THE COMPUTATIONAL BENEFITS OF EPISODIC MEMORY IN SPATIAL LEARNING

Xiangshuai Zeng
Institute for Neural Computation
Faculty of Computer Science
Ruhr University Bochum

Laurenz Wiskott
Institute for Neural Computation
Faculty of Computer Science
Ruhr University Bochum

Sen Cheng
Institute for Neural Computation
Faculty of Computer Science
Ruhr University Bochum

November 24, 2021

ABSTRACT

Episodic memory has been studied extensively in the past few decades, but so far little is understood about how it is used to affect behavior. Here we postulate three learning paradigms: one-shot learning, replay learning, and online learning, where in the first two paradigms episodic memory is retrieved for decision-making or replayed to the neocortex for extracting semantic knowledge, respectively. In the third paradigm, the neocortex directly extracts information from online experiences as they occur, but does not have access to these experiences afterwards. By using visually-driven reinforcement learning in simulations, we found that whether an agent is able to solve a task by relying on the three learning paradigms depends differently on the number of learning trials and the complexity of the task. Episodic memory can, but does not always, have a major benefit for spatial learning, and its effect differs for the two modes of accessing episodic information. One-shot learning is initially faster than replay learning, but the latter reaches a better asymptotic performance. We believe that understanding how episodic memory drives behavior will be an important step towards elucidating the nature of episodic memory.

Introduction

Even though there is wide consensus that episodic memory (EM) is the memory of personally experienced episodes, (Tulving, 1972), the precise conceptualization of episodic memory has been difficult to come by. One source of this difficulty might be a dominant focus on the properties of EM, whereas little is known about its function. While it seems self-evident that information from the past is useful for, and required by, any complex behavior, it is much less clear what exactly the benefit is of storing the experience of a *particular* episode in long-term memory.

Research on human memory often addresses this question at an abstract, conceptual level. For instance, Klein et al. (2009) suggests that maintaining a pool of episodic memories enables its owner to reevaluate an individual's past behavior in light of new information, thus serving an important role in social interaction. Mahr and Csibra (2017) discusses the communicative function of episodic memory from a philosophical point of view. While social functions are undoubtedly important, especially for human beings, it is unclear how EM informs actions in these suggested cases. Another influential idea is based on the survival processing benefit (Nairne et al., 2007), which means that subjects can remember names of objects that are relevant for survival in the wilderness better than non-relevant words. Nairne and Pandeirada (2016) therefore argues that episodic memory has adapted to ensure our survival in the kind of environments that our stone-age ancestors lived in (adaptive memory theory). While the potential benefit of remembering survival-related material better than other material is evident, it remains unknown how the recalled memory drives behavior. Moreover, EM probably has other functions beyond remembering survival-relevant items that are not covered by the adaptive memory theory. A broader suggestion by Suddendorf and Corballis (1997, 2007) is that EM is an epiphenomenon of the capacity to mentally travel into the future. Forecasting the future, they argue, enables us to take the appropriate actions in the present to ensure a favorable outcome in the future. However, the mental time travel hypothesis leaves two fundamental questions about the function of EM unanswered. How exactly do memories

about past episodes inform upcoming or future actions? Why do we need memories of particular episodes for mental time travel into the future?

In computational neuroscience, a specific suggestion is that EM provides the data to extract regularities from multiple, repeated experiences (Nadel and Moscovitch, 1998; Cheng, 2017). A related suggestion was made in the Complementary Learning Systems (CLS) theory, where they suggested that replay of episodic memory supports integration of new information into the semantic network (McClelland et al., 1995). According to CLS theory, replay facilitates interleaved training, i.e., the alternating presentation of new and old information, which avoids catastrophic interference (McCloskey and Cohen, 1989) in the semantic network. Although hippocampal replay was hypothesized to play a role in learning more than three decades ago (Buzsaki, 1989), and CLS theory provides a specific suggestion for its computational function, it is still lacking a functional role of EM in driving behavior, which is required for a measure of performance. A link to behavior is provided by the studies in reinforcement learning (RL) where agents need to take sequences of actions in an environment to maximize the expected accumulated reward. Early work used online learning exclusively, i.e., an experience drove learning exactly once. Later it was found that replaying earlier experiences speeds up learning in many RL tasks (Lin, 1992). Recent advances in utilizing episodic-like memory led to superhuman performance on many (video) games (Mnih et al., 2015).

In this paper, we first propose three learning paradigms based on two different ways in which EM is used, or the situation when it is absent, and hypothesize what the learning curve might look like in these paradigms. Then we build computational models of the three hypothesized learning paradigms in the framework of RL and conduct simulations of spatial learning to test our hypothesis. We focus on spatial learning because so much is known about neural activity in the hippocampus during spatial navigation (the hippocampus is closely linked to episodic memory) and there is much interest in spatial navigation in RL. The results of our experiments show that agents deploying the three learning paradigms separately are able to solve a given task at different speeds, and that the harder the task is, the more profound the difference is. In addition, the agents also show different behavior patterns and reach different asymptotic performance during and at the end of the learning. Our results then make a few predictions about the nature and functions of episodic memory in spatial learning.

Hypotheses

One key to understanding the function of EM is to recognize that EM can provide information that is useful for learning in at least two fundamentally different modes. First, we suggest that the most direct way of using EM to drive behavior is to retrieve a sequence of events that compose the episode and that include information about the actions performed, and use that information directly to optimize a sequence of actions. For instance, a rat might go to one arm of a T-maze because it remembers that it found a piece of cheese in that arm once before (Fig. 1A). We term this kind of learning *one-shot learning*. Second, we suggest that EM can be replayed offline repeatedly to drive learning in the neocortex, as suggested by (Cheng, 2017; Nadel and Moscovitch, 1998; McClelland et al., 1995). Replay enables the neocortex to acquire the information about how to solve a cognitive task from multiple memory traces of similar experiences. For instance, a rat running in a T-maze might learn that a piece of cheese is always located in the right arm. We term this kind of learning *replay learning*. In addition, learning can also occur without employing EM, since the neocortex can directly extract semantic information from online experiences as they occur (*online learning*).

We hypothesize that the learning speed of the three learning paradigms have the relationship: one-shot learning $>$ replay learning $>$ online learning. The direct use of episodic memory does not require multiple updates so it is one-shot, which depends on the hippocampus (Cheng, 2013). By contrast, it takes time for the neocortex, a large and distributed network, to extract semantic information from memory replay (Cheng, 2017). Hence, we hypothesize that replay learning is slower than one-shot learning, but still faster than online learning, because old experiences can be extensively reused via replay. One-shot and replay learning paradigms coexist in a healthy brain, whereas hippocampal animals or patients, whose EM system is impaired, can rely only on online learning. Indeed, hippocampal patients were still able to acquire new semantic memories but generally required much more learning trials to master the new knowledge, while controls could learn the same contents after a single trial (Rosenbaum et al., 2005; O'Kane et al., 2004). Similar observations have been made in rodents as well (Wiltgen et al., 2006; Kosaki et al., 2014).

