately preceding, upcoming or more remote episodes, depending on the behavioral state of the animal at the time of replay (Pfeiffer & Foster, 2013; Ólafsdóttir et al., 2017). Moreover, even single replay events can depict more than one trajectory, such as the next one and the path the animal will take after reaching the goal location (Pfeiffer & Foster, 2013), as if representing a multi-step planning process (Foster, 2017; Miller & Venditto, 2021). Another reason why experiences from a more distant past might be replayed is simply that the agent is using a period during which it does not have to engage with the environment to optimize memory. This is particularly apparent for replay during sleep, when the brain has idle time to process experiences while not being actively engaged with any task. Sleep replay has frequently been observed in animals and humans, and been linked in particular to memory consolidation. Following sleep, memory interference is reduced (Baran et al., 2010; McDevitt et al., 2015) and memory integration or differentiation has been found in fMRI patterns after a delay period with sleep (Favila et al., 2016; Tompary & Davachi, 2017).

A final aspect of re-inventing the past relates to re-considering the usefulness of past experiences in light of one's knowledge about a goal. The RL framework presented thus far is aimed at the pursuit of a single goal (e.g., the single reward location in our grid world, see Figure 1). However, in many real-world applications, such as the movements of a robotic arm that needs to pick and place objects, an RL strategy incorporating multiple goals would be far more beneficial. Consider again the grid world example in Figure 1, but this time the agent can only move a finite number of steps. Since there is only one goal state that returns a reward, most of the transitions do not land in the goal state and therefore receive no reward. In such a sparse binary reward situation, where success only results from those sequences of transitions ending in the goal state, most sequences of transitions end in uninformative failures, often related to early termination without reward ("giving up"). For instance, when an agent gives up because a goal was not found after a particular amount of time, it can not know how close it was to the goal. Humans, however, can learn from failure as well as success. Inspired by this idea, an ML technique known as hindsight experience replay (Andrychowicz et al., 2017) is used to relabel the unsuccessful transitions by simply changing the goal state, such that the transitions would now be considered as successful under the new goal, thus contributing to the agent's learning. To the best of our knowledge, no directly equivalent observation has been made in the brain.

2.4 Planning for a better future

So far, we have mainly focused on the various ways in which replay serves learning and memory. We have shown, for instance, how replay can be used to speed up learning Q-values, which can be used for behavior that maximizes reward (Lin, 1991, see Figure 1). Yet, computational, psychological and neuroscientific research has pointed out the importance of another mechanism that is crucial for goal-directed behavior: planning. A core aspect of this process is the prospective, i.e., future-oriented, evaluation during which an agent deliberates which of the available sequences of actions and states leads to the best among several potential outcomes. Notably, planning requires a mental model, or cognitive map (Tolman, 1938, 1948), of the environment, that describes the agent's knowledge about the transition structure of events (e.g., paths between locations in a maze or contingencies between non-spatial task states)

including the outcomes at each potential location. Knowledge about the causal structure of the environment allows an agent to predict and compare the outcomes of sequences of states and actions and to choose the one that yields most reward. Figure 2 provides an illustration of how planning differs from the other two aspects of cognition we considered so far, acting and learning. While the distinction between learning and planning is useful, we will see below how agents can learn from planning processes.
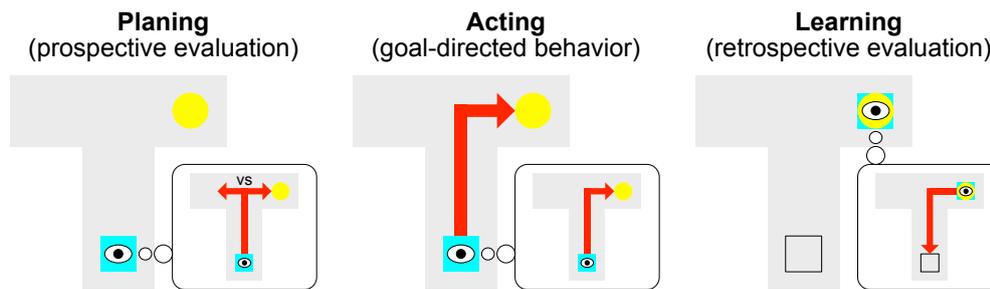
Figure 2: **Illustration of replay content.** The agent is represented by the blue square. The goal location is indicated by the yellow circle. Red arrows indicate behavioral trajectories of the agent in the T-maze environment or internally generated trajectories during replay. During planning (left panel) the agent engages in prospective evaluation of potential behavioral sequences in order to select the one that leads to the goal location, using forward replay. During goal-directed behavior (center panel) the agent instantiates the behavioral trajectory that is immediately relevant to prepare action using forward replay. During learning (right panel), the agent retrospectively evaluates its previous behavior, usually upon reaching a goal location, using backward replay. © Wittkuhn et al., doi: 10.6084/m9.figshare.14261636.v1, CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

Within the RL framework, the difference between acting based on learned (cached) values versus acting based on an internal planning process is embodied by the distinction between model-free and model-based systems (Sutton & Barto, 2018; Daw et al., 2005). In model-based RL, the agent uses experience to learn a model of the environment that is described by a function that relates the current state s_t and action a_t to the next state s_{t+1} and the reward r_t (see Box 1). The agent can use this model at decision time or during offline periods to simulate experience (Sutton, 1990), i.e., individual (s_t, a_t, r_t, s_{t+1})-tuples, or longer episodes spanning several transitions. This allows the agent to update its learned values or to determine which action would be best to execute next, considering the rewards obtained and how the environment would change if a particular action was taken. This deliberation

process has two advantages. First, planning allows the agent to remain in the safety of mental imagination and avoid the risk of suffering from potentially harmful consequences. Second, planning can be used to decide between never-experienced, purely hypothetical courses of action (Liu et al., 2019), a feat which would not be possible with purely experience-based replay.

Despite these differences between planning and learning, much work in RL has emphasized their similarities (Sutton, 1990, 1991; Sutton et al., 2012; van Seijen & Sutton, 2015), and points to a function of planning that goes beyond deliberation. As we have seen in our discussion of Lin, 1992, the same learning mechanisms can be applied to real or simulated experience. Planning can thus not only be used to determine immediate behavior, but also to shape value functions, a process referred to as background planning (as opposed to the decision-time planning and deliberation considered above, see Pezzulo et al., 2019). A case in point regarding the similarities between planning and learning is illustrated by the Dyna architecture (Sutton, 1990, 1991). Just as any model-free RL agent, a Dyna agent selects actions according to learned Q-values, and uses experiences to update these Q-values. But it also uses experiences to observe rewards and subsequent states that follow each state-action pair, to update an internal model of the environment. Importantly, in Dyna the model is then used to train the model-free agent by producing simulated episodes that will be used to update the agent's Q-values, just like real experiences. Two implications from this perspective should be noted. First, replay might be used not only to update value functions, but also to update a model of the environment (see also, Penagos et al., 2017; Pezzulo et al., 2019, 2017; Momennejad, 2020). Second, such a mechanism also suggests that replay elicits reward prediction error signals, which could explain why striatal replay has been observed before (Lansink et al., 2008, 2009; Pennartz, 2004). Notably, reward prediction errors might be accompanied by state-prediction errors, which in the brain might both be conveyed by dopaminergic signals (see e.g., Sharpe et al., 2017; Gardner et al., 2018).

It should be noted that an interaction between deliberation and learning may arise. Entering a novel environment or changes in a familiar environment should increase deliberation because a new model has to be learned or an existing model has to be updated. But if the contingencies in the environment remain unchanged, the need for the agent to update its model and deliberate its choices diminishes

with learning. At the same time, decisions become more habitual and less deliberate (Dolan & Dayan, 2013). Thus, replayed trajectories in the hippocampus that evaluate all potential trajectories might be only predictive of behavior during earlier phases of learning (Singer et al., 2013) or vanish from the hippocampus (e.g., Wimmer & Büchel, 2019) when behavior becomes stereotyped. Findings by Papale et al. (2016) demonstrate an inverse relationship between SWRs at reward sites and deliberation at choice points.

While the potential benefits of replay for planning have been recognized early on in RL (Sutton, 1990), serious consideration of this aspect only appeared in neuroscience later, where studies demonstrated replay events in the awake state, often during short pauses from active behavior (e.g., Kudrimoti et al., 1999; Foster & Wilson, 2006; Csicsvari et al., 2007; Diba & Buzsáki, 2007; Kurth-Nelson et al., 2016; Eldar et al., 2020). This allowed researchers to draw closer correspondence between the replayed and the behavioral trajectories, and has resulted in a wealth of findings supporting the idea that replay supports model-based planning in animals as well as humans (for reviews, see e.g., Yu & Frank, 2015; Pezzulo et al., 2019; Wang et al., 2020; Tambini & Davachi, 2019; Carr et al., 2011; Ólafsdóttir et al., 2018). Disruption of awake hippocampal SWRs during a spatial alternation task specifically impaired the ability to decide between two trajectories to alternating goal locations, whereas place field representations, reactivation during rest, and other navigation behavior remained intact (Jadhav et al., 2012). Replay events in the awake state predominantly co-occur with SWRs during short pauses from ongoing exploratory behavior. Replay trajectories during awake SWRs often start at the current location of the animal (Diba & Buzsáki, 2007; Davidson et al., 2009; Karlsson & Frank, 2009; Pfeiffer & Foster, 2013; Singer et al., 2013; Ambrose et al., 2016) and end at the goal location (Dupret et al., 2010; Pfeiffer & Foster, 2013).

