nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00950-3

Sequential memory improves sample and memory efficiency in episodic control

Received: 6 October 2022

Accepted: 5 November 2024

Published online: 31 December 2024

Check for updates

Ismael T. Freire[1,3] ✉, Adrián F. Amil[1] ✉ & Paul F. M. J. Verschure[2] ✉

Deep reinforcement learning algorithms are known for their sample inefficiency, requiring extensive episodes to reach optimal performance. Episodic reinforcement learning algorithms aim to overcome this issue by using extended memory systems to leverage past experiences. However, these memory augmentations are often used as mere buffers, from which isolated events are resampled for offline learning (for example, replay). In this Article, we introduce Sequential Episodic Control (SEC), a hippocampal-inspired model that stores entire event sequences in their temporal order and employs a sequential bias in their retrieval to guide actions. We evaluate SEC across various benchmarks from the Animal-AI testbed, demonstrating its superior performance and sample efficiency compared to several state-of-the-art models, including Model-Free Episodic Control, Deep Q-Network and Episodic Reinforcement Learning with Associative Memory. Our experiments show that SEC achieves higher rewards and faster policy convergence in tasks requiring memory and decision-making. Additionally, we investigate the effects of memory constraints and forgetting mechanisms, revealing that prioritized forgetting enhances both performance and policy stability. Further, ablation studies demonstrate the critical role of the sequential memory component in SEC. Finally, we discuss how fast, sequential hippocampal-like episodic memory systems could support both habit formation and deliberation in artificial and biological systems.

The increasing popularity of deep reinforcement learning (DRL) has been driven in recent years by its ability to reach human-level performance in several domains that were traditionally considered hallmarks of human intelligence: for instance, beating world champions in games like chess and Go[1] or, more recently, in complex real-time multiplayer games like Starcraft[2] and DOTA[3]. However, to achieve such remarkable feats, these algorithms require orders of magnitude more data to learn from than humans[4]. This illustrates the difference in sampling efficiency between DRL and the human brain.

The sample inefficiency problem in DRL refers to the large amounts of data and episodes that these methods require to reach the level of performance of humans[5]. In the case of the classic Atari games, DRL systems require millions of samples to reach human-level results[6]. In recent scenarios involving more complex tasks, the numbers of samples can reach up to several billion[7]. A number of sources of inefficiency and slowness in DRL have been already identified[8]. One pertains to an intrinsic feature of such systems: the gradient-based updates that slowly drive the learning of both policy and value functions limit the speed at which these systems can achieve optimal performance.

Recently, variants of episodic reinforcement learning (ERL) algorithms have been proposed to overcome the intrinsic limitations of the gradient-based methods of DRL, in particular the introduction

[1]Donders Institute for Brain Cognition and Behaviour – Centre for Neuroscience (DCN-FNWI), Radboud University, Nijmegen, the Netherlands. [2]Alicante Institute of Neuroscience, Department of Health Psychology, Universidad Miguel Hernandez de Elche, Elche, Spain. [3]Present address: Institute of Intelligent Systems and Robotics, Sorbonne University, Paris, France. ✉e-mail: ismael.freire@donders.ru.nl; adrian.fernandezamil@donders.ru.nl; pverschure@umh.es

of a memory system that allows the algorithm to make use of previously successful experiences to speed up the learning of an optimal policy. Within ERL, two main approaches have been followed: first, enhancing and bootstrapping the offline learning capacity of a Deep Q-Network (DQN) by replaying past experiences stored in a memory buffer[9–12]; and second, using the stored events in memory for direct control by generating the policy directly from the memory buffer[13–15]. Both of these solutions are partially inspired by some features of the hippocampal episodic memory in biological systems, in particular the known phenomenon of the acquisition, retention and replay of stored behavioural sequences[16].

Although capable of improving the sample efficiency of DRL, ERL methods also face a problem of memory efficiency. The efficient use of a limited memory buffer capacity is especially critical when embedding such algorithms in embodied systems such as robots that face strict real-time computational and storage limitations[17]—a problem that also brains must deal with[18,19]. However, studying the effects of incorporating memory constraints in ERL models is rarely considered[20], with the recent exception of ref. 15, which studied the role of forgetting in episodic control agents in two-dimensional discrete grid-world settings.