These hypotheses are best summarized and formalized in the learning curves of the three learning paradigms (Fig. 1B). If the task is simple and the animal is given a sufficient number of learning trials (dashed line), all three learning paradigms reach the performance criterion. However, if the animal is given fewer trials, hippocampal animals might not be able to solve it because they only have online learning at their disposal -– consistent with experimental reports (Wiltgen et al., 2006; Kosaki et al., 2014). By contrast, control animals can use replay learning or one-shot learning based on EM and, therefore, learn the same task in fewer trials. We expect little difference in the performance between replay and one-shot learning in simple tasks, because both are able to learn the task quickly. Therefore, it would not be possible to distinguish which type of learning was responsible in this case. In other words, although a hippocampal

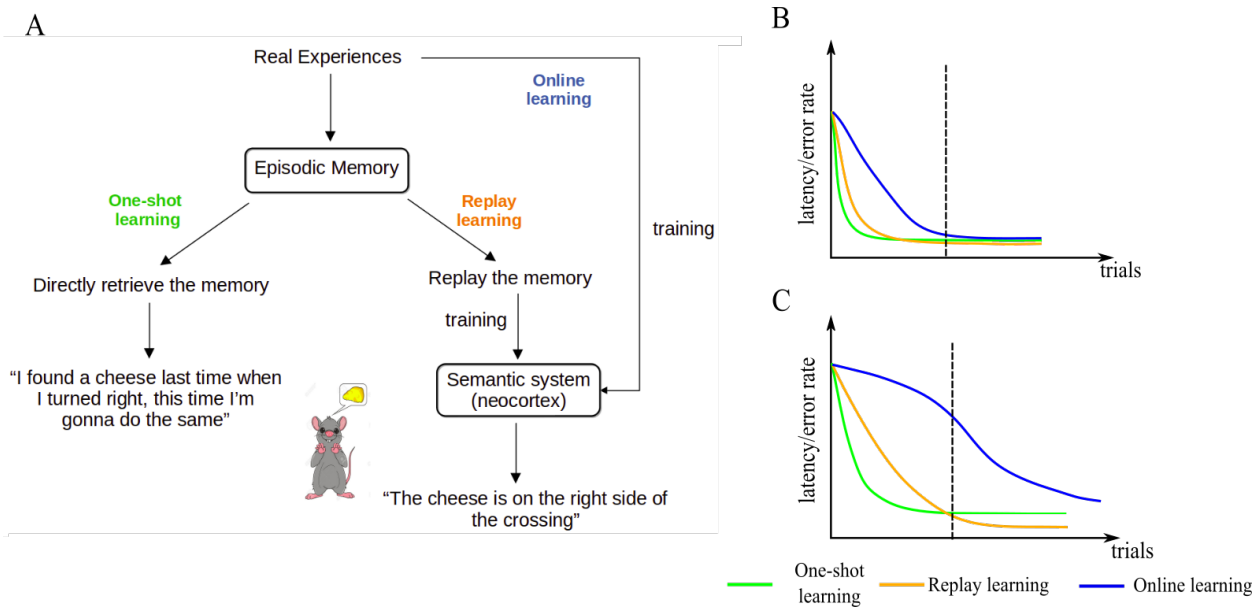

Figure 1: **The three hypothesized learning paradigms and their learning curves. A**: A schematic illustration of the three learning paradigms for the example of a rat running in a T-maze. The hypothesized learning curves of the three learning paradigms for **B**: a simple task and **C**: a difficult task. The vertical dash line indicates the time limits of a typical experiment. Note that the asymptotic performance of semantic learning with replay is better than that of episodic learning.

lesion would abolish rapid learning in a simple task, interfering with replay would not, because one-shot learning is still intact.

As the task complexity increases, more subtle features of the three learning paradigms are revealed (Fig. 1C). All learning paradigms start at worse performance (higher latency/error rates) and the rate of improvement will also become lower, too. However, the different learning paradigms are affected to different degrees. Online learning is affected the most, yielding a learning curve that is quite flat at the beginning of learning (Fig. 1C, blue line). This means that a hippocampal animal, which uses only online learning paradigm, cannot solve the more complex task within the time limits of a typical experiment. We further hypothesize that replay learning extracts asymptotic solutions that are more efficient than those extracted by one-shot learning, which is more prone to persist on a sub-optimal solution because of its higher reliance on particular experiences. Therefore, we expect that abolishing replay has different effects on task performance depending on the timing of the test trial. If tested early during learning, abolishing replay might not have much of an effect because one-shot learning is as fast as, or faster than, replay learning. However, after extensive training, replay learning performs better than one-shot leaning, so that animals whose hippocampal replay has been disrupted cannot reach the same performance as that of controls. This dissociates the dependence on the hippocampus from the dependence on hippocampal replay.

In this paper, we investigate the quantitative properties of the three learning paradigms in a computational model applied to spatial learning. These studies allow us to test, and confirm, the above hypotheses about the learning curves of the three learning paradigms.

Materials and Methods

Computational modeling framework

All simulations were performed in a virtual-reality modeling framework (Fig. 2), which was developed to study models of rodents behavior in spatial navigation and extinction learning tasks (Walther et al., 2021) and named CoBeL-RL (**C**losed-l**o**op simulator of complex **Be**havior and **L**earning based on **R**einforcement **L**earning). The virtual environments were designed with the unity game engine (`https://unity.com/`), while the remaining parts of the framework were developed using Python. In the simulations, an artificial agent equipped with a wide-view camera represented an animal navigating in the virtual environment. The World Topology module determined the spatial

positions and orientations, which the agent could be placed in, and the allowed transitions between them. This topology was only used to determine which transitions were valid in the simulation and were not known to the learning agent. The OpenAI Gym Interface (`https://gym.openai.com/`) module formed the interface between the virtual environments and the RL agent, transmitting information and control commands between the two.

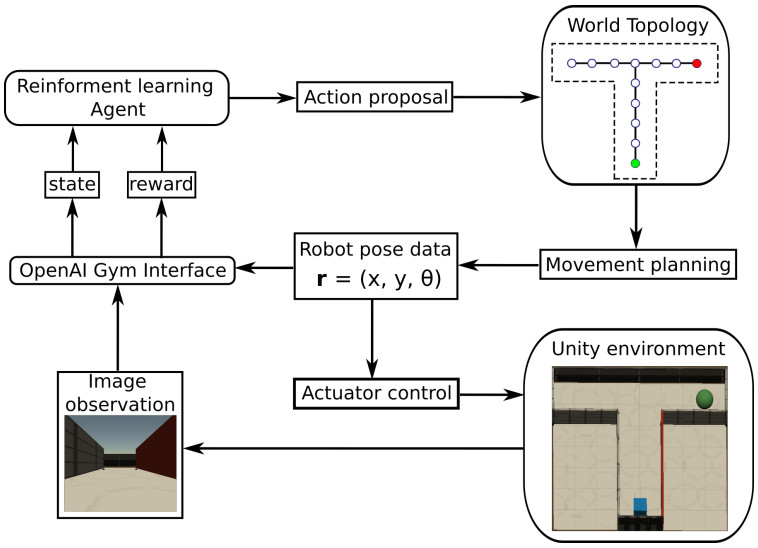

Figure 2: **Schematic of the CoBeL-RL simulation framework.** Modules are sketched with rounded rectangles. Other rectangles represent data and commands. The information or control flows are depicted with arrows. State and reward are concepts in reinforcement learning. `https://gitlab.ruhr-uni-bochum.de/cns/1-frameworks/CoBeL-RL`

Simulated spatial learning tasks

We designed a series of virtual environments for spatial navigation in the Unity simulator to test our hypotheses (Fig. 3A-E). The agent was always placed in the same starting location (blue square) and had to find a goal (green disk). The sensory input to the learning agent consisted of an 84×84 RGB image (Fig. 3F). The agent could face four different orientations at each node: north, west, south, or east; it could take six different actions: go forward, backwards, left, right, rotate 90 degrees clockwise, or rotate 90 degrees counterclockwise. The complexity of the spatial navigation tasks increased from Fig. 3A to Fig. 3E, because both the total number of nodes and the minimum number of transitions that the agent needed to reach the goal increase. Once the agent reached the goal location, or the trials timed out after 600 time steps, we ended the current trial and returned the agent to the initial position. At the beginning of the experiment the agent had no prior knowledge about the environment, hence it was compelled to randomly explore it at the beginning of the training.

Reinforcement learning (RL)

We modeled spatial learning in the framework of RL, where an agent interacts with an environment. At each time step t, the agent observes the state of the environment represented by $s_t \in S$, where S is the set of all possible states, and takes an action $a_t \in A(s_t)$ in response, where $A(s_t)$ represents the set of all possible actions the agent can take in state s_t. These actions, in combination with the dynamics of the environment, lead to a new state in the environment at the next time step s_{t+1}, which the agent also observes. In addition, the agent receives a reward r_{t+1}. These steps are repeated until a terminal state is reached, which indicates an episode (or a trial in the convention of neuroscience) has ended. For the spatial navigation tasks (Fig. 3), the state was defined by the RGB image that the agent received from the camera. At each time step, the agent received no reward ($r_t = 0$) unless it found the goal and ended the episode, where it was rewarded with $r_T = +1.0$ with T indicating the last step of the current episode.