A behavioral correlate of deliberation was already described in the 1930s in rodents (Tolman, 1926; Muenzinger & Fletcher, 1936), who tend to pause at a decision point to look back and forth between possible paths, a behavior called vicarious trial and error (VTE) (for review, see Redish, 2016). Later studies found that during VTE events, hippocampal place cells associated with theta sequences sweep ahead from the animal's current location (Johnson & Redish, 2007; Wikenheiser & Redish, 2015b; Amemiya & Redish, 2016; Papale et al., 2016). It was also found that during

VTE-like behavior, place cell activity influenced the formation of place fields thought to stabilize the cognitive map (Monaco et al., 2014). Note that VTE-like replay is often accompanied by theta sequences, which differ from SWR in their neurophysiology. Nonetheless, both can be described as sequential activation of hippocampal cell populations, a simplifying assumption that is helpful from a computational perspective (Foster, 2017; Pezzulo et al., 2019). Recently, theta sequences have been shown to quickly cycle between possible future trajectories (Kay et al., 2020), and increases in theta power in the MTL have been observed in humans in a spatial planning task (Kaplan et al., 2020). In human fMRI, blood-oxygen-level dependent (BOLD) activity in the hippocampus has been shown to increase with deliberation time when deciding between two food items with similar value (Bakkour et al., 2019) and hippocampal activity patterns reflect routes to navigational goal locations (Brown et al., 2016). Another study has found that when humans re-learn outcomes associated with choices at lower levels of a decision tree, the extent to which higher levels of the decision tree are reactivated during rest correlates with how much their decisions change, to reach the new downstream reward states (Momennejad et al., 2018).

It should be noted that although awake replay during deliberation of future choices is often related to improved task performance, the replayed trajectories do not necessarily reveal the behavioral trajectory the animal would subsequently take (Johnson & Redish, 2007; Singer et al., 2013). In the study by Singer et al. (2013), hippocampal replay during SWRs that preceded correct choices reflected trajectories for the correct *and* incorrect option in a two-alternative W-maze. Once correct performance became stable (at 85% correct), replayed trajectories shifted to represent the correct future choice more frequently than the incorrect one (Singer et al., 2013). One interpretation of these findings can be that the hippocampus uses replay to evaluate all potential trajectories and the behaviorally relevant trajectory is instantiated in a different brain region.

One additional complicating factor regarding the relation between replay and subsequent behavior concerns the task setting and motivational state of the animal (e.g. Carey et al., 2019; Wu et al., 2017). As discussed above, replayed place cell sequences can represent the trajectory into a shock zone which is subsequently avoided (Wu et al., 2017). This might have the purpose to learn strongly from and not forget about significant outcomes, and thus in this circumstance replay

is related to avoiding rather than initiating trajectories. In line with this idea, a growing literature on computational psychiatry considers that replay could underlie symptoms like avoidance and rumination that characterize psychiatric disorders like anxiety (Gagne et al., 2018; Heller & Bagot, 2020; Mobbs et al., 2020).

If replay is related to planning, but the ultimate determination of behavior also depends on other brain areas, then replayed trajectories might be influenced by concurrent reactivation outside the hippocampus, such as the amygdala in the case of aversive outcomes (Girardeau et al., 2017). A number of studies sheds light on how replay in the hippocampus is coordinated with other brain regions to instantiate behavior. Replay is known to be coordinated with PFC (Jadhav et al., 2016; Pezzulo et al., 2014; Peyrache et al., 2009; Tang et al., 2017), and some work has placed particular focus on the interaction of hippocampal replay and the OFC (Steiner & Redish, 2012; Schuck & Niv, 2019). Indeed, disruption of nearby medial PFC attenuated components of hippocampal theta sequences representing the current location of the animal (Schmidt et al., 2019) and suppression of hippocampal input impaired the integration of task state structure in the OFC (Wikenheiser et al., 2017). Similarly, a recent study in humans has found that hippocampal replay at rest was not directly linked to behavior during a task (Schuck & Niv, 2019). Rather, replay at rest was linked to how well the different task-states were represented in the OFC, which in turn were linked to behavior (Schuck & Niv, 2019). Outside of the PFC, entorhinal grid cells that are thought to enable vector-based spatial navigation likely contribute to planning, as implicated in computational work (see e.g., Erdem & Hasselmo, 2012; Bush et al., 2015).

Finally, we want to highlight that most of the above interpretations of replay in the neuroscience literature rest on the idea that replay is shaped by previously experienced behavioral trajectories. Yet, this account was challenged by a series of findings reporting apparent "preplay" of place cell sequences before the environment was ever experienced (Dragoi & Tonegawa, 2011, 2013). These findings highlight that apparent sequentiality can also reflect hippocampal cell assemblies that are connected in a way that constrains sequential firing, even prior to experience of a new maze. Nevertheless, other findings indicate that previous experience is required for such spontaneous sequential activation to occur (Silva et al., 2015). Recent computational work has suggested a link between preplay and efficient learning, arguing that

preexisting internal sequences could be used as a dynamical reservoir (Leibold, 2020). The extent to which the hippocampus is able to seemingly preplay novel experiences could depend on the similarity between pre-existing hippocampal representations and new memories about to be formed (Eichenbaum, 2015). Methodologically, this nevertheless highlights the necessity to compare pre- versus post-task replay (e.g. Buhry et al., 2011), as shown by recent research that observed pre-vs.-post changes in replay can indeed be explained by cell activation and firing rate correlations during experience (Farooq et al., 2019).

2.5 Inference and generalization

Apart from its role in learning and planning, recent developments in ML and neuroscience research suggest that replay also contributes to inference and generalization (for previous reviews and perspectives, see Kumaran, 2012; Kumaran & McClelland, 2012; Cazé et al., 2018; Herszage & Censor, 2018; Lewis et al., 2018; Momennejad, 2020). A recent theme in this domain has been to build artificial agents that learn generative models from experience, which can then be used to infer new connections based on latent structural rules (Evans & Burgess, 2019), infer the correct context when given new data (Stoianov et al., 2020) and generalize information to new tasks to mitigate performance losses (Shin et al., 2017). In the model proposed by Stoianov et al. (2020), for instance, trajectories through a maze are used to learn a generative model, which can produce new trajectories consistent with the current maze structure during offline periods. As new mazes are learned, novel trajectories continue to be generated offline, but from all the mazes that have been experienced, preventing information about any one maze from being lost (similar to our considerations about forgetting above). The hierarchical structure of the model results in trajectories being clustered into distinct maze contexts, which allows maze categories to be inferred when presented with new data. Unlike replay used in other contexts, the model by Stoianov et al. (2020) does not suggest that prioritized replay helps to improve behavioral outcomes. Generative replay that was prioritized based on how surprising observations were under the generative model increased the number of reactivation events that contained important goal locations but did not further improve inference performance (Stoianov et al., 2020).

The clustering of trajectories seen in the model above is related to a broader

theoretical view, which has emphasized that separate encoding of transition information and sensory information during learning will allow knowledge about transitions to be reused across situations with structural similarities but new sensory specifics (Behrens et al., 2018; Liu et al., 2019; Baram et al., 2020; Whittington et al., 2020). Because replay provides a strong candidate mechanism for learning about transition structure (Stoianov et al., 2020), replay of abstract (sensory-independent) transition information could help to build representations of task structure that can be generalized and used to guide behaviour in new sensory environments (Liu et al., 2019) or combined with sensory observations to make inferences about the current environment (Evans & Burgess, 2019; Stoianov et al., 2020).

Another major computational approach has focused on replay as a mechanism to learn successor representations, a predictive representation that reflects the expected future visitation of states, given the current state (Dayan, 1993). Unlike the one-step transition matrices that are known as models in model-based RL, the successor matrix can reflect non-adjacent dependencies. This allows the agent to understand relationships between a state and multiple successor states, knowledge which can be used to solve inference problems, such as finding the shortest path to a new reward location (Russek et al., 2017; Momennejad et al., 2017). Successor representations have similar properties to place fields, skewing in the opposite direction of travel and over-representing goal locations, while the eigenvectors of a successor matrix can partition the environment into clusters that help planning, resembling entorhinal grid cells in some spatial contexts (Stachenfeld et al., 2017). Hippocampal-entorhinal fMRI signals have also been shown to reflect relationships between successive non-spatial objects organized in a graph (Schapiro et al., 2013; Garvert et al., 2017), consistent with the SR (Stachenfeld et al., 2017). Critically, replay of past experiences could be used to update the successor matrix during offline periods (Russek et al., 2017). In recent work, Russek et al. (2017) proposed replay of (s_t, a_t, s_{t+1})-tuples that are prioritized by recency, in which rewards are not needed to update the successor matrix. Using this approach, it was shown that learning SRs with offline replay gives an agent unique benefits compared to agents without replay. In particular, the fast updating of successor states through replay allowed the agent to quickly infer policy updates needed to adapt to changes in the task environment, like a new barrier, that affect the state transition structure.

A potential role of replay in generalization and inference has also been suggested by neuroscientific studies. In a recent study from Barron et al. (2020), hippocampal cells selective to cues and rewarding outcomes that have not been directly experienced together, but whose relationship can be inferred based on sensory pre-conditioning (Brogden, 1939), were found to be co-active during SWR events (Barron et al., 2020). These cells also tended to be reactivated during SWR events in a specific order, with reward selective cells reactivated prior to cue selective cells akin to backward replay. Using MEG, researchers have discovered that visual stimuli are reactivated in a non-experienced order that was based on prior learning of a rule about how items should be reported (Liu et al., 2019). This indicates that sequential reactivation is able to combine prior learning with new sensory inputs to produce behavior relevant to new environments. This aligns with the theoretical proposal that decomposing information into separate structural and sensory representations can be used to generalize structural relationships to situations with different sensory specifics but similar underlying structure (Behrens et al., 2018; Baram et al., 2020; Whittington et al., 2020). Nonetheless, this theoretical view is not restricted to replay events and, as a result, there is still limited work testing this idea in the context of replay. Finally, recordings in the hippocampus and PFC have shown that hippocampal place cells are reactivated with subsets of prefrontal cells that encode generalizable task elements (Yu et al., 2018) and that at least some medial PFC neurons involved in replay have generalized firing fields that cover multiple starting locations or multiple goal locations within a maze (Kaefer et al., 2020). Together, these results suggest that replay could also contribute to generalization through coordinating the appropriate reactivation of PFC neurons.