Thus far, ERL models do not capitalize on the sequential nature of the temporal unfolding of events that are reflected in the structure of hippocampal memory[21–23]. Notably, states in ERL are treated as being independent samples of a distribution of world states[13–15] instead of being experienced, stored and retrieved in a sequential manner due to the embodiment of the agent and the continuity of real-world interaction[21,24]. Indeed, the idea that integrated episodes may guide behaviour, echoing the efficiency of human episodic memory, has been supported by neuroscientific evidence[25–27]. These studies suggest that behaviour is informed by compression of sensory data into integrated episodes, reflecting trajectories through environmental and cognitive spaces, rather than isolated state–action pairs. The implications of treating experiences as integrated sequences have been discussed in the literature, emphasizing the importance of sequentiality in both animal and human cognition[28–30].

Based on this evidence, we propose that the performance of ERL methods can be further improved by incorporating additional aspects of mammalian episodic memory[22,31,32]. In particular, we suggest that the hippocampal memory system implements core features that solve the sampling efficiency challenge. Notably, the hippocampus has been argued to compress sensory data to generate efficient representations of the world's state[32]. Beyond this autoencoder-like function that explains many features of hippocampal dynamics[33,34], we also consider the conjunctive nature of event representations[35] and their sequential scaffolding and chaining into coherent, repeatable and goal-oriented memory episodes[21,36]. This sequential linking of events is in turn enabled by a particular winner-take-all (WTA) selection mechanism[37] that filters the unfolding sequence of behavioural events[38]. These sequences are time-multiplexed at faster neuronal timescales and maintained in short-term memory via nested frequencies or the theta-gamma code[18], which allows for fast long-term consolidation and retrieval and serves key cognitive functions such as mind-travel at decision points[39] and sequential decision-making[40,41]. Further, empirical evidence points to the prospective replay of rewarded sequences[42] playing a key role in bootstrapping learning and biasing decision-making towards previously successful actions[36,43] while they are coordinated by the active pursuit of goals by the agent[44]. Here we incorporate these features to enhance episodic control of reinforcement learning (RL) agents, focusing on the sequential binding of events into goal-oriented episodes and the way the retrieval of those affects learning dynamics.

In summary, in this Article, we explore the role of sequentiality, memory capacity and forgetting in ERL. We present a new algorithm, Sequential Episodic Control (SEC), that leverages the sequential nature of stored experiences for direct control. We demonstrate how imposing a sequential inductive bias on the retrieval of memories for action selection favours both overall performance and sample efficiency in several naturalistic foraging benchmarks. Moreover, we also show how this sequential inductive bias enhances memory efficiency by an order of magnitude compared to the same episodic control model without such a feature. Finally, we show that forgetting mechanisms also enhance both efficiency and policy stability.

Background and related work

Following Thorndike's law of effect, the goal of an RL agent is to maximize reward through its interaction with the environment[45]. This is usually operationalized as maximizing the expected discounted return $R_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$, where T is the length of an episode and $\gamma \varepsilon (0, 1]$ is the discount factor. Given state $s_t \varepsilon S$ of the environment, the agent takes action $a_t \varepsilon A$ following its policy $\pi(s_t, a_t)$, which brings about a reward $r_{t+1} \varepsilon R$ leading to a state of the environment s_{t+1}. In Q-learning[46], the agent learns the action-value function $Q^{\pi}(s, a) = \mathbb{E}[R_t | s_t = s, a]$ by computing the expected rewards obtained by acting on a given state. In DRL methods, such as DQN[6], this function is parametrized by a deep neural network to approximate the optimal action-value function.

Most of the work on ERL has tried to improve the sample efficiency of parametric models like DQN, either by bootstrapping their learning through an extended memory system or by value propagation methods that capitalize more on the experiences that yielded higher rewards in the past. The so-called episodic memory DQN adds a memory buffer parallel to a DQN and shows that having coordinated systems leads to faster reward propagation and higher sample efficiency as compared to the standard DQN[11]. Conversely, the Episodic Backward Update model propagates the value of a state to its previous states after sampling a complete episode[12]. This modification allows the Episodic Backward Update model to achieve the same mean human normalized performance on several Atari benchmarks as DQN while using only 5% of the data. An alternative model called ERL with Associative Memory (ERLAM) stores the different behavioural trajectories in a graph instead of a dictionary and associates different nodes to create an instance-based reasoning model that is more sample efficient[10]. Indeed, this follows an earlier proposal by Kubie on the graph-like features of hippocampal memory[47], which has been successfully translated to robot models of foraging[48].