In RL, the behavior of the agent and learning is driven by the discounted cumulative reward function received in a trial:

$$R_t = \sum_{\tau=t+1}^{T} \gamma^{\tau-t-1} r_\tau, \tag{1}$$

where γ is a discount factor that determines the relative values of immediate reward vs. those that are more distant in time with $0 < \gamma <= 1$. The objective of the agent is to develop a policy π that maximizes the expected cumulative

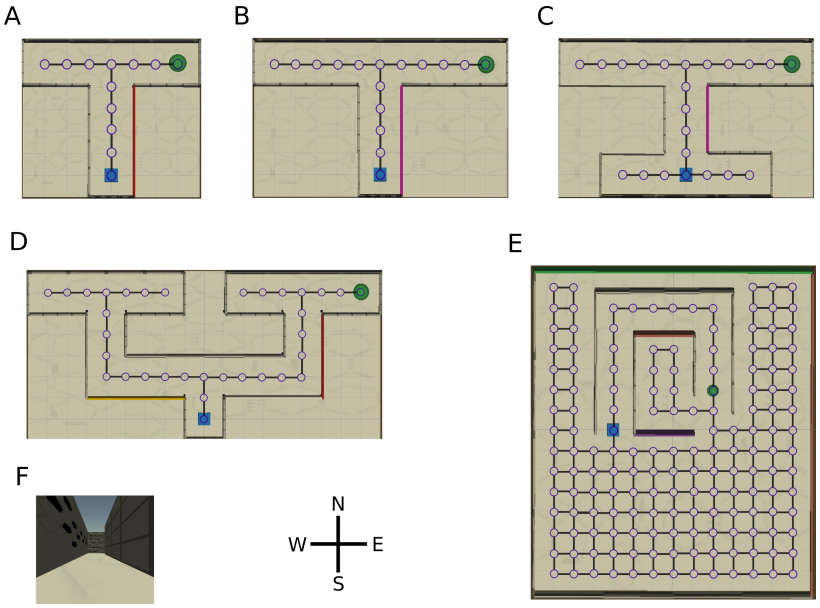

Figure 3: **Overview of the virtual environments with their topology graph.** The starting location is represented by a blue square and the goal by a green disk. Nodes in the graph represent allowed positions for the agent and solid lines allowed transitions between positions. On each node, the agent can face in four orientations: north, west, south, and east. The starting position of the agent and the location of the goal in the graph remain constant during the simulation. Note the topology graph of the environments are only drawn on this figure for demonstration purpose, they are not visible to the agent. **A**: T-maze, **B**: long T-maze, **C**: H-maze, **D**: double T-maze, **E**: tunnel maze, and **F**: an example view collected by the camera placed on the top of the agent.

reward

$$G_t = E_\pi \left(R_t \right). \tag{2}$$

One method for solving this task is Q learning, a class of RL algorithms, where the agent learns a state-action value function, the so-called Q function. The scalar function $Q(s_t, a_t)$ measures how desirable the state-action pair (s_t, a_t) is to the agent. If learned correctly, the larger the value of the Q function, the larger a cumulative reward the action a_t in state s_t will yield. Mathematically, the Q function can be expressed as

$$Q(s_t, a_t) = E_\pi \left(R_t | s_t, a_t \right) \tag{3}$$

In state s_t, the agent selects the action a_t with the highest Q value. To balance exploration and exploitation, we used the ϵ-greedy algorithm: with probability ϵ the agent will randomly select an action from A regardless of the Q values, otherwise the agent selects the action that yields the highest Q value. Throughout our simulations we set $\epsilon = 0.1$. The discount factor is set as $\gamma = 0.9$ to encourage the agent to find the shortest path from the initial position to the goal.

The three learning algorithms

We selected three RL algorithms based on Q learning to model our hypothesized learning paradigms (Fig. 4): Model-free Episodic Control (EC) (Blundell et al., 2016) for one-shot learning, Deep Q Learning with memory replay (DQN) (Mnih et al., 2015) for replay learning, and online Deep Q Learning (online DQN) for online learning. Real experiences were modeled as a sequence of (state, action, next state, reward) tuples, i.e., $(s_t, a_t, s_{t+1}, r_{t+1})$. Like in EM (Cheng and Werning, 2016), these sequences of events were stored in memory for one-shot learning in EC and for replay learning in the DQN algorithm.

Model-free episodic control (EC). The Q function is represented as a table whose entries are directly computed using sequences of past experiences stored in memory by using the following equation

$$Q(s_t, a_t) = \begin{cases} R_t & \text{if the state is unknown} \\ \max\{R_t, Q(s_t, a_t)\} & \text{if the state is known} \end{cases} \tag{4}$$

Note that the sequential ordering of experience tuples is critical in EC because of equation 1. When making a decision later, the agent will look up the table and find the state that is the same as the current one and then select an action based

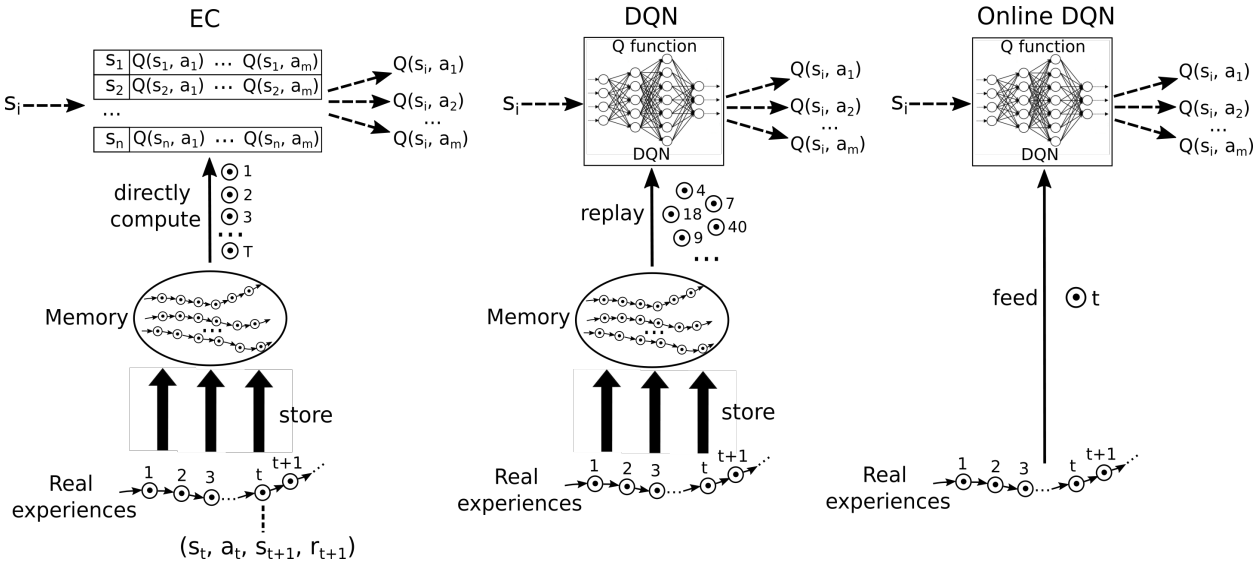

Figure 4: **A schematic illustration of the three reinforcement learning algorithms used in this study.** Each circle with a dot inside represents an experience, defined as a sequence of (state, action, reward, next state) tuples. Real experiences are collected sequentially and stored in memory. Model free episodic control (EC) uses this sequential information to systematically extract reward information, which are stored in a table of Q values. The Deep Q network (DQN) selects experiences randomly for replay to train a deep neural network, which then represents the Q function. The online DQN does not store experiences in memory and uses each experience only once to train the deep neural network representing the Q function.

on the retrieved Q values. The $\max$ operation in Eq. 4 guarantees that the sequence starting from (s_t, a_t) that leads to the largest cumulative reward encountered so far will be followed by the agent. For states that have never been visited before, the Q values are approximated by averaging the values of the k-nearest states. We used $k = 5$.