2.6 Representation learning

The analogy between hippocampal replay and replay in RL models sheds light on another important aspect: states – the agent's internal representation of its environment. State representations are crucial for RL because they can be optimized for efficient learning and structure the task at hand as an MDP (see Box 1). In this last section, we highlight findings indicating that replay is not specific to spatial locations and could instead involve such task-dependent state representations. We argue that replaying states has unique benefits as opposed to replaying only observations.

Moreover, we speculate that replay might also have a role in *learning* the representations that guide behavior. As such, replay could offer a window into the operations the brain performs to craft useful representations of the possible task states.

The main idea of states is that they meaningfully delineate different situations in which a specific action has a particular value. In a game of chess, for instance, the states are the different configurations of the pieces on the board. Given a board configuration, one can learn how beneficial it is to move a particular piece. Making the same move in a different position would have very different consequences, and differentiating the value of possible moves by board configuration is therefore paramount. The color of the shirt of your opponent, in contrast, may be a salient visual observation, but it has no implications for the values of the moves, and it therefore need not be part of the state. This implies that task state representations can be shaped by attentional mechanisms that focus on task-relevant dimensions (e.g., Niv et al., 2015; Leong et al., 2017). Besides filtering irrelevant information, good state representations have to overcome the incompleteness of sensory inputs, which by themselves sometime do not contain all information needed to determine which action will lead to which outcome. Outcomes, for instance, often depend on non-observable context, such as past events, but internal representations can be used to combine current experience with observations of the past to determine the current task state (Wilson et al., 2014; Schuck et al., 2018). For example, knowledge about how your opponent has played in the past can lead you to select a different move in the same situation as in a previous match. Finally, state representations should be robust to noise, leverage similarity among the states to determine what information can be generalized, and identify the hidden causes that may underlie sensory observations (e.g., Gershman & Niv, 2010). Different chess positions might for instance share structural similarities that can be exploited with the same strategy. Defining the states of a task is therefore crucial for the success of an agent. As Dayan (1993) has put it: "difficult problems can be rendered trivial if looked at in the correct way" (p. 613). Importantly, since the agent does not know from the start "how to look at" the problem in the right way, learning useful state representations constitutes a major challenge for intelligent agents (Bengio et al., 2013; Niv, 2019).

In traditional RL models, the input to the algorithm is often engineered by hand to reflect the true state of the environment. Replay in these models therefore always

reflects states, because no differentiation between sensory observations and states is made. In DNNs, the model is equipped with a convolutional network that converts sensory observations into states, over which Q-learning operates (e.g., Mnih et al., 2015). Replay in DNNs involves a memory buffer of observations, which are converted into states inside the network during replay, allowing efficient learning. These approaches therefore suggest that replay either involves only states, or observations as well as states.

In animals, several findings indicate that a large variety of representations, including non-spatial sensory as well as state-like representations, might be replayed. First, the firing of hippocampal "place" cells can reflect a number of non-spatial aspects of the environment, if they are task-relevant, such as sounds (Aronov et al., 2017), time (MacDonald et al., 2011) or successor representations (Stachenfeld et al., 2017; but see O'Keefe & Krupic, 2021). More directly, one fMRI study by Schuck and Niv (2019) has found that sequential hippocampal replay during post-task rest reflected the non-spatial states of a sequential decision-making task. Importantly, observed transitions between decoded replay events were best explained by replay of states that include non-observable task aspects, such as information from the previous trial, rather than by replay of sensory features of the task stimuli alone. This study therefore provides direct evidence for the idea that replay involves state representations that are optimized for the operation of RL algorithms. In an MEG-study by Liu et al. (2019), human participants first learned an abstract rule how objects should be ordered in a sequence and later replayed a novel set of objects according to the learned rule rather than in order of experience. Replayed sequences consisted of factorized representations of sensory objects, the identity of the sequence they belonged to, as well as the position within that sequence, supporting the notion that replay is not limited to one kind of information. Moreover, Jadhav et al. (2012) showed that disruption of SWRs in a spatial alternation task impaired navigation when it required unobservable knowledge of the previous trial, thus hinting at the activation of state representations rather than observations during replay.

Other evidence suggests that replay involves multiple representations which are reactivated in parallel. These representations reflect visual (Ji & Wilson, 2006; Wittkuhn & Schuck, 2021), auditory (Rothschild et al., 2016) or grid-like (Ólafsdóttir et al., 2016, 2017; O'Neill et al., 2017) information and have been observed during

interactions between hippocampus and OFC (for reviews, see Wikenheiser & Redish, 2015a; Wikenheiser & Schoenbaum, 2016), which might relate to the agent's task state representations (Kaplan et al., 2017). Disruption of the medial PFC particularly attenuated components of hippocampal theta sequences representing the current location of the animal (Schmidt et al., 2019) and suppression of hippocampal input to the OFC impaired the integration of task state structure (Wikenheiser et al., 2017). Conversely, disruption of SWRs during sleep impaired the integrity of hippocampal maps but they re-emerged following re-learning (Gridchyn et al., 2020) suggesting that relevant maps are stored in brain areas other than the hippocampus (Niethard & Born, 2020). It should also be noted that a number of investigations have shown replay events outside the hippocampus need not be coordinated with hippocampal activity (O'Neill et al., 2017; Kaefer et al., 2020; Wittkuhn & Schuck, 2021). In sum, evidence is mounting that hippocampal "place cell" firing can reflect a variety of non-spatial but task-relevant aspects (e.g., Aronov et al., 2017), replay occurs in a wide variety of brain areas that encode different informational aspects, which at least partially reflect an animals' understanding of what is task relevant and even that replay directly reflects partially observable task states (Schuck & Niv, 2019).

Replaying states and replaying different levels of representation might convey unique benefits. RL models benefit from transition information between states, but they could be affected adversely if transitions of task-irrelevant aspects affect the agents internal model. For example, representing states as specific locations in physical space will result in a transition matrix that is different from a transition matrix of more abstract task states, but if spatial position is irrelevant to the task at hand, then transitions between locations are not needed for learning and planning. In other words: the transition matrix, and therefore the foundation of an agent's ability to plan and simulate actions, fully depends on its internal state representations (Guo et al., 2020). In addition, observations are, under most circumstances, much more high dimensional than states, and storing them might require a much larger memory buffer than an approach in which states are stored (e.g., Shin et al., 2017). Thus, replay that has access to and operates on highly adaptive state representations seems computationally sensible.

If replay indeed operates on the level of states, and learning good state repre-

sentations is a major challenge for any agent, this begs another question: can replay also help to change the content of the representation itself? Some evidence suggests this might indeed be the case. Schuck and Niv (2019) observed that replay in the hippocampus during rest was related to better decodability of partially observable state representations from the OFC during the task. Moreover, decoding of state representations in the OFC increased over time. Although this evidence is correlational, it hints at some relation between replay and state representation learning. How could replay help to form better state representations? One important instance concerns successor representations (SRs), which provide an efficient way to incorporate knowledge about the transitions between states into the state definition. Computational work by Russek et al. (2017) has shown that SRs are not only useful state representations, but they can also be learned and updated through replay. More generally, information about state transitions can give rise to further graph analytical insights which are known to provide a good basis for state representations (Mahadevan & Maggioni, 2007). Sequential replay is a natural match as a mechanism to learn states encoding transitional information, and possibly might be involved in extracting graph properties from experienced transition structures, such as bottleneck states. This knowledge could then be used to form better states. Other work has highlighted that representations which predict latent embeddings of future observations are particularly useful (Guo et al., 2020). An evaluation of predictiveness could therefore be an important contribution of replay to state representation learning.

Some evidence suggests that the role of replay for state learning could go beyond the focus on information about transitions. SRs, for instance, can be extended to deal with partially observable task environments (Vértes & Sahani, 2019). Caselles-Dupré and colleagues proposed another interesting account that involves variational autoencoders (VAEs) (Caselles-Dupré et al., 2018, 2019). Building on earlier work that used generative models to circumvent the memory requirements of observational replay (Shin et al., 2017), Caselles-Dupré et al. (2019) proposed storing latent representations rather than observations, and using past experiences in this form of continually training a VAE that acts as a state model. Importantly, only by replaying past episodes can the VAE learn to form a state representation that allows the agent to act efficiently across more than one environment.

While these computational models represent intriguing proposals, to the best of our knowledge, not many empirical studies have shed light on whether state learning mechanisms provide a realistic account for biological replay. As we have shown, however, some of the above discussed neuroscientific results do suggest that replay is involved in the construction and maintenance of hidden state representations.

3   Conclusion and outlook

In this review, we have summarized the literature on replay in neuroscience and ML to showcase which computational benefits biological and artificial agents can gain from replaying previous experience. We have discussed five main computational benefits that, although overlapping, provide useful categories for thinking about what might motivate an agent to employ replay: faster learning and increased data efficiency, less forgetting, the reorganization of experience, planning and generalization. In addition we have argued that replayed content is much richer than a sequence of locations, and could reflect the agent's current state representation. State representations are often task- and context-dependent and are influenced by a range of factors, including the goal-relevant aspects of the agents observations, the transition structure of states, the location, number and value of goal locations and the motivational and metabolic state of the animal. We have argued that RL theory provides useful guidance to understand which form state representations might take in a given task, and which implications a particular state representation would have for an agent's behavior. Finally, we have discussed how replay might not only reflect but could help the agent to learn those states to begin with. While many questions in particular regarding the latter idea still remain, considering these factors will greatly help to determine what replayed representations represent and how replay updates decision-making policies that are used to control behavior.

4 Declaration of competing interest

The authors declare no competing interests.

5 Acknowledgements

Glossary

BOLD blood-oxygen-level dependent.

CLS complementary learning systems.

DNN deep neural network.

DQN Deep Q-Network.

EEG electroencephalography.

fMRI functional magnetic resonance imaging.

i.i.d. independent and identically distributed.

MDP Markov decision process.

MEG magnetoencephalography.

ML machine learning.

MTL medial temporal lobe.

OFC orbitofrontal cortex.

PE prediction error.

PFC prefrontal cortex.

RL reinforcement learning.

SGD stochastic gradient descent.

SR successor representation.