In contrast to ERL approaches that capitalize on different forms of episodic memory to accelerate learning, episodic control models such as Model-Free Episodic Control (MFEC) remove all gradient-based methods using a non-parametric instance-based way of learning[13]. MFEC records rewarding experiences in a tabular memory and follows a policy that capitalizes on those stored events. MFEC updates its action-value estimates by storing in memory the highest Q-values experienced in a state. When encountering a state, the system consults this memory and picks the state–action pair that gave the highest reward. When faced with new states, their value is approximated by averaging the action values of the k state neighbours, using Euclidean distance as a similarity metric. Neural Episodic Control builds upon MFEC by adding a so-called differentiable neural dictionary that stores slow-changing state representations and fast-updating value estimates, retrieving those values for efficient action selection by using context-based lookup[14].

The SEC model that we present in this Article (Fig. 1) departs from the previous literature on episodic control in several aspects. First, it considers state–action pairs as integrated representational primitives that reflects hippocampal coding[35]. Second, it stores the complete sequence of state–action pairs (that is, events) leading to goal states (for example, rewards), conserving their serial order, instead of storing world states and actions as isolated memory elements. A partial exception here is ERLAM, which does not treat events as being completely independent. However, the main difference between SEC and ERLAM is that whereas the latter builds a graph based on the state transitions of the stored experiences to bootstrap the learning of a parametric

https://doi.org/10.1038/s42256-024-00950-3

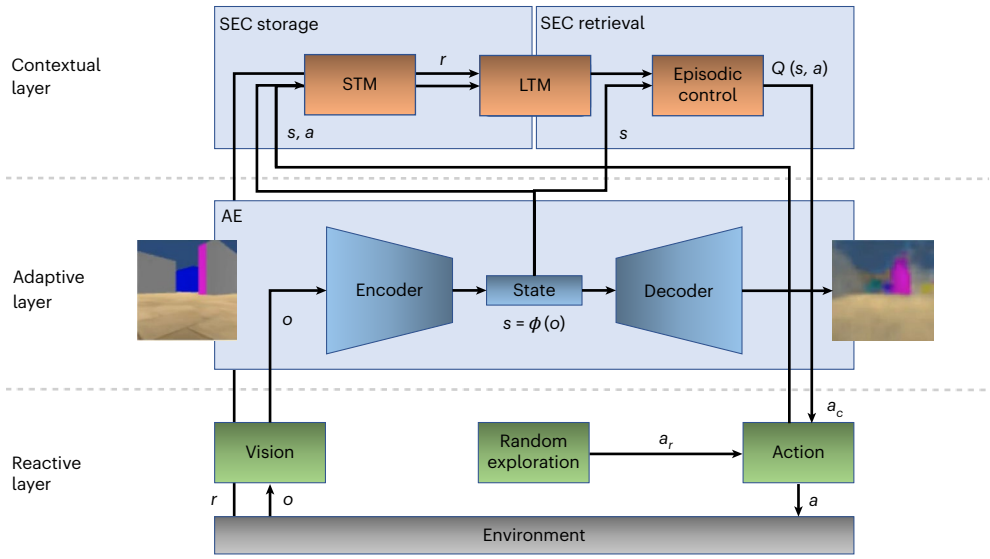


Fig. 1 | SEC architecture. Following the DAC framework (see ref. 23 for a review), SEC can be functionally divided into three layers: reactive, adaptive and contextual. The reactive layer (green) implements a predefined random exploration algorithm. The adaptive layer (blue) acquires states of the world through an autoencoder (AE), whereas the contextual layer (red) comprises short- and long-term episodic memory buffers and an action-selection algorithm.

RL agent, SEC stores memories in a sequential goal-oriented manner, conserving the temporal structure of action and using this memory buffer directly for action selection and control.

Like other episodic control models[13,14], SEC also follows a non-parametric approach as it uses an episodic tabular memory to store previously reinforced experiences and their predecessor states (for example behavioural sequences) to guide decision-making when encountering similar states. However, SEC deals with memory retrieval in a different way, through a combination of a perceptual similarity metric and a WTA mechanism—akin to attractor-based dynamics in perception (for example, ref. 49). Moreover, SEC computes the action-value function of a given state based on the combination of three factors: perceptual similarity between perceived and retrieved states, sequential bias between memory states and discounted reward value (Fig. 2). All these aspects are explained in detail in Methods.

To assess the effectiveness of the SEC model and the impact of its unique features, we conducted a series of four experiments (see Methods for a detailed description of the model and the experimental setup). Experiment 1 benchmarks SEC against established state-of-the-art algorithms across four challenging tasks from the Animal-AI testbed to evaluate its overall performance and sample efficiency (Fig. 3). In the subsequent experiments, we focused on the Double T-Maze task to investigate specific properties of SEC. Experiment 2 examines how varying memory capacity affects the learning efficiency and performance stability of SEC. Experiment 3 explores the contribution of each component within the model's valuation function—perceptual similarity, sequential bias and discounted reward value—to the decision-making process. Finally, Experiment 4 investigates the effect of different forgetting mechanisms on SEC's ability to adapt and maintain performance over time.