Blundell et al. (2016) originally utilized two different dimension-reduction methods to pre-process the raw inputs in order to decrease the computational requirements: Variational Autoencoder (VAE, Kingma and Welling, 2014) and random projection. We chose the latter in our implementation. Specifically, raw images generated from Unity of size $84 \times 84 \times 3$ were projected onto a lower-dimensional space, i.e., $\phi : x \rightarrow Mx$ where $M \in \mathbb{R}^{F \times D}$ and D is the dimension of the original inputs. The entries of matrix M were drawn from a standard Gaussian and according to the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss, 1984), this transformation preserves relative distances in the original space. In our implementation, the dimension of the projected space was $F = 256$. To further speed up inference and learning, the states stored in memory were used to construct a KD tree so that the search of closest neighbours to a given state becomes efficient. Lastly, we set the maximum number of experiences that can be stored to $50,000$.

Deep Q network (DQN). The DQN represents the Q function as an artificial deep neural network (DNN), which maps a state s_i to the Q values of all the possible actions on this state. During learning, a mini-batch of experience tuples are randomly selected from memory and used to construct a loss function according to

$$L = \sum_i \left(r_{i+1} + \max_{a'} Q(s_{i+1}, a') - Q(s_i, a_i) \right)^2 \tag{5}$$

The backpropagation algorithm is used to minimize this loss function by adjusting the DNN's weights. We consider the DNN to loosely represent the semantic network in the neocortex, and the whole learning process as replaying past experiences from episodic memory to the semantic system for extracting information. Unlike that in EC, the sequential ordering of experience tuples is not important in DQN, because the items are randomly chosen for replay.

We used a DNN with the same architecture as Mnih et al. (2015). The input, $84 \times 84 \times 3$ image, was first passed to a convolutional layer consisting of 16 8×8 filters with stride 4 and then to a second convolutional layer consisting of 32 4×4 filters with stride 2. The last hidden layer consisted of 256 fully connected units. The output layer had 6 units, each corresponding to the Q value of one of the possible actions. A rectifying linear unit (ReLU) was used as activation function in all layers except for the output layer, where a linear activation function was used. At each

time step, a mini-batch of 32 samples was randomly drawn from the memory to update the network using the Adam optimizer with a learning rate of 0.0001.

Online Deep Q network. To model spatial learning without EM, online learning was based only on the experiences as they occur. Specifically, at each time step t, a loss function was constructed using only the current experience tuple according to Eq. 6. The DNN then minimized this loss function in order to find the optimal Q function. Note that each experience tuple was used only once for learning.

$$L_t = \Big(r_{t+1} + \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big)^2 \tag{6}$$

The hyper-parameters for the online DQN agent were the same as for the DQN agent except that there was no experience replay.

Results

To test our hypotheses regarding the function of EM in learning, we applied each of the three RL algorithms (EC: one-shot learning, DQN: replay learning, online DQN: online learning) to solve the five different spatial navigation tasks (Fig. 3). The agents trained using the three algorithms are referred to as EC agent, DQN agent, and online DQN agent, respectively. To reduce the variance caused by the randomness in the training, we performed 25 independent runs for each task-algorithm combination.

Episodic memory facilitates spatial learning in two different modes

The learning curves, for the three agents in the five tasks qualitatively, confirm all of our hypotheses (Fig. 5). Namely, the learning speeds of the three algorithms have the relationship: EC $>$ DQN $>$ online DQN in all of the tested environments. The complexity of the task indeed increases from the T-maze (Fig. 5A) to the double T-maze (Fig. 5D), since the learning curves of the three RL algorithms all start at a larger number of time steps required for completing a trial. Also, as the complexity increases, the learning curves shift to the right, which means that more learning trials are required to find the shortest path. It is worth noting that the online DQN algorithm is more sensitive to the change of the task complexity, while the learning curves of EC and DQN shift up and rightwards only slightly, hence they are less sensitive. For instance, if we limited the number of trials to 100 (Fig. 5, dashed line), then online DQN would not be able to solve the more complex double T-maze, even though it could solve the simpler T-maze, whereas both the EC and DQN agents could solve both tasks. Note that if we decreased the number of trials to one, then the only paradigm that could solve any task is EC (one-shot learning). Finally, the DQN agent exhibits better asymptotic performance than the EC agent, as we hypothesized. We show below that this difference exists in all tasks, even though this effect does not stand out in the learning curves due to the scale of the plot.

We hypothesize that there are two main reasons why online learning is so slow. First, since the agent is only rewarded where it finds the goal (sparse rewards) and updates of the DNN are based only on the current experience that connects two consecutive states, it takes many repeated experiences to propagate reward information from the goal to other states in the environment, particularly to the starting point. Second, learning is based on gradient decent on the loss function, so performing updates with single data points results in noisy updates of the DNN. This makes learning particularly unstable and accounts for the much larger variance in the performance of the online DQN agent. This is why mini-batch updating is prevalent to reduce the noise in gradient descent in the field of machine learning (Masters and Luschi, 2018). Therefore, replaying past experiences of the agent from EM not only speeds up the propagation of the reward information but also stabilizes the learning.

The learning curves of the three agents are the main results. In the following, we analyze in more detail how and what the three agents learn so that we gain more insight into how EM affects learning. The analyses also test the two new hypotheses about the mechanisms by which EM speeds up learning.

Replay learning finds better asymptotic solutions as compared to one-shot learning

To demonstrate the differences between the solutions found by the EC and DQN agent, we visualized single example trajectories that the agents took inside the tunnel maze during the test trial (Fig. 6). The EC agent took the green trajectory, which is longer than the orange one taken by the DQN agent. In the tunnel maze, the green trajectory can be found more easily by random exploration of the agent because once the agent moves into the tunnel, it is very likely that the agent moves along the tunnel and reaches the goal since its movement is constrained by the walls. By contrast, the agent has to explore the open space in the bottom half of the maze to find the shorter orange trajectory, which takes more trials and errors. The EC agent learns quickly from single experiences. Therefore, once it has found the green

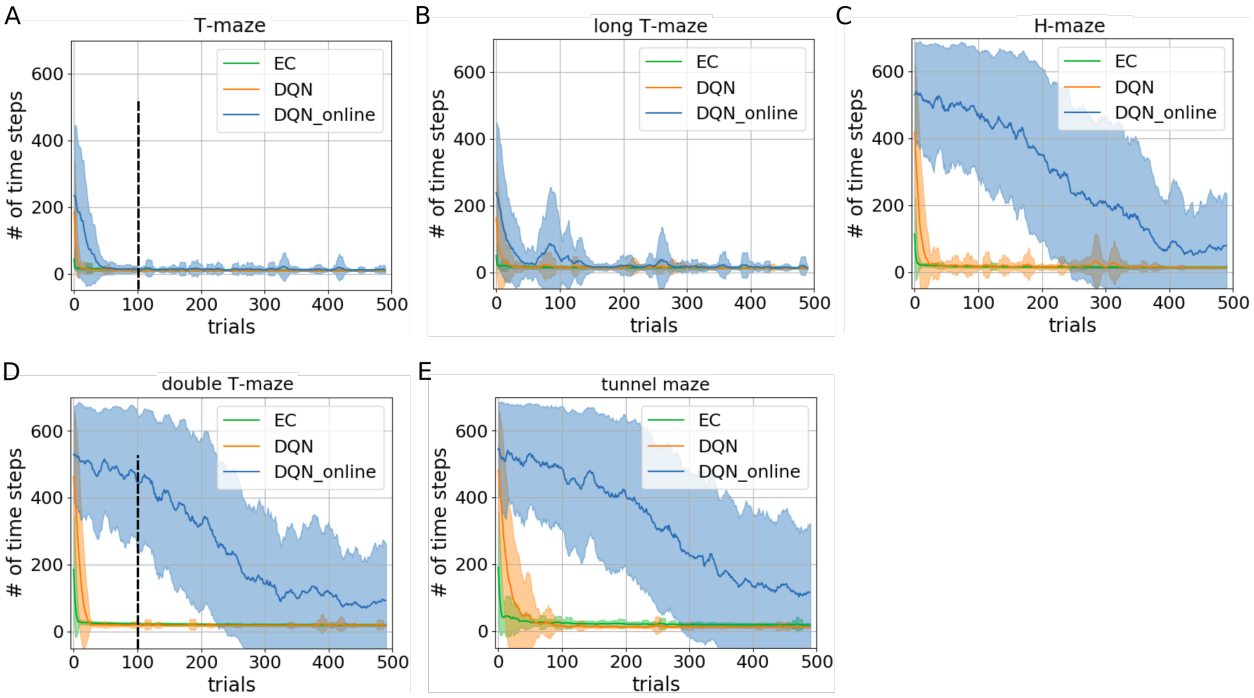

Figure 5: **Learning curves for the three learning paradigms in different environments.** Each panel shows the number of time steps that the agent takes to find the goal in a specific trial. The curves and shaded areas indicate the mean and the standard deviation, respectively, obtained from 25 independent runs. Model-free Episodic Control (EC) represents one-shot learning, Deep Q Learning with memory replay (DQN) represents replay learning, and online Deep Q Learning (online DQN) represents online learning.