STDP spike timing dependent plasticity.

SWR sharp wave-ripple.

TD temporal difference.

TMR targeted memory reactivation.

VAE variational autoencoder.

VTE vicarious trial and error.

6   References

Ambrose, R. E., Pfeiffer, B. E., & Foster, D. J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron*, *91*(5), 1124–1136. doi: 10.1016/j.neuron.2016.07.047

Amemiya, S., & Redish, A. D. (2016). Manipulating decisiveness in decision making: Effects of clonidine on hippocampal search strategies. *Journal of Neuroscience*, *36*(3), 814–827. doi: 10.1523/JNEUROSCI.2595-15.2016

Andre, D., Friedman, N., & Parr, R. (1998). Generalized prioritized sweeping. In *Advances in neural information processing systems.* MIT Press.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., . . . Zaremba, W. (2017). Hindsight experience replay. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 5048–5058). Curran Associates, Inc.

Antony, J. W., & Schapiro, A. C. (2019). Active and effective replay: systems consolidation reconsidered again. *Nature Reviews Neuroscience*, *20*(8), 506–507.   doi: 10.1038/s41583-019-0191-8

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719–722.   doi: 10.1038/nature21692

Axmacher, N., Elger, C. E., & Fell, J. (2008). Ripples in the medial temporal lobe are relevant for human memory consolidation. *Brain*, *131*(7), 1806–1817.   doi: 10.1093/brain/awn103

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., . . . Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, *8*, e46080. doi: 10.7554/eLife.46080

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2020). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*.   doi: 10.1016/j.neuron.2020.11.024

Baran, B., Wilson, J., & Spencer, R. M. C. (2010). REM-dependent repair of competitive memory suppression. *Experimental Brain Research*, *203*(2), 471–477.   doi: 10.1007/s00221-010-2242-2

Barron, H. C., Reeve, H. M., Koolschijn, R. S., Perestenko, P. V., Shpektor, A., Nili, H., . . . Dupret, D. (2020). Neuronal computation underlying inferential reasoning in humans and mice. *Cell*, *183*(1), 228 - 243.e21.   doi: 10.1016/j.cell.2020.08.035

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509.   doi: 10.1016/j.neuron.2018.10.002

Bellman, R. (1957). *Dynamic programming.* Princeton University Press.

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, *362*(6415).   doi: 10.1126/science.aat6766

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature Neuroscience*, *15*(10), 1439–1444.   doi: 10.1038/nn.3203

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. doi: 10.1109/tpami.2013.50

Bhattarai, B., Lee, J. W., & Jung, M. W. (2019). Distinct effects of reward and navigation history on hippocampal forward and reverse replays. *Proceedings of the National Academy of Sciences*, 201912533.   doi: 10.1073/pnas.1912533117

Bird, C. M. (2020). How do we remember events? *Current Opinion in Behavioral Sciences*, *32*, 120–125. doi: 10.1016/j.cobeha.2020.01.020

Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, *361*(6407), 31–39. doi: 10.1038/361031a0

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*(1), 15958. doi: 10.1038/ncomms15958

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. doi: 10.1038/nn.4573

Bottini, R., & Doeller, C. F. (2020). Knowledge across reference frames: Cognitive maps and image spaces. *Trends in Cognitive Sciences*, *24*(8), 606 - 619. doi: https://doi.org/10.1016/j.tics.2020.05.008

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408 - 422. doi: https://doi.org/10.1016/j.tics.2019.02.006

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*. doi: 10.1016/j.neuron.2020.06.014

Bragin, A., Engel, J., Wilson, C. L., Fried, I., & Buzsáki, G. (1999). High-frequency oscillations in human brain. *Hippocampus*, *9*(2), 137–142. doi: 10.1002/(sici)1098-1063(1999)9:2<137::aid-hipo5>3.0.co;2-0

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*(4), 323. doi: 10.1037/h0058944

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

Brunec, I. K., Moscovitch, M., & Barense, M. D. (2018). Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences*, *22*(7), 637–650. doi: 10.1016/j.tics.2018.03.013

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology*, *61*(1), 27-48. doi: 10.1146/annurev.psych.60.110707.163508

Buhry, L., Azizi, A. H., & Cheng, S. (2011). Reactivation, replay, and preplay: How it might all fit together. *Neural Plasticity*, 1–11. doi: 10.1155/2011/203462

Bush, D., Barry, C., Manson, D., & Burgess, N. (2015). Using grid cells for navigation. *Neuron*, *87*(3), 507 - 520. doi: https://doi.org/10.1016/j.neuron.2015.07.006

Buzsáki, G. (1989). Two-stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience*, *31*(3), 551–570. doi: 10.1016/0306-4522(89)90423-5

Buzsáki, G. (2015). Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, *25*(10), 1073–1188. doi: 10.1002/hipo.22488

Cairney, S. A., Lindsay, S., Sobczak, J. M., Paller, K. A., & Gaskell, M. G. (2016). The benefits of targeted memory reactivation for consolidation in sleep are contingent on memory accuracy and direct cue-memory associations. *Sleep*, *39*(5), 1139-1150. doi: 10.5665/sleep.5772

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

Carr, M. F., Jadhav, S. P., & Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, *14*(2), 147–

1293   153.   doi: 10.1038/nn.2732

1294 Caselles-Dupré, H., Garcia-Ortiz, M., & Filliat, D. (2018). Continual state representation learning for

1295   reinforcement learning using generative replay. *arXiv e-prints*, arXiv:1810.03880.

1296 Caselles-Dupré, H., Garcia-Ortiz, M., & Filliat, D. (2019). S-TRIGGER: Continual State Representa-

1297   tion Learning via Self-Triggered Generative Replay. *arXiv e-prints*, arXiv:1902.09434.

1298 Cazé, R., Khamassi, M., Aubin, L., & Girard, B. (2018). Hippocampal replays under the

1299   scrutiny of reinforcement learning models. *Journal of Neurophysiology*, *120*(6), 2877–2896.

1300   doi: 10.1152/jn.00145.2018

1301 Cheng, S., & Frank, L. M. (2008). New experiences enhance coordinated neural activity in the

1302   hippocampus. *Neuron*, *57*(2), 303 - 313.   doi: 10.1016/j.neuron.2007.11.035

1303 Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories

1304   are constructed from experience. *Hippocampus*, *29*(3), 162–183.   doi: 10.1002/hipo.23074

1305 Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. MIT Press.

1306 Constantinescu, A. O., OReilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge

1307   in humans with a gridlike code. *Science*, *352*(6292), 1464–1468.   doi: 10.1126/science.aaf0941

1308 Csicsvari, J., O'Neill, J., Allen, K., & Senior, T. (2007). Place-selective firing contributes to the

1309   reverse-order reactivation of CA1 pyramidal cells during sharp waves in open-field exploration.

1310   *European Journal of Neuroscience*, *26*(3), 704–716.   doi: 10.1111/j.1460-9568.2007.05684.x

1311 Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience.

1312   *Neuron*, *63*(4), 497–507.   doi: 10.1016/j.neuron.2009.07.027

1313 Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and

1314   dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

1315   doi: 10.1038/nn1560

1316 Dayan, P. (1993). Improving generalization for temporal difference learning: The successor represen-

1317   tation. *Neural Computation*, *5*(4), 613–624.   doi: 10.1162/neco.1993.5.4.613

1318 Deng, X., Chen, S., Sosa, M., Karlsson, M. P., Wei, X.-X., & Frank, L. M. (2020). A variable clock

1319   underlies internally generated hippocampal sequences. *bioRxiv*.   doi: 10.1101/2020.04.10.035980

1320 Denovellis, E. L., Gillespie, A. K., Coulter, M. E., Sosa, M., Chung, J. E., Eden, U. T., &

1321   Frank, L. M. (2020). Hippocampal replay of experience at real-world speeds. *bioRxiv*.

1322   doi: 10.1101/2020.10.20.347708

1323 Deuker, L., Olligs, J., Fell, J., Kranz, T. A., Mormann, F., Montag, C., . . . Axmacher, N. (2013).

1324   Memory consolidation by replay of stimulus-specific neural activity. *Journal of Neuroscience*,

1325   *33*(49), 19373–19383.   doi: 10.1523/jneurosci.0414-13.2013

1326 Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples.

1327   *Nature Neuroscience*, *10*(10), 1241–1242.   doi: 10.1038/nn1961

1328 Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*,

1329   *11*(2), 114–126.   doi: 10.1038/nrn2762

1330 Diekelmann, S., Born, J., & Wagner, U. (2010). Sleep enhances false memories depend-

1331   ing on general memory performance. *Behavioural Brain Research*, *208*(2), 425 - 429.

1332   doi: https://doi.org/10.1016/j.bbr.2009.12.021

1333 Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312 - 325.

1334   doi: 10.1016/j.neuron.2013.09.007

1335 Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral

1336   cortex? *Neural Networks*, *12*(7), 961-974.   doi: 10.1016/S0893-6080(99)00046-5

Dragoi, G., & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature*, *469*(7330), 397–401.   doi: 10.1038/nature09633

Dragoi, G., & Tonegawa, S.   (2013).   Distinct preplay of multiple novel spatial experiences in the rat.   *Proceedings of the National Academy of Sciences*, *110*(22), 9100–9105.   doi: 10.1073/pnas.1306031110

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General*, *136*(2), 169.   doi: 10.1037/0096-3445.136.2.169

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current Opinion in Behavioral Sciences*, *17*, 141-146.   doi: 10.1016/j.cobeha.2017.08.003

Duncan, K. D., & Shohamy, D. (2016). Memory states influence value-based decisions. *Journal of Experimental Psychology: General*, *145*(11), 1420.