Results

Improved episodic control via sequentiality

Experiment 1 evaluated the SEC model across four benchmarks: Double T-Maze, Object Permanence, Cylinder and Detour tasks within the Animal-AI environment (Fig. 4). The performance of SEC was compared against several state-of-the-art models including Non-Sequential Episodic Control (NSEC), MFEC with an autoencoder (MFEC-ae), MFEC with random projections (MFEC-rp), DQN and ERLAM. A comprehensive statistical analysis is available in Supplementary Section 3.

In the Double T-Maze task, the SEC model demonstrated superior performance, achieving higher average rewards faster and sustaining those rewards over time compared to the other models. This benchmark emphasizes the model's proficiency in tasks requiring the encoding and retrieval of sequential information to navigate toward a goal.

For the Object Permanence task, which tests the model's ability to remember and act upon the location of unseen objects, SEC again outperformed the competing algorithms, showing a rapid increase in average rewards. This indicates the model's effectiveness in scenarios where indirect cues must guide decision-making. The results of DQN and ERLAM seem to indicate that they have fallen into the local minima of capturing the small visible reward (+1) instead of pursuing the hidden but greater reward (+3), which was more difficult to find.

The Cylinder task, requiring the discernment between opaque and transparent obstacles, saw a similar trend, with the SEC model matching or outperforming other models, albeit with a closer margin. The results suggest that although SEC is adept at tasks requiring visual discrimination, the advantage is less pronounced in this context, with several models performing equally well.

In the Detour task, designed to assess the ability to plan and execute a path around an obstruction, SEC outshone all other models notably, as evidenced by the steep and consistent rise in average reward. The SEC model's performance in this task underscores its capacity for handling complex spatial navigation challenges.

Across all tasks, the NSEC model performed notably worse than its sequential counterpart, emphasizing the critical role of sequentiality in the SEC model's success. The MFEC variants, although competitive, did not reach the performance peaks of SEC. DQN and ERLAM trailed behind, particularly in tasks that demanded more sophisticated episodic memory capabilities.

The results from Experiment 1 underscore the SEC model's robustness and versatility across a variety of tasks that challenge different cognitive skills. The consistent outperformance of SEC over NSEC and other models across all benchmarks highlights the efficacy of SEC mechanisms in complex navigational and cognitive tasks.

Efficient performance under memory constraints

To analyse the effect that limited memory capacity imposes on the SEC algorithm, we tested the standard (SEC) and ablated (NSEC) versions of the model with varying long-term memory (LTM) capacities