trajectory it is more likely to stick with it. A DQN agent however is more likely to find the better solution, because it learns more slowly and, thus, explores the environment more extensively (more on this below).

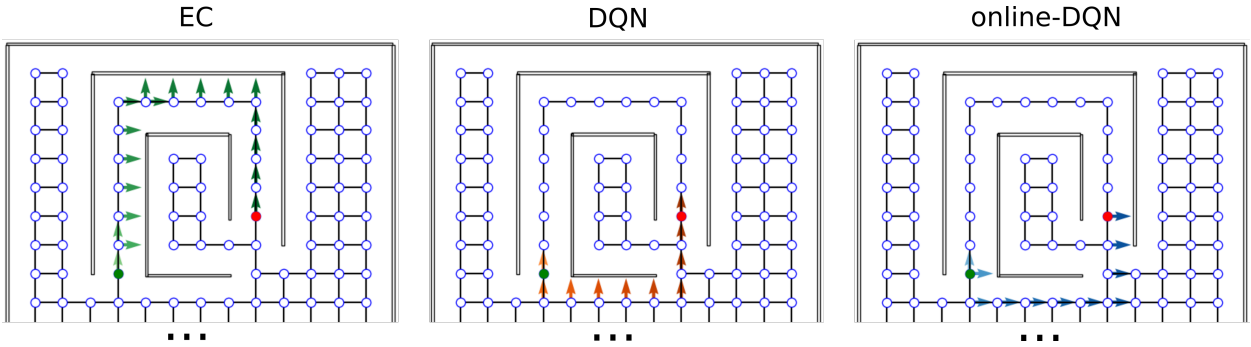

Figure 6: **Sample trajectories in the tunnel maze for three learning algorithms.** An arrow is attached to each node that the agent visits in one test trial. The direction of the arrow indicates the orientation of the agent, not the chosen action; lighter colors represent visits earlier in the trajectory and darker colors later in the trajectory. The starting and goal node are depicted in green and red, respectively. Model-free Episodic control (EC), Deep Q network (DQN), and online Deep Q network (online-DQN).

Beyond single example trajectories, we find that in 52% (13/25) of the runs, the EC agent finds the trajectory along the tunnel, while in only 12% (3/25) of the runs, does the DQN agent employ the same strategy to reach the goal. In short, the EC agent (one-shot learning) learns the fastest but has a disadvantage: it is more prone to get stuck in a sub-optimal solution when there are multiple solutions and the globally optimal one is more difficult to find.

To compare the asymptotic solutions found by the three learning algorithms more systematically, we placed the trained agents on the starting position inside each maze and recorded the number of time steps they took to find the goal (Fig. 7).

This analysis revealed two interesting results. Firstly, there were always a few simulations where the online DQN agent could not find a solution from the starting node to the goal, indicated by the data points around 600 time steps (the time out value). It is these unsuccessful test trials that contribute to the high average number of time steps in the solutions found by the online-DQN agent. Nevertheless, even the online DQN agent discovered trajectories with lengths similar to those found by the EC and DQN agent, as indicated by the data points inside the blue bars. This further demonstrates that the online learning paradigm is a very unstable process whose success depends on the randomness in the training.

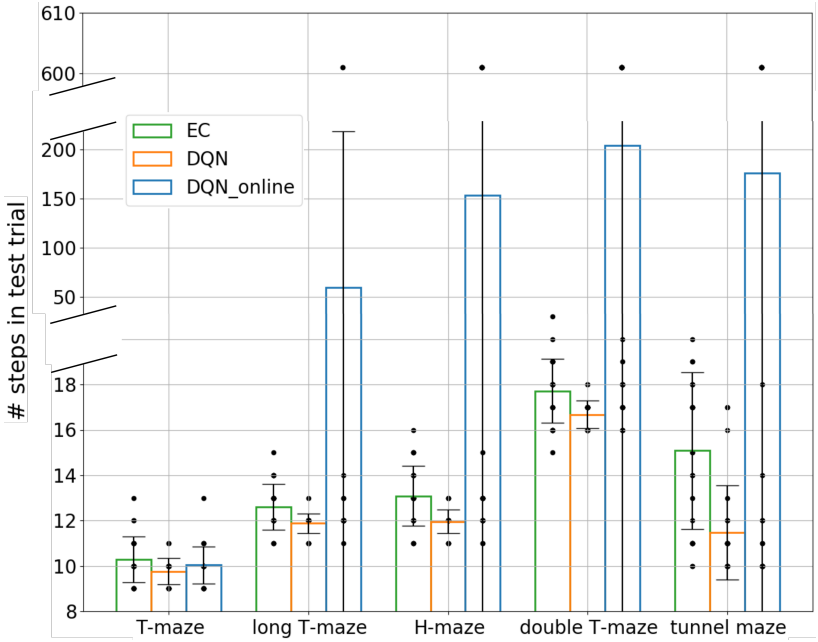

Figure 7: **The number of time steps that the agent takes to find the goal in test trial.** Bars represent the average number of steps for each algorithm-task combination over 25 runs. Each black dot represents a single run.

Secondly, in every environment, on average the DQN agent was able to find better solutions, i.e., shorter trajectories, compared to the EC agent (Fig. 7). This difference is especially pronounced in the tunnel maze where there are multiple different routes from the start to the goal (see Fig. 3E), because of how EC and DQN utilize episodic memory. EC directly retrieves a sequence of past experiences for making decisions, and therefore follows the first solution it finds, which is discovered through randomly exploring the maze and, hence, is usually not an optimal trajectory. By contrast, DQN constantly extracts the optimal solution based on all its past experiences, which is a gradual process and enables the agent to explore the environment extensively and yields more diverse experiences compared to EC. Thus, it is more likely for the DQN agent to extract a near-optimal solution. Of course, it is possible that the first solution found by the EC agent is near optimal, but the probability of this incident is small, especially when the state space is large. This also means that the solutions found by EC between independent runs can be very different, which explains the larger variance of time steps required to reach the goal for the EC as compared to the DQN agent (green and orange bars, respectively, Fig. 7).

One-shot learning propagates reward information more efficiently than replay learning does

Next, we examined how extensively the three agents explored the environments during training and how well the reward information was propagated from the goal to other states (Fig. 8). In the training phase, we recorded the position (represented by a node index) and orientation of the agent inside the mazes at each time step and extracted only the unique combinations to represent the set of states that the agent visited during training. Note that each state (location, orientation) is uniquely associated with an RGB image. After the training was complete, we placed the agent in every state it had visited during training and checked whether the agent would navigate to the goal from that initial state, or not. If so, we considered the reward to have propagated to that initial state and call that state a solution state.