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

Ego-Stengel, V., & Wilson, M. A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, *20*(1), 1–10.   doi: 10.1002/hipo.20707

Eichenbaum, H. (2015). Does the hippocampus preplay memories? *Nature Neuroscience*, *18*(12), 1701–1702.   doi: 10.1038/nn.4180

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

Eldar, E., Lièvre, G., Dayan, P., & Dolan, R. J. (2020). The roles of online and offline replay in planning. *eLife*, *9*.   doi: 10.7554/elife.56911

Epstein, R. A., Patai, E. Z., Julian, J. B., & Spiers, H. J. (2017). The cognitive map in humans: spatial navigation and beyond. *Nature Neuroscience*, *20*(11), 1504–1513.   doi: 10.1038/nn.4656

Erdem, U. M., & Hasselmo, M. (2012). A goal-directed spatial navigation model using forward trajectory planning based on grid cells. *European Journal of Neuroscience*, *35*(6), 916-931. doi: 10.1111/j.1460-9568.2012.08015.x

Eschenko, O., Ramadan, W., Mölle, M., Born, J., & Sara, S. J. (2008). Sustained increase in hippocampal sharp-wave ripple activity during slow-wave sleep after learning. *Learning & Memory*, *15*(4), 222-228.   doi: 10.1101/lm.726008

Euston, D. R., Tatsuno, M., & McNaughton, B. L.   (2007).   Fast-forward playback of recent memory sequences in prefrontal cortex during sleep.   *Science*, *318*(5853), 1147–1150. doi: 10.1126/science.1148979

Evans, T., & Burgess, N. (2019). Coordinated hippocampal-entorhinal replay as structural inference. *Advances in Neural Information Processing Systems*, 1729–1741.

Farooq, U., Sibille, J., Liu, K., & Dragoi, G. (2019). Strengthened temporal coordination within pre-existing sequential cell assemblies supports trajectory replay. *Neuron*, *103*(4), 719–733.e7. doi: 10.1016/j.neuron.2019.05.040

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*(1), 11066.   doi: 10.1038/ncomms11066

Favila, S. E., Lee, H., & Kuhl, B. A. (2020). Transforming the concept of memory reactivation. *Trends*

in Neurosciences, *43*(12), 939–950. doi: 10.1016/j.tins.2020.09.006

Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., & Dabney, W. (2020). Revisiting Fundamentals of Experience Replay. *arXiv e-prints*, arXiv:2007.06700.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322. doi: 10.1073/pnas.1800755115

Foster, D. J. (2017). Replay comes of age. *Annual Review of Neuroscience*, *40*(1), 581-602. doi: 10.1146/annurev-neuro-072116-031538

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

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

French, R. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. doi: 10.1016/s1364-6613(99)01294-2

Fyhn, M., Hafting, T., Treves, A., Moser, M.-B., & Moser, E. I. (2007). Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, *446*(7132), 190–194. doi: 10.1038/nature05601

Gagne, C., Dayan, P., & Bishop, S. J. (2018). When planning to survive goes wrong: predicting the future and replaying the past in anxiety and PTSD. *Current Opinion in Behavioral Sciences*, *24*, 89–95. doi: 10.1016/j.cobeha.2018.03.013

Gardner, M. P. H., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1891), 20181645. doi: 10.1098/rspb.2018.1645

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, *6*. doi: 10.7554/elife.17086

Genzel, L., Dragoi, G., Frank, L., Ganguly, K., de la Prida, L., Pfeiffer, B., & Robertson, E. (2020). A consensus statement: defining terms for reactivation analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1799), 20200001. doi: 10.1098/rstb.2020.0001

Gerrard, J. L., Kudrimoti, H., McNaughton, B. L., & Barnes, C. A. (2001). Reactivation of hippocampal ensemble activity patterns in the aging rat. *Behavioral Neuroscience*, *115*(6), 1180–1192. doi: 10.1037/0735-7044.115.6.1180

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*(1), 101–128. doi: 10.1146/annurev-psych-122414-033625

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology*, *20*(2), 251-256. doi: 10.1016/j.conb.2010.02.008

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLOS Computational Biology*, *10*(11), 1-13. doi: 10.1371/journal.pcbi.1003939

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience*, *12*(10), 1222–1223. doi: 10.1038/nn.2384

Girardeau, G., Inema, I., & Buzsáki, G. (2017). Reactivations of emotional memory in the hippocampus–amygdala system during sleep. *Nature Neuroscience*, *20*(11), 1634–1642. doi: 10.1038/nn.4637

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

Gridchyn, I., Schoenenberger, P., O'Neill, J., & Csicsvari, J. (2020). Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron*. doi: 10.1016/j.neuron.2020.01.021

Grieves, R. M., & Jeffery, K. J. (2017). The representation of space in the brain. *Behavioural Processes*, *135*, 113 - 131. doi: 10.1016/j.beproc.2016.12.012

Gruber, M. J., Ritchey, M., Wang, S.-F., Doss, M. K., & Ranganath, C. (2016). Post-learning hippocampal dynamics promote preferential retention of rewarding events. *Neuron*, *89*(5), 1110– 1120. doi: 10.1016/j.neuron.2016.01.017

Gulati, T., Guo, L., Ramanathan, D. S., Bodepudi, A., & Ganguly, K. (2017). Neural reactivations during sleep determine network credit assignment. *Nature Neuroscience*, *20*(9), 1277–1284. doi: 10.1038/nn.4601

Guo, Z. D., Pires, B. A., Piot, B., Grill, J.-B., Altché, F., Munos, R., & Azar, M. G. (2020). Bootstrap latent-predictive representations for multitask reinforcement learning. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 3875– 3886). PMLR.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695 - 705. doi: 10.1016/j.neuron.2010.01.034

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806. doi: 10.1038/nature03721

Haga, T., & Fukai, T. (2018). Recurrent network model for learning goal-directed sequences through reverse replay. *eLife*, *7*. doi: 10.7554/elife.34171

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: the role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111-120. doi: 10.1016/j.tics.2013.01.001

Harvey, C. D., Coen, P., & Tank, D. W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, *484*(7392), 62–68. doi: 10.1038/nature10918

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. doi: 10.1016/j.neuron.2017.06.011

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299 - 306. doi: 10.1016/j.tics.2007.05.001

Helfrich, R. F., Lendner, J. D., Mander, B. A., Guillen, H., Paff, M., Mnatsakanyan, L., ... Knight, R. T. (2019). Bidirectional prefrontal-hippocampal dynamics organize information transfer during sleep in humans. *Nature Communications*, *10*(1). doi: 10.1038/s41467-019-11444-x

Heller, A. S., & Bagot, R. C. (2020). Is hippocampal replay a mechanism for anxiety and depression? *JAMA Psychiatry*. doi: 10.1001/jamapsychiatry.2019.4788

Herszage, J., & Censor, N. (2018). Modulation of learning and memory: A shared framework for interference and generalization. *Neuroscience*, *392*, 270 - 280. doi: 10.1016/j.neuroscience.2018.08.006

Hoffman, K. L., & McNaughton, B. L. (2002). Coordinated reactivation of distributed memory traces in primate neocortex. *Science*, *297*(5589), 2070–2073. doi: 10.1126/science.1073538

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., van Hasselt, H., & Silver, D. (2018). Distributed prioritized experience replay. *arXiv e-prints*, arXiv:1803.00933.

Høydal, Ø. A., Skytøen, E. R., Andersson, S. O., Moser, M.-B., & Moser, E. I. (2019). Object-vector coding in the medial entorhinal cortex. *Nature*, *568*(7752), 400–404. doi: 10.1038/s41586-019-

1077-7

Igloi, K., Gaggioni, G., Sterpenich, V., & Schwartz, S. (2015). A nap to recap or how reward regulates hippocampal-prefrontal memory networks during daytime sleep in humans. *eLife*, *4*, e07903. doi: 10.7554/eLife.07903

Jackson, J. C., Johnson, A., & Redish, A. D. (2006). Hippocampal sharp waves and reactivation during awake states depend on repeated sequential experience. *Journal of Neuroscience*, *26*(48), 12415–12426. doi: 10.1523/jneurosci.4118-06.2006

Jadhav, S. P., Kemere, C., German, P. W., & Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science*, *336*(6087), 1454–1458. doi: 10.1126/science.1217230

Jadhav, S. P., Rothschild, G., Roumis, D. K., & Frank, L. M. (2016). Coordinated excitation and inhibition of prefrontal ensembles during awake hippocampal sharp-wave ripple events. *Neuron*, *90*(1), 113–127. doi: 10.1016/j.neuron.2016.02.010

Jafarpour, A., Penny, W., Barnes, G., Knight, R. T., & Duzel, E. (2017). Working memory replay prioritizes weakly attended events. *eNeuro*, *4*(4), 1–11. doi: 10.1523/eneuro.0171-17.2017

Ji, D., & Wilson, M. A. (2006). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, *10*(1), 100–107. doi: 10.1038/nn1825

Johnson, A., & 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. doi: 10.1523/jneurosci.3761-07.2007

Joo, H. R., & Frank, L. M. (2018). The hippocampal sharp wave–ripple in memory retrieval for immediate use and consolidation. *Nature Reviews Neuroscience*, *19*(12), 744–757. doi: 10.1038/s41583-018-0077-1

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2019.07.012

Kaefer, K., Nardin, M., Blahna, K., & Csicsvari, J. (2020). Replay of behavioral sequences in the medial prefrontal cortex during rule switching. *Neuron*. doi: 10.1016/j.neuron.2020.01.015

Kaplan, R., Campo, A. T., Bush, D., King, J., Principe, A., Koster, R., . . . Friston, K. J. (2020). Human hippocampal theta oscillations reflect sequential dependencies during spatial planning. *Cognitive Neuroscience*, *11*(3), 122-131. (PMID: 31617790) doi: 10.1080/17588928.2019.1676711

Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, *40*(5), 256 - 259. doi: 10.1016/j.tins.2017.03.002

Karlsson, M. P., & Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nature Neuroscience*, *12*(7), 913–918. doi: 10.1038/nn.2344

Kay, K., Chung, J. E., Sosa, M., Schor, J. S., Karlsson, M. P., Larkin, M. C., . . . Frank, L. M. (2020). Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell*. doi: 10.1016/j.cell.2020.01.014

King, C., Henze, D. A., Leinekugel, X., & Buzsáki, G. (1999). Hebbian modification of a hippocampal population pattern in the rat. *The Journal of Physiology*, *521*(1), 159–167. doi: 10.1111/j.1469-7793.1999.00159.x

Kudrimoti, H. S., Barnes, C. A., & McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and eeg dynamics. *The Journal of Neuroscience*, *19*(10), 4090–4101. doi: 10.1523/jneurosci.19-10-04090.1999

Kuhl, B. A., Shah, A. T., DuBrow, S., & Wagner, A. D. (2010). Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nature Neuroscience*, *13*(4), 501–506.

doi: 10.1038/nn.2498

Kumaran, D. (2012). What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience*, *6*, 157. doi: 10.3389/fnhum.2012.00157

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. doi: 10.1016/j.tics.2016.05.004

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. doi: 10.1037/a0028681

Kuriyama, K., Stickgold, R., & Walker, M. P. (2004). Sleep-dependent learning and motor-skill complexity. *Learning & Memory*, *11*(6), 705-713. doi: 10.1101/lm.76304

Kurth-Nelson, Z., Economides, M., Dolan, R. J., & Dayan, P. (2016). Fast sequences of non-spatial state representations in humans. *Neuron*, *91*(1), 194–204. doi: 10.1016/j.neuron.2016.05.028

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., Joosten, R. N. J. M. A., McNaughton, B. L., & Pennartz, C. M. A. (2008). Preferential reactivation of motivationally relevant information in the ventral striatum. *Journal of Neuroscience*, *28*(25), 6372–6382. doi: 10.1523/jneurosci.1054-08.2008

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, *7*(8), e1000173. doi: 10.1371/journal.pbio.1000173

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194. doi: 10.1016/s0896-6273(02)01096-6

Lee, S. W., O'Doherty, J. P., & Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. *PLOS Biology*, *13*(4), 1-36. doi: 10.1371/journal.pbio.1002137

Leibold, C. (2020). A model for navigation in unknown environments based on a reservoir of hippocampal sequences. *Neural Networks*, *124*, 328–342. doi: 10.1016/j.neunet.2020.01.014

Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: The third way. In *Proceedings of the 20th international conference on neural information processing systems* (pp. 889–896). Red Hook, NY, USA: Curran Associates Inc.

Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*(2), 451 - 463. doi: 10.1016/j.neuron.2016.12.040

Lewis, P. A., & Bendor, D. (2019). How targeted memory reactivation promotes the selective strengthening of memories in sleep. *Current Biology*, *29*(18), R906–R912. doi: 10.1016/j.cub.2019.08.019

Lewis, P. A., Knoblich, G., & Poe, G. (2018). How memory replay in sleep boosts creative problem-solving. *Trends in Cognitive Sciences*, *22*(6), 491–503. doi: 10.1016/j.tics.2018.03.009

Lin, L. J. (1991). Programming robots using reinforcement learning and teaching. In *Association for the advancement of artificial intelligence* (pp. 781–786).

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*(3-4), 293–321. doi: 10.1007/bf00992699

Lin, L.-J. (1993). Reinforcement learning for robots using neural networks. *Carnegie Mellon University*.

Liu, B., Ye, X., Gao, Y., Dong, X., Wang, X., & Liu, B. (2019). Forward-looking imaginative planning framework combined with prioritized-replay double DQN. In *2019 5th international conference on control, automation and robotics (iccar)* (p. 336-341). doi: 10.1109/ICCAR.2019.8813352

Liu, Y., Dolan, R., Kurth-Nelson, Z., & Behrens, T. E. (2020). Measuring sequences of representations with temporally delayed linear modelling. *bioRxiv*. doi: 10.1101/2020.04.30.066407

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652. doi: 10.1016/j.cell.2019.06.012

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156. doi: 10.1016/s0896-6273(01)00186-6

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737 - 749. doi: 10.1016/j.neuron.2011.07.012

Magee, J. C. (1997). A synaptically controlled, associative signal for hebbian plasticity in hippocampal neurons. *Science*, *275*(5297), 209–213. doi: 10.1126/science.275.5297.209

Mahadevan, S., & Maggioni, M. (2007). Proto-value functions: A laplacian framework for learning representation and control in markov decision processes. *Journal of Machine Learning Research*, *8*(74), 2169-2231.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *262*(841), 23–81. doi: 10.1098/rstb.1971.0078

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

Maurer, A. P., & Nadel, L. (2021). The continuity of context: A role for the hippocampus. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2020.12.007

McClelland, J. L., McNaughton, B. L., & 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), 419–457. doi: 10.1037/0033-295x.102.3.419

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, *24*, 109–165. doi: 10.1016/s0079-7421(08)60536-8

McDevitt, E. A., Duggan, K. A., & Mednick, S. C. (2015). REM sleep rescues learning from interference. *Neurobiology of Learning and Memory*, *122*, 51–62. doi: 10.1016/j.nlm.2014.11.015

McNamara, C. G., Tejero-Cantero, Á., Trouche, S., Campo-Urriza, N., & Dupret, D. (2014). Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nature Neuroscience*, *17*(12), 1658–1660. doi: 10.1038/nn.3843

Michon, F., Sun, J.-J., Kim, C. Y., Ciliberti, D., & Kloosterman, F. (2019). Post-learning hippocampal replay selectively reinforces spatial memory for highly rewarded locations. *Current Biology*, *29*(9), 1436–1444.e5. doi: 10.1016/j.cub.2019.03.048

Miller, K. J., & Venditto, S. J. C. (2021). Multi-step planning in the brain. *Current Opinion in Behavioral Sciences*, *38*, 29 - 39. doi: 10.1016/j.cobeha.2020.07.003

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*(1), 8-30. doi: 10.1109/JRPROC.1961.287775

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M.

(2013). Playing Atari with Deep Reinforcement Learning. *arXiv e-prints*, arXiv:1312.5602.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236

Mobbs, D., Headley, D. B., Ding, W., & Dayan, P. (2020). Space, time, and fear: Survival computations along defensive circuits. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2019.12.016

Momennejad, I. (2020). Learning structures: Predictive representations, replay, and generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166. doi: 10.1016/j.cobeha.2020.02.017

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *eLife*, *7*, e32548. doi: 10.7554/eLife.32548

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

Monaco, J. D., Rao, G., Roth, E. D., & Knierim, J. J. (2014). Attentive scanning behavior drives one-trial potentiation of hippocampal place fields. *Nature Neuroscience*, *17*(5), 725–731. doi: 10.1038/nn.3687

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, *13*(1), 103–130. doi: 10.1007/bf00993104

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*(1), 69–89. doi: 10.1146/annurev.neuro.31.061307.090723

Muenzinger, K. F., & Fletcher, F. M. (1936). Motivation in learning. vi. escape from electric shock compared with hunger-food tension in the visual discrimination habit. *Journal of Comparative Psychology*, *22*(1), 79–91.

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., & Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *The Journal of Neuroscience*, *19*(21), 9497–9507. doi: 10.1523/jneurosci.19-21-09497.1999

Niethard, N., & Born, J. (2020). A backup of hippocampal spatial code outside the hippocampus? new light on systems memory consolidation. *Neuron*, *106*(2), 204 - 206. doi: doi.org/10.1016/j.neuron.2020.03.034

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. doi: 10.1038/s41593-019-0470-8

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015

Norman, Y., Yeagle, E. M., Khuvis, S., Harel, M., Mehta, A. D., & Malach, R. (2019). Hippocampal sharp-wave ripples linked to visual episodic recollection in humans. *Science*, *365*(6454), eaax1030. doi: 10.1126/science.aax1030

O'Keefe, J., & Conway, D. H. (1978). Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, *31*(4), 573–590. doi: 10.1007/BF00239813

O'Keefe, J., & 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. doi: 10.1016/0006-8993(71)90358-1

O'Keefe, J., & Krupic, J. (2021). Do hippocampal pyramidal cells respond to non-spatial stimuli?

1645   *Physiological Reviews*. doi: 10.1152/physrev.00014.2020

1646 O'Keefe, J., & Nadel, L. (1974). Maps in the brain. *New Scientist*, *62*(903), 749–751.

1647 O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map.* Oxford: Clarendon Press.

1648 Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. (2015). Hip-
1649   pocampal place cells construct reward related sequences through unexplored space. *eLife*, *4*.
1650   doi: 10.7554/elife.06063

1651 Ólafsdóttir, H. F., Bush, D., & Barry, C. (2018). The role of hippocampal replay in memory and
1652   planning. *Current Biology*, *28*(1), R37–R50. doi: 10.1016/j.cub.2017.10.073

1653 Ólafsdóttir, H. F., Carpenter, F., & Barry, C. (2016). Coordinated grid and place cell replay during
1654   rest. *Nature Neuroscience*, *19*(6), 792–794. doi: 10.1038/nn.4291

1655 Ólafsdóttir, H. F., Carpenter, F., & Barry, C. (2017). Task demands predict a dy-
1656   namic switch in the content of awake hippocampal replay. *Neuron*, *96*(4), 925–935.e6.
1657   doi: 10.1016/j.neuron.2017.09.035

1658 O'Neill, J., Boccara, C., Stella, F., Schoenenberger, P., & Csicsvari, J. (2017). Superficial layers of the
1659   medial entorhinal cortex replay independently of the hippocampus. *Science*, *355*(6321), 184–188.
1660   doi: 10.1126/science.aag2787

1661 O'Neill, J., Pleydell-Bouverie, B., Dupret, D., & Csicsvari, J. (2010). Play it again: Re-
1662   activation of waking experience and memory. *Trends in Neurosciences*, *33*(5), 220–229.
1663   doi: 10.1016/j.tins.2010.01.006

1664 O'Neill, J., Senior, T. J., Allen, K., Huxter, J. R., & Csicsvari, J. (2008). Reactivation of experience-
1665   dependent cell assembly patterns in the hippocampus. *Nature Neuroscience*, *11*(2), 209–215.
1666   doi: 10.1038/nn2037

1667 O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning
1668   systems. *Cognitive Science*, *38*(6), 1229-1248. doi: 10.1111/j.1551-6709.2011.01214.x