In one example in the tunnel maze (Fig. 8), the EC agent explored the least, but spread the reward information to most of the states it had visited. It is also apparent that the agent always goes into the tunnel to reach the goal no matter which state it starts from, since the reward information is not propagated to the node below the goal node (marked in red). The

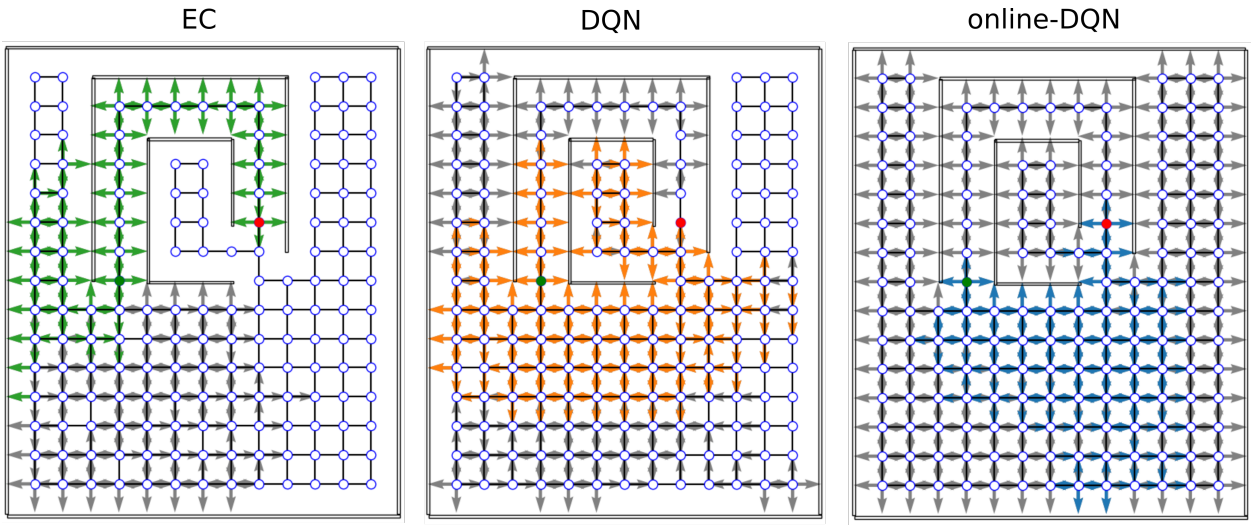

Figure 8: **Sample coverage and propagation of reward information in the tunnel maze.** An arrow (gray/colored) indicates that the agent has visited the node. The direction of the arrow indicates the orientation of the agent, not the chosen action. A colored (green/orange/blue) arrow indicates that the agent will navigate to the goal, if placed on that node in that orientation. The starting and goal nodes are depicted in green and red, respectively. Model-free Episodic control (EC), Deep Q network (DQN), online deep Q network (online-DQN).

presence of gray arrows indicates that there are states that were visited by the EC agent, from where it cannot find the goal. Since EC propagates the reward information in one shot at the end of each trial (Eq. 4), these states were probably visited in the very early stage of training when the agent failed to find the goal before the trial timed out. Finally, due to rapid learning, the behavior of the EC agent soon becomes highly repetitive after the first solution is discovered. By contrast, the DQN agent explored the maze more extensively, and also propagated the reward information to a large portion of the visited states as well. It is not easy to tell at which stage of the training the agent visits a gray state since the DQN propagates the reward signal gradually. Meanwhile, we do see an advantage of memory replay since the online DQN agent, who learns only from online experiences, explores nearly the entire maze, but spreads the reward information to the least range among the three algorithms.

A more systematic analysis (Fig. 9) shows that on average, the fractions of visited states in each maze for the three learning algorithms are inversely related to the learning speed, namely, online DQN $>$ DQN $>$ EC. That is, the faster the agent learns, the less it explores the environments. Although the online DQN agent explored almost 100% percent of the state space in each maze, it propagated the reward information the least among the three algorithms. There does not appear to be a significant difference in the fraction of states from where the EC and DQN agent could navigate to the goal. However, since the EC agent explores a smaller fraction of the environment, it seems to have propagated the reward information more efficiently to the states it had visited. To analyze this advantage of EC, we plotted the ratio of the number of solution states to the total number of visited states for each algorithm-maze combination (Fig. 10). Indeed, in every maze, the EC agent propagated the reward information to a larger fraction of the explored states than the DQN agent did.

Finally, we analyzed whether the DQN and online DQN agents would propagate the reward information further, if given more training trials. By comparing the training outcome after 250 trials (Fig. 10A) to that after 500 trials (Fig. 10B), we see that the solution-states-to-visited-states ratio for the EC and DQN agents have converged, but it is quite possible that the online DQN agent would spread the reward information to a larger fraction of states, if it were given more training.

Discussion

To study the functional role of episodic memory (EM), we studied three learning paradigms that differ in how they access information stored in EM: one-shot learning, replay learning and online learning. To compare the three paradigms quantitatively, we chose three reinforcement learning algorithms and applied them to spatial learning tasks in simulated maze environments. The three agents received no prior information about the mazes and had to solve the navigation tasks based on raw sensory inputs by trial and errors. We found that whether an agent is able to on the number of learning trials and the complexity of the task. One-shot learning initially solves the task very quickly, but cannot reach

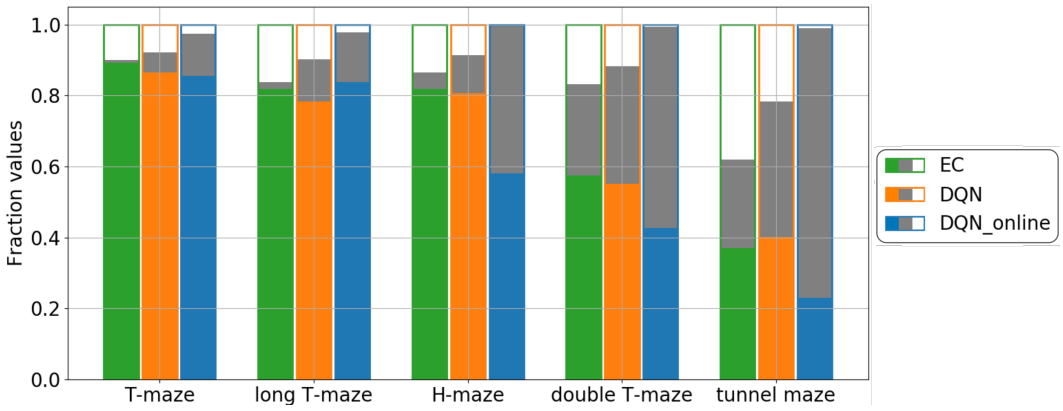

Figure 9: **Summary of training outcomes.** For each algorithm-task combination, bars show the fraction of states that were never visited during training (white fill), from which the agent will navigate to the goal, i.e., solution states (green/orange/blue fill), and were visited at least once during training, but from which the agent will not reach the goal (gray fill). The visited states were collected from the training of 500 trials and the solution states from 1 test trial after the training. The results represent an average over 25 runs.