1669 O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall:
1670   Avoiding a trade-off. *Hippocampus*, *4*(6), 661-682. doi: 10.1002/hipo.450040605

1671 Oudiette, D., & Paller, K. A. (2013). Upgrading the sleeping brain with targeted memory reactivation.
1672   *Trends in Cognitive Sciences*, *17*(3), 142 - 149. doi: 10.1016/j.tics.2013.01.006

1673 Papale, A. E., Zielinski, M. C., Frank, L. M., Jadhav, S. P., & Redish, A. D. (2016). Interplay between
1674   hippocampal sharp-wave-ripple events and vicarious trial and error behaviors in decision making.
1675   *Neuron*, *92*(5), 975–982. doi: 10.1016/j.neuron.2016.10.028

1676 Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with
1677   neural networks: A review. *Neural Networks*, *113*, 54 - 71. doi: 10.1016/j.neunet.2019.01.012

1678 Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on
1679   the activity of these cells during subsequent sleep episodes. *The Journal of Neuroscience*, *9*(8),
1680   2907–2918. doi: 10.1523/jneurosci.09-08-02907.1989

1681 Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2020). Structuring knowledge with cogni-
1682   tive maps and cognitive graphs. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2020.10.004

1683 Penagos, H., Varela, C., & Wilson, M. A. (2017). Oscillations, neural computations and learning during
1684   wake and sleep. *Current Opinion in Neurobiology*, *44*, 193–201. doi: 10.1016/j.conb.2017.05.009

1685 Peng, J., & Williams, R. J. (1993). Efficient learning and planning within the dyna framework.
1686   *Adaptive Behavior*, *1*(4), 437–454. doi: 10.1177/105971239300100403

1687 Pennartz, C. M. A. (2004). The ventral striatum in off-line processing: Ensemble reactivation during
1688   sleep and modulation by hippocampal ripples. *Journal of Neuroscience*, *24*(29), 6446–6456.

doi: 10.1523/jneurosci.0575-04.2004

Peters, K. R., Smith, V., & Smith, C. T. (2007). Changes in sleep architecture following motor learning depend on initial skill level. *Journal of Cognitive Neuroscience*, *19*(5), 817–829. doi: 10.1162/jocn.2007.19.5.817

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

Pezzulo, G., Donnarumma, F., Maisto, D., & Stoianov, I. (2019). Planning at decision time and in the background during spatial navigation. *Current Opinion in Behavioral Sciences*, *29*, 69–76. doi: 10.1016/j.cobeha.2019.04.009

Pezzulo, G., Kemere, C., & van der Meer, M. A. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Annals of the New York Academy of Sciences*, *1396*(1), 144–165. doi: 10.1111/nyas.13329

Pezzulo, G., van der Meer, M. A., Lansink, C. S., & Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, *18*(12), 647–657. doi: 10.1016/j.tics.2014.06.011

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

Qin, Y.-L., Mcnaughton, B. L., Skaggs, W. E., & Barnes, C. A. (1997). Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1360), 1525–1533. doi: 10.1098/rstb.1997.0139

Ramanathan, D. S., Gulati, T., & Ganguly, K. (2015). Sleep-dependent reactivation of ensembles in motor cortex promotes skill consolidation. *PLOS Biology*, *13*(9), 1–25. doi: 10.1371/journal.pbio.1002263

Ranck, J. B. J. (1984). Head direction cells in the deep layer of dorsal presubiculum in freely moving rats. *Society of Neuroscience*, *10*, 599.

Rasch, B., & Born, J. (2007). Maintaining memories by reactivation. *Current Opinion in Neurobiology*, *17*(6), 698–703. doi: 10.1016/j.conb.2007.11.007

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*(2), 285–308. doi: [10.1037/0033-295x.97.2.285](10.1037/0033-295x.97.2.285)

Redish, A. D. (1999). *Beyond the cognitive map: From place cells to episodic memory.* MIT Press.

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*(3), 147–159. doi: [10.1038/nrn.2015.30](10.1038/nrn.2015.30)

Redish, A. D., & Touretzky, D. S. (1998). The role of the hippocampus in solving the morris water maze. *Neural Computation*, *10*(1), 73–111. doi: [10.1162/089976698300017908](10.1162/089976698300017908)

Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences*, *21*(12), 962–980. doi: 10.1016/j.tics.2017.08.005

Roscow, E. L., Jones, M. W., & Lepora, N. F. (2019). Behavioural and computational evidence for memory consolidation biased by reward-prediction errors. *bioRxiv*. doi: [10.1101/716290](10.1101/716290)

Rothschild, G., Eban, E., & Frank, L. M. (2016). A cortical–hippocampal–cortical loop of information processing during memory consolidation. *Nature Neuroscience*, *20*(2), 251–259. doi: [10.1038/nn.4457](10.1038/nn.4457)

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create

event boundaries in memory. *Cognition*, *203*, 104269. doi: 10.1016/j.cognition.2020.104269

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

Schafer, M., & Schiller, D. (2018). Navigating social space. *Neuron*, *100*(2), 476–489. doi: 10.1016/j.neuron.2018.10.006

Schapiro, A. C., McDevitt, E. A., Chen, L., Norman, K. A., Mednick, S. C., & Rogers, T. T. (2017). Sleep benefits memory for semantic category structure while preserving exemplar-specific information. *Scientific Reports*, *7*(1), 14869. doi: 10.1038/s41598-017-12884-5

Schapiro, A. C., McDevitt, E. A., Rogers, T. T., Mednick, S. C., & Norman, K. A. (2018). Human hippocampal replay during rest prioritizes weakly learned information and predicts memory performance. *Nature Communications*, *9*(1). doi: 10.1038/s41467-018-06213-1

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

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. doi: 10.1098/rstb.2016.0049

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv e-prints*, arXiv:1511.05952.

Schmidt, B., Duin, A. A., & Redish, A. D. (2019). Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision making. *Journal of Neurophysiology*, *121*(6), 1981–2000. doi: 10.1152/jn.00793.2018

Schmidt, C., Peigneux, P., Muto, V., Schenkel, M., Knoblauch, V., Münch, M., . . . Cajochen, C. (2006). Encoding difficulty promotes postlearning changes in sleep spindle activity during napping. *Journal of Neuroscience*, *26*(35), 8976–8982. doi: 10.1523/JNEUROSCI.2464-06.2006

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402 - 1412. doi: 10.1016/j.neuron.2016.08.019

Schuck, N. W., & Niv, Y. (2019). Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), eaaw5181. doi: 10.1126/science.aaw5181

Schuck, N. W., Wilson, R., & Niv, Y. (2018). A state representation for reinforcement learning and decision-making in the orbitofrontal cortex. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making* (1st ed., p. 259-278). Academic Press. doi: 10.1016/B978-0-12-812098-9.00012-7

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11–21. doi: 10.1136/jnnp.20.1.11

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., . . . Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, *20*(5), 735–742. doi: 10.1038/nn.4538

Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 2990–2999). Curran Associates, Inc.

Shin, J. D., Tang, W., & Jadhav, S. P. (2019). Dynamics of awake hippocampal-prefrontal re-

play for spatial learning and memory-guided decision making. *Neuron*, *104*(6), 1110–1125. doi: 10.1016/j.neuron.2019.09.012

Shin, Y. S., & DuBrow, S. (2020). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science*. doi: 10.1111/tops.12505

Silva, D., Feng, T., & Foster, D. J. (2015). Trajectory events across hippocampal place cells require previous experience. *Nature Neuroscience*, *18*(12), 1772–1779. doi: 10.1038/nn.4151

Singer, A. C., Carr, M. F., Karlsson, M. P., & Frank, L. M. (2013). Hippocampal swr activity predicts correct decisions during the initial learning of an alternation task. *Neuron*, *77*(6), 1163–1173. doi: 10.1016/j.neuron.2013.01.027

Singer, A. C., & Frank, L. M. (2009). Rewarded outcomes enhance reactivation of experience in the hippocampus. *Neuron*, *64*(6), 910–921. doi: 10.1016/j.neuron.2009.11.016

Sio, U. N., Monaghan, P., & Ormerod, T. (2013). Sleep on it, but only if it is difficult: Effects of sleep on problem solving. *Memory & Cognition*, *41*(2), 159–166. doi: 10.3758/s13421-012-0256-7

Skaggs, W. E., & McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, *271*(5257), 1870–1873. doi: 10.1126/science.271.5257.1870

Spiers, H. J. (2020). The hippocampal cognitive map: One space or many? *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2019.12.013

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231.

Staba, R. J., Wilson, C. L., Bragin, A., Fried, I., & Engel, J. (2002). Quantitative analysis of high-frequency oscillations (80–500 hz) recorded in human epileptic hippocampus and entorhinal cortex. *Journal of Neurophysiology*, *88*(4), 1743–1752. doi: 10.1152/jn.2002.88.4.1743

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. doi: 10.1038/nn.4650

Staresina, B. P., Alink, A., Kriegeskorte, N., & Henson, R. N. (2013). Awake reactivation predicts memory in humans. *Proceedings of the National Academy of Sciences*, *110*(52), 21159–21164. doi: 10.1073/pnas.1311989110

Staresina, B. P., Bergmann, T. O., Bonnefond, M., van der Meij, R., Jensen, O., Deuker, L., . . . Fell, J. (2015). Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nature Neuroscience*, *18*(11), 1679–1686. doi: 10.1038/nn.4119

Steiner, A., & Redish, A. D. (2012). The road not taken: Neural correlates of decision making in orbitofrontal cortex. *Frontiers in Neuroscience*, *6*, 131. doi: 10.3389/fnins.2012.00131

Stella, F., Baracskay, P., O'Neill, J., & Csicsvari, J. (2019). Hippocampal reactivation of random trajectories resembling brownian diffusion. *Neuron*, *102*(2), 450–461. doi: 10.1016/j.neuron.2019.01.052

Stoianov, I., Maisto, D., & Pezzulo, G. (2020). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *bioRxiv*. doi: 10.1101/2020.01.16.908889

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nature Neuroscience*. doi: 10.1038/s41593-020-0614-x

Sutherland, G. R., & McNaughton, B. (2000). Memory trace reactivation in hippocampal and neocortical neuronal ensembles. *Current Opinion in Neurobiology*, *10*(2), 180 - 186. doi: 10.1016/S0959-4388(00)00079-9

1821 Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approxi-
1822   mating dynamic programming. In B. Porter & R. Mooney (Eds.), *Machine learning proceedings*
1823   *1990* (pp. 216–224). San Francisco (CA): Morgan Kaufmann. doi: 10.1016/B978-1-55860-141-
1824   3.50030-4

1825 Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM*
1826   *SIGART Bulletin*, *2*(4), 160–163. doi: 10.1145/122344.122377

1827 Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. *MIT Press*.