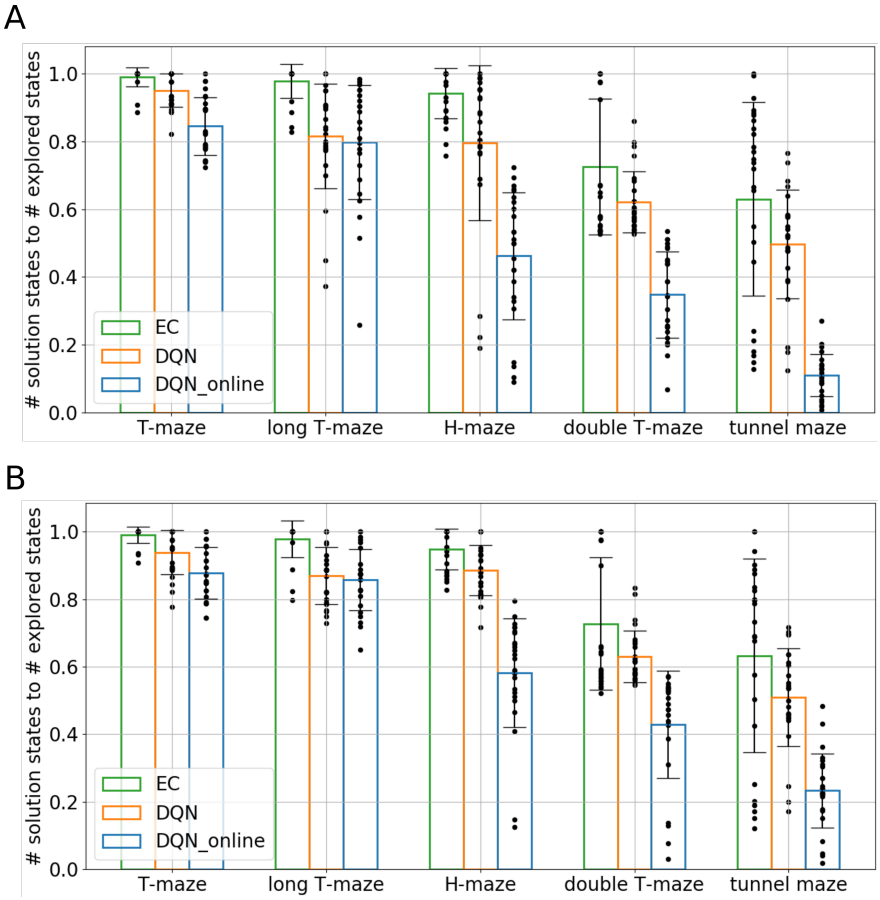

Figure 10: **Efficiency of reward propagation.** The ratio between the number of the solution states, i.e., states from which the agent will navigate to the goal, and the number of explored states during the training. Each bar represents an average over 25 simulations. **A**: after 250 learning trials. **B**: after 500 learning trials. While more learning trials allows replay (DQN) and online learning to propagate the reward information to a slightly larger fraction of visited nodes, the improvement is small.

the same asymptotic performance as that of replay learning, which converges more slowly but explores the environments more extensively. Online learning without EM is the most sensitive to changes in task complexity and is unstable. It therefore leads to large variability in learning performance. We further discuss the implications of our results below.

On the importance of studying the learning dynamics

While in hippocampal research it is fairly common to report learning curves, these are generally averaged over animals and/or blocks of learning. However, such averaging can produce misleading results (Gallistel et al., 2004; Smith et al., 2004) For instance, Cheng and Sabes (2006, 2007) show that studying the dynamics of learning reveals an enhanced picture of the process of adaptation, and Donoso et al. (2021) uncovered a large variability of behavior in extinction learning and renewal by focusing on individual learning curves. None of these findings could have been revealed by simply comparing the differences in performance before and after learning. Our modeling in this study suggests that it is key to experimentally measure the learning curves in simple and more complex tasks in control and hippocampal animals, as well as, when hippocampal replay or memory retrieval is inhibited separately. Comparing learning curves reveals much more about the learning process and the role of episodic memories, than comparing pairs of data points, e.g., comparing the performance of hippocampal to control animals in one task after a fixed number of learning trials. Hence, our modeling is another example in which much more information can be gained from studying the trial-by-trial dynamics of learning, rather than merely comparing before and after learning.

Implications for the role of the hippocampus in learning

Many tasks in animal experiments have been categorized as being hippocampal-dependent, such as, e.g., spatial navigation (Morris et al., 1982), contextual fear conditioning (Maren et al., 2013) and trace conditioning (McEchron et al., 1998), because a hippocampal animal performs worse in these tasks than a control animal. However, the literature suggests that the hippocampus might play diverse roles in these different tasks. In spatial navigation, the most widely held view suggests that the hippocampus provides a cognitive map (Moser et al., 2008), so that animals know where they are located and can plan where to go. However, the same brain region is also thought to be responsible for constructing a contextual representation in contextual fear conditioning (Maren et al., 2013), or bridging a temporal gap between the CS and US in trace conditioning (Bangasser et al., 2006).

Our current modeling results suggest a rather different picture. First, the function of the hippocampus is to serve as a crucial part of the biological substrate of the EM system (Cheng, 2013; Cheng and Werning, 2016), so that animals can perform one-shot learning and replay learning by accessing EM in two different modes. We suggest that all other cognitive functions that the hippocampus might be involved in can be traced back to its function in EM. Second, it might not be useful to categorically label a task as hippocampally-dependent, or not, since the dependence is determined by the complexity of the task itself and how extensive the training is. For instance, a simple spatial navigation task might not depend on the hippocampus, if a sufficient number of training trials are given to the hippocampal animal, which can still use online learning. Third, hippocampal lesion might have more widespread impact on behavior than traditional measures of success suggest. In our simulations, hippocampal lesions not only significantly decreases the learning speed of the animal, but also leads to a more thorough exploration of the environment, as well as, more variability in the learning process and outcomes due to the high instability of online learning.

Two different modes of accessing episodic memory

Two different modes of accessing episodic memory, retrieval and replay, have been studied before. On the one hand, retrieval entails the direct use of episodic memory and supports one-shot learning. Recently, Banino et al. (2020) has shown that retrieving EM with a recurrent attention mechanism enables transitive inference. On the other hand, replay allows the neocortex to extract semantic information from repeated experiences. It has also been demonstrated in computational models (McClelland et al., 1995; van de Ven et al., 2020) that replay can prevent catastrophic interference (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017) in semantic networks by enabling interleaved training, i.e., the alternating presentation of old and new information. This latter aspect of replay might be more important than the mere repetition of the experience and explains why increasing the learning rate in the online learning paradigms does not improve performance to the level of replay learning.

To our knowledge, there has not been an effort to directly and quantitatively compare the function of EM in the two modes. In our simulations, the two modes of accessing EM have different speeds and, as a consequence, let the agent explore the environment to different extents. Particularly, replay learning enables the agent to find better asymptotic solutions since the agent explores a larger fraction of the state space. By contrast, the one-shot learning agent shows highly repetitive behavior patterns immediately after finding the first solution. Therefore, our results suggest that impairing replay, but leaving EM intact otherwise, leads to very specific learning deficits that are different from deficits

due to abolishing EM altogether, i.e., anterograde amnesia. However, it might not be easy to observe this distinction at the behavioral level in animal experiments since the two modes coexist in a healthy brain and are normally impaired at the same time, usually by a hippocampal lesion.

To dissociate the functional role of replay from that of EM in general, one could disrupted Sharp wave–ripples (SWR) in the hippocampus selectively to prevent the replay of place cell activities, while keeping EM intact (Girardeau et al., 2009; Ego-Stengel and Wilson, 2010). Such studies found that indeed the asymptotic performance of the experimental group was impaired relative to that of controls. However, the experimental group also learned more slowly, suggesting that disrupting SWRs might impede one-shot learning as well. This discrepancy to our prediction might arise from a mixing of the two modes in learning tasks that span multiple days in the experiment. In our model, we studied the learning paradigms separately to get a clear view of the functional implications of each paradigm. In the future, it would be interesting to combine learning paradigms so that the agent can solve tasks faster and/or reach better asymptotic performance. Alternatively, we expect that if the task were simpler, so that it could be learned in a single-shot, then SWR disruption might reveal only a reduction in asymptotic performance but not in the initial learning rate.

Another way to test our hypothesis is to inhibit memory retrieval selectively, but leave memory replay intact. In this case, we predict that the experimental animals learn more slowly as compared to controls, but explore the environment more extensively and find a better solution at a late stage of training.

The sequentiality of episodic memory

While there are exceptions (Levy, 1996; Bayati et al., 2018; Cheng, 2013), sequentiality has not been widely accepted as an important feature of EM and many studies model EM as static memory patterns (Treves and Rolls, 1992; Hasselmo et al., 2002; Káli and Dayan, 2004). In our model, the sequentiality of episodic memory is key in the one-shot learning paradigm. This is because when there is a temporal distance between the agent's actions and their consequences, the sequential order of past experiences have to be maintained in memory to credit the past states and actions when updating the Q values. The importance of sequentiality is consistent with recent observations that theta sequences in the hippocampus are modulated by the animal's previous behaviour at the spatial location where the theta sequences occur (Parra-Barrero et al., 2021).