1828 Sutton, R. S., Szepesvári, C., Geramifard, A., & Bowling, M. (2012). Dyna-style planning with linear
1829   function approximation and prioritized sweeping. *arXiv e-prints*, *abs/1206.3285*.

1830 Swanson, R. A., Levenstein, D., McClain, K., Tingley, D., & Buzsáki, G. (2020). Variable speci-
1831   ficity of memory trace reactivation during hippocampal sharp wave ripples. *Current Opinion*
1832   *in Behavioral Sciences*, *32*, 126–135. (Understanding memory: Which level of analysis?)
1833   doi: 10.1016/j.cobeha.2020.02.008

1834 Talamini, L. M., Nieuwenhuis, I. L. C., Takashima, A., & Jensen, O. (2008). Sleep directly following
1835   learning benefits consolidation of spatial associative memory. *Learning & Memory*, *15*(4), 233–
1836   237. doi: 10.1101/lm.771608

1837 Tambini, A., Berners-Lee, A., & Davachi, L. (2017). Brief targeted memory reactivation during the
1838   awake state enhances memory stability and benefits the weakest memories. *Scientific Reports*,
1839   *7*(1). doi: 10.1038/s41598-017-15608-x

1840 Tambini, A., & Davachi, L. (2013). Persistence of hippocampal multivoxel patterns into postencoding
1841   rest is related to memory. *Proceedings of the National Academy of Sciences*, *110*(48), 19591–
1842   19596. doi: 10.1073/pnas.1308499110

1843 Tambini, A., & Davachi, L. (2019). Awake reactivation of prior experiences consolidates memories and
1844   biases cognition. *Trends in Cognitive Sciences*, *23*(10), 876–890. doi: 10.1016/j.tics.2019.07.008

1845 Tambini, A., Ketz, N., & Davachi, L. (2010). Enhanced brain correlations during rest are related to
1846   memory for recent experiences. *Neuron*, *65*(2), 280–290. doi: 10.1016/j.neuron.2010.01.001

1847 Tang, W., & Jadhav, S. P. (2019). Sharp-wave ripples as a signature of hippocampal-prefrontal
1848   reactivation for memory during sleep and waking states. *Neurobiology of Learning and Memory*,
1849   *160*, 11–20. doi: 10.1016/j.nlm.2018.01.002

1850 Tang, W., Shin, J. D., Frank, L. M., & Jadhav, S. P. (2017). Hippocampal-prefrontal reactivation
1851   during learning is stronger in awake compared with sleep states. *Journal of Neuroscience*, *37*(49),
1852   11789–11805. doi: 10.1523/JNEUROSCI.2291-17.2017

1853 Tang, W., Shin, J. D., & Jadhav, S. P. (2021). Multiple time-scales of decision making in the
1854   hippocampus and prefrontal cortex. *eLife*, *10*, e66227. doi: 10.7554/eLife.66227

1855 Tang, Y., Nyengaard, J. R., De Groot, D. M., & Gundersen, H. J. G. (2001). Total regional
1856   and global number of synapses in the human brain neocortex. *Synapse*, *41*(3), 258-273.
1857   doi: 10.1002/syn.1083

1858 Taube, J., Muller, R., & Ranck, J. (1990). Head-direction cells recorded from the postsubiculum in
1859   freely moving rats. ii. effects of environmental manipulations. *Journal of Neuroscience*, *10*(2),
1860   436–447. doi: 10.1523/JNEUROSCI.10-02-00436.1990

1861 Tolman, E. C. (1926). A behavioristic theory of ideas. *Psychological Review*, *33*(5), 352.
1862   doi: 10.1037/h0070532

1863 Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological Review*, *45*(1),
1864   1–41. doi: 10.1037/h0062733

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. doi: 10.1037/h0061626

Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, *96*(1), 228–241. doi: 10.1016/j.neuron.2017.09.005

Trettel, S. G., Trimper, J. B., Hwaun, E., Fiete, I. R., & Colgin, L. L. (2019). Grid cell co-activity patterns during sleep reflect spatial overlap of grid fields during active behaviors. *Nature Neuroscience*, *22*(4), 609–617. doi: 10.1038/s41593-019-0359-6

van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, *11*(1), 4069. doi: 10.1038/s41467-020-17866-2

van de Ven, G. M., & Tolias, A. S. (2018). Generative replay with feedback connections as a general strategy for continual learning. *arXiv e-prints*, arXiv:1809.10635.

van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K., & Dupret, D. (2016). Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron*, *92*(5), 968–974. doi: 10.1016/j.neuron.2016.10.020

van Seijen, H., & Sutton, R. (2015). A deeper look at planning as learning from replay. In *International conference on machine learning* (pp. 2314–2322).

Vaz, A. P., Inati, S. K., Brunel, N., & Zaghloul, K. A. (2019). Coupled ripple oscillations between the medial temporal lobe and neocortex retrieve human memory. *Science*, *363*(6430), 975–978. doi: 10.1126/science.aau8956

Vaz, A. P., Wittig, J. H., Inati, S. K., & Zaghloul, K. A. (2020). Replay of cortical spiking sequences during human memory retrieval. *Science*, *367*(6482), 1131–1134. doi: 10.1126/science.aba0672

Vértes, E., & Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 13714–13724). Curran Associates, Inc.

Wang, S., Feng, S. F., & Bornstein, A. (2020). Mixing memory and desire: How memory reactivation supports deliberative decision-making. *PsyArXiv*. doi: 10.31234/osf.io/5vksj

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249 - 1263.e23. doi: 10.1016/j.cell.2020.10.024

Wikenheiser, A. M., Marrero-Garcia, Y., & Schoenbaum, G. (2017). Suppression of ventral hippocampal output impairs integrated orbitofrontal encoding of task structure. *Neuron*, *95*(5), 1197-1207.e3. doi: 10.1016/j.neuron.2017.08.003

Wikenheiser, A. M., & Redish, A. D. (2013). The balance of forward and backward hippocampal sequences shifts across behavioral states. *Hippocampus*, *23*(1), 22-29. doi: doi.org/10.1002/hipo.22049

Wikenheiser, A. M., & Redish, A. D. (2015a). Decoding the cognitive map: ensemble hippocampal sequences and decision making. *Current Opinion in Neurobiology*, *32*, 8–15. doi: 10.1016/j.conb.2014.10.002

Wikenheiser, A. M., & Redish, A. D. (2015b). Hippocampal theta sequences reflect current goals. *Nature Neuroscience*, *18*(2), 289–294. doi: 10.1038/nn.3909

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps

in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. doi: 10.1038/nrn.2016.56

Wilson, M., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676–679. doi: 10.1126/science.8036517

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. doi: 10.1016/j.neuron.2013.11.005

Wimmer, E. G., & Büchel, C. (2019). Learning of distant state predictions by the orbitofrontal cortex in humans. *Nature Communications*, *10*(1). doi: 10.1038/s41467-019-10597-z

Wimmer, G. E., & Büchel, C. (2020). Reactivation of pain-related patterns in the hippocampus from single past episodes relates to successful memory-based decision making. *bioRxiv*. doi: 10.1101/2020.05.29.123893

Wittkuhn, L., & Schuck, N. W. (2021). Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nature Communications*, *12*(1), 1795. doi: 10.1038/s41467-021-21970-2

Wolosin, S. M., Zeithamova, D., & Preston, A. R. (2012). Reward modulation of hippocampal subfield activation during successful associative encoding and retrieval. *Journal of Cognitive Neuroscience*, *24*(7), 1532–1547. doi: 10.1162/jocn{_}a{_}00237

Wood, E. R., Dudchenko, P. A., Robitsek, R., & Eichenbaum, H. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron*, *27*(3), 623–633. doi: 10.1016/s0896-6273(00)00071-4

Wu, C.-T., Haggerty, D., Kemere, C., & Ji, D. (2017). Hippocampal awake replay in fear memory retrieval. *Nature Neuroscience*, *20*(4), 571–580. doi: 10.1038/nn.4507

Yu, J. Y., & Frank, L. M. (2015). Hippocampal–cortical interaction in decision making. *Neurobiology of Learning and Memory*, *117*, 34–41. doi: 10.1016/j.nlm.2014.02.002

Yu, J. Y., Liu, D. F., Loback, A., Grossrubatscher, I., & Frank, L. M. (2018). Specific hippocampal representations are linked to generalized cortical representations in memory. *Nature Communications*, *9*(2209). doi: 10.1038/s41467-018-04498-w

Zhang, H., Deuker, L., & Axmacher, N. (2017). Replay in humans - first evidence and open questions. In N. Axmacher & B. Rasch (Eds.), *Cognitive neuroscience of memory consolidation* (pp. 251–263). Springer. doi: 10.1007/978-3-319-45066-7_15

Zhang, H., Fell, J., & Axmacher, N. (2018). Electrophysiological mechanisms of human memory consolidation. *Nature Communications*, *9*(1). doi: 10.1038/s41467-018-06553-y

Zhang, S., & Sutton, R. S. (2017). A Deeper Look at Experience Replay. *arXiv e-prints*, arXiv:1712.01275.

Zielinski, M. C., Tang, W., & Jadhav, S. P. (2018). The role of replay and theta sequences in mediating hippocampal-prefrontal interactions for memory and cognition. *Hippocampus*. doi: 10.1002/hipo.22821