Although an agent is still able to extract semantic knowledge through memory replay, its learning speed is relatively slower and it cannot solve a task, especially a difficult one, in a one-shot manner. However, since life-threatening experiences should not be repeated, one-shot learning is crucial for the survival of an animal or species. In turn, one-shot learning in a complex task requires that sequences of experiences are stored in EM.

The role of sequential replay

In our model, experiences are replayed individually in random order while hippocampal activity patterns are replayed in sequential order (Louie and Wilson, 2001; Buhry et al., 2011). We did not model a biologically-plausible replay mechanism because, first, an artificial neural network is notoriously hard to train with correlated data (Bengio, 2012), and replaying training samples randomly breaks the serial correlations among the data, thus stabilizing the learning. Therefore, in the original implementation of deep Q learning (Mnih et al., 2015), experiences are randomly replayed and we chose to leave this part unchanged in our modeling. Second, the statistics of hippocampal replay in rodents is complex (Stella et al., 2019; Gupta et al., 2010; Ólafsdóttir et al., 2015) and difficult to model. Since the primary focus of this paper is about the functional consequence of memory replay, not the statistics of replay *per se*, we decided to simplify our modeling at the current stage. However, we believe there is a functional role of the sequential replay in the hippocampus and will investigate it in our future work.

In conclusion, our modeling study made a first attempt at studying the computational role of EM in spatial learning and quantitatively comparing two different modes of accessing EM: retrieval and replay. While many open questions remain, the computational framework allows us to clearly delineate episodic from semantic information and therefore could help resolve the controversial issue of what, if anything, separates episodic from semantic memory.

Acknowledgements

This work was supported by grants from the German Research Foundation (DFG), project number 419037518 – FOR 2812, P2 (S.C.) and project number 419039588 - FOR 2812, P5 (L.W.).

References

Bangasser, D. A., Waxler, D. E., Santollo, J., and Shors, T. J. (2006). Trace conditioning and the hippocampus: The importance of contiguity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(34):8702–6.

Banino, A., Chadwick, M. J., Botvinick, M., Badia, A. P., Köster, R., Zambaldi, V., Hassabis, D., Barry, C., Kumaran, D., and Blundell, C. (2020). MEMO: A Deep Network for Flexible Combination of Episodic Memories. *arXiv*, page 2001.10913.

Bayati, M., Neher, T., Melchior, J., Diba, K., Wiskott, L., and Cheng, S. (2018). Storage fidelity for sequence memory in the hippocampal circuit. *PLOS ONE*, 13(10):e0204685.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *arXiv:1206.5533 [cs]*.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., and Hassabis, D. (2016). Model-Free Episodic Control. *arXiv:1606.04460 [cs, q-bio, stat]*.

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

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

Cheng, S. (2013). The CRISP theory of hippocampal function in episodic memory. *Frontiers in Neural Circuits*, 7(May):88.

Cheng, S. (2017). Consolidation of Episodic Memory: An Epiphenomenon of Semantic Learning. In Axmacher, N. and Rasch, B., editors, *Cognitive Neuroscience of Memory Consolidation*, chapter 4, pages 57–72. Springer International Publishing, Cham, Switzerland.

Cheng, S. and Sabes, P. N. (2006). Modeling Sensorimotor Learning with Linear Dynamical Systems. *Neural computation*, 18(4):760–793.

Cheng, S. and Sabes, P. N. (2007). Calibration of Visually Guided Reaching Is Driven by Error-Corrective Learning and Internal Dynamics. *Journal of Neurophysiology*, 97(4):3057–3069.

Cheng, S. and Werning, M. (2016). What is Episodic Memory if it is a Natural Kind? *Synthese*, 193(5):1345–85.

Donoso, J. R., Packheiser, J., Pusch, R., Lederer, Z., Walther, T., Uengoer, M., Lachnit, H., Güntürkün, O., and Cheng, S. (2021). Emergence of complex dynamics of choice due to repeated exposures to extinction learning. *Animal Cognition*, 24(6):1279–1297.

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

Gallistel, C. R., Fairhurst, S., and Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13124–13131.

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

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, 65(5):695–705.

Hasselmo, M. E., Bodelón, C., and Wyble, B. P. (2002). A Proposed Function for Hippocampal Theta Rhythm: Separate Phases of Encoding and Retrieval Enhance Reversal of Prior Learning. *Neural Computation*, 14(4):793–817.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(1):189–206.

Káli, S. and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience*, 7(3):286–294.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *arXiv:1612.00796 [cs, stat]*.

Klein, S. B., Cosmides, L., Gangi, C. E., Jackson, B., Tooby, J., and Costabile, K. A. (2009). Evolution and Episodic Memory: An Analysis and Demonstration of a Social Function of Episodic Recollection. *Social Cognition*, 27(2):283–319.

Kosaki, Y., Lin, T.-C. E., Horne, M. R., Pearce, J. M., and Gilroy, K. E. (2014). The Role of the Hippocampus in Passive and Active Spatial Learning. *Hippocampus*, 24(12):1633–1652.

Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, 6(6):579–590.

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

Louie, K. and Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29(1):145–156.

Mahr, J. and Csibra, G. (2017). Why do we remember? the communicative function of episodic memory. *The Behavioral and Brain Sciences*, pages 1–93.

Maren, S., Phan, K. L., and Liberzon, I. (2013). The contextual brain: Implications for fear conditioning, extinction and psychopathology. *Nature Reviews Neuroscience*, 14(6):417–28.

Masters, D. and Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. *arXiv:1804.07612 [cs, stat]*.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Bower, G., editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, Inc., New York, NY, USA.

McEchron, M. D., Bouwmeester, H., Tseng, W., Weiss, C., and Disterhoft, J. F. (1998). Hippocampectomy disrupts auditory trace fear conditioning and contextual fear conditioning in the rat. *Hippocampus*, 8(6):638–646.

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

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

Moser, E. I., Kropff, E., and Moser, M.-B. (2008). Place Cells, Grid Cells, and the Brain's Spatial Representation System. *Annual Review of Neuroscience*, 31(1):69–89.

Nadel, L. and Moscovitch, M. (1998). Hippocampal contributions to cortical plasticity. *Neuropharmacology*, 37(4-5):431–9.

Nairne, J. S. and Pandeirada, J. N. S. (2016). Adaptive Memory: The Evolutionary Significance of Survival Processing. *Perspectives on Psychological Science*, 11(4):496–511.

Nairne, J. S., Thompson, S. R., and Pandeirada, J. N. S. (2007). Adaptive Memory: Survival Processing Enhances Retention. *Journal of experimental psychology*, 33(2):263–273.

O'Kane, G., Kensinger, E. A., and Corkin, S. (2004). Evidence for semantic learning in profound amnesia: An investigation with patient H.M. *Hippocampus*, 14(4):417–425.

Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., and Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *eLife*, 4(JUNE2015):e06063.

Parra-Barrero, E., Diba, K., and Cheng, S. (2021). Neuronal sequences during theta rely on behavior-dependent spatial maps. *eLife*, 10:e70296.

Rosenbaum, R. S., Köhler, S., Schacter, D. L., Moscovitch, M., Westmacott, R., Black, S. E., Gao, F., and Tulving, E. (2005). The case of K.C.: Contributions of a memory-impaired person to memory theory. *Neuropsychologia*, 43(7):989–1021.

Smith, A. C., Frank, L. M., Wirth, S., Yanike, M., Hu, D., Kubota, Y., Graybiel, A. M., Suzuki, W. A., and Brown, E. N. (2004). Dynamic Analysis of Learning in Behavioral Experiments. *The Journal of Neuroscience*, 24(2):447–461.

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

Suddendorf, T. and Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, social, and general psychology monographs*, 123(2):133–67.

Suddendorf, T. and Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, 30(03):299–313.

Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, 2(2):189–199.

Tulving, E. (1972). Episodic and semantic memory. In *Organization of Memory*, pages xiii, 423–xiii, 423. Academic Press, Oxford, England.

van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069.

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

Wiltgen, B. J., Sanders, M. J., Anagnostaras, S. G., Sage, J. R., and Fanselow, M. S. (2006). Context fear learning in the absence of the hippocampus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(20):5484–91.