https://doi.org/10.1038/s42256-024-00950-3

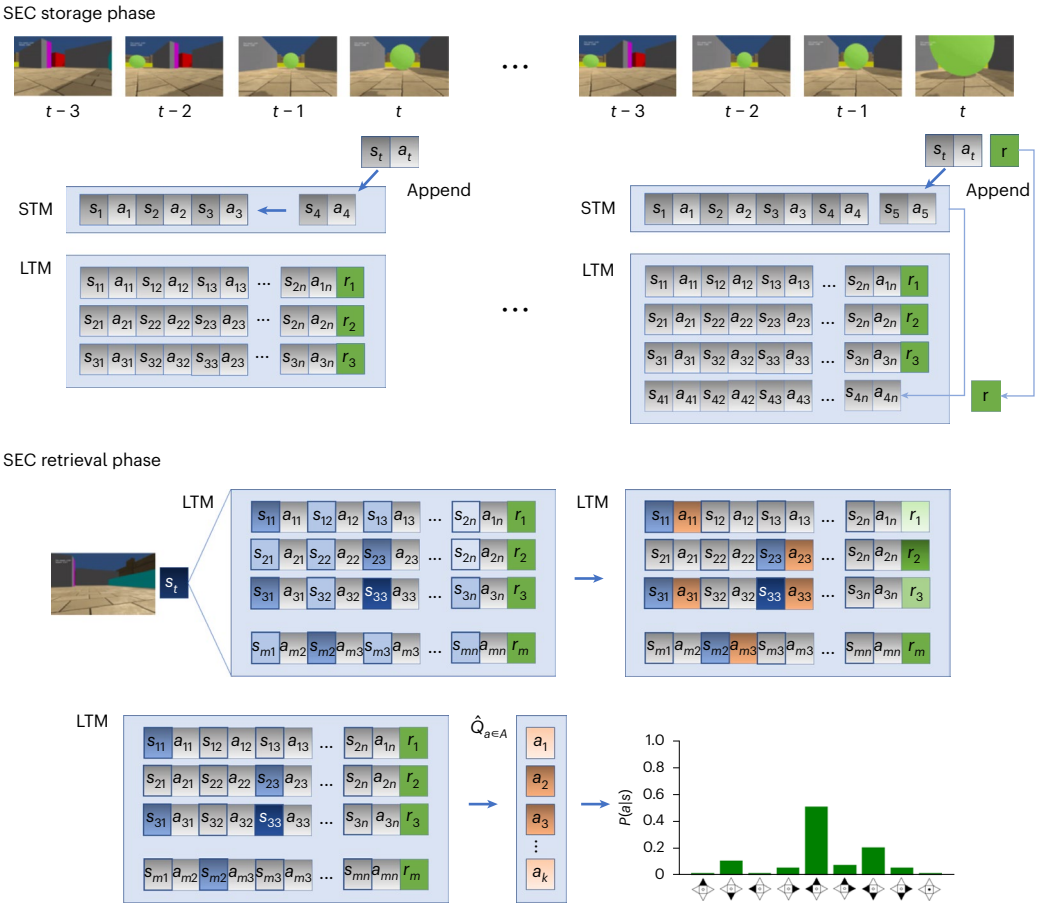


Fig. 2 | SEC memory storage and retrieval phases. During the storage phase, state–action couplets are stored in STM on a FIFO basis at every time step (top-left). Upon encountering a reward, the content of the STM is transferred to the LTM buffer, along with the reward value (top-right). During the retrieval phase, first, following equation (1), the current observed state is compared with the stored states in the LTM and the most similar ones are retrieved (middle). After that, following equation (5), the action-value function for the observed state is computed by taking the actions attached to the retrieved states along with their discounted relative reward values (bottom). This creates a probability distribution in action space from which the final action is drawn. See Methods for more details.

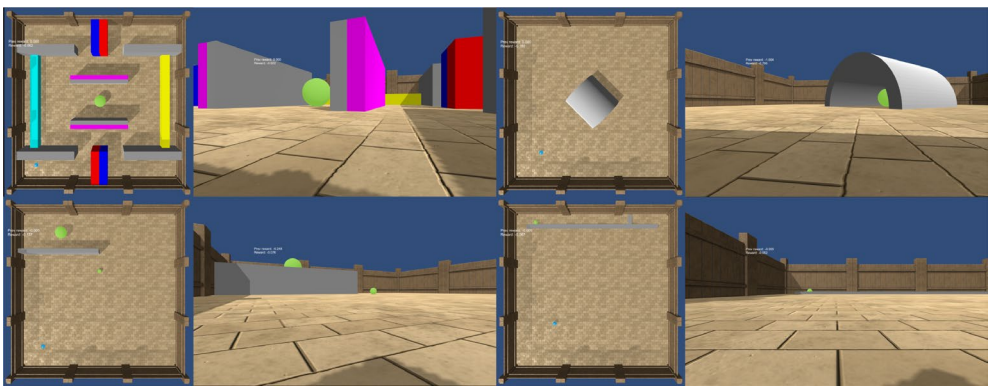


Fig. 3 | Illustration of the benchmarks from the Animal-AI environment. The displayed environments are Double T-Maze (top-left), Cylinder (top-right), Object Permanence (bottom-left) and Detour (bottom-right). See ref. 75. For each benchmark, the left side of the panel provides a third-person, bird's-eye view of the environment, whereas the right side offers the first-person perspective as seen by the agent navigating the scenario.

(that is, a fixed LTM memory of 125, 250, 500 and 1,000 sequences) in the Double T-maze task (see Supplementary Section 3 for a detailed statistical analysis). Therefore, once the memory was filled, no further memories could be stored. As in Experiment 1, we performed 20 simulations of 5,000 episodes (approximately 5 million frames) per model and memory condition.

The results of Experiment 2 show that SEC reaches a plateau reward acquisition upon arriving at 1,000 episodes for all tested memory conditions (Fig. 5). In terms of comparative performance, the SEC model obtains a clear advantage with respect to its non-sequential version, NSEC, across all memory conditions. As before, SEC not only reaches higher levels of accumulated reward but also reaches

https://doi.org/10.1038/s42256-024-00950-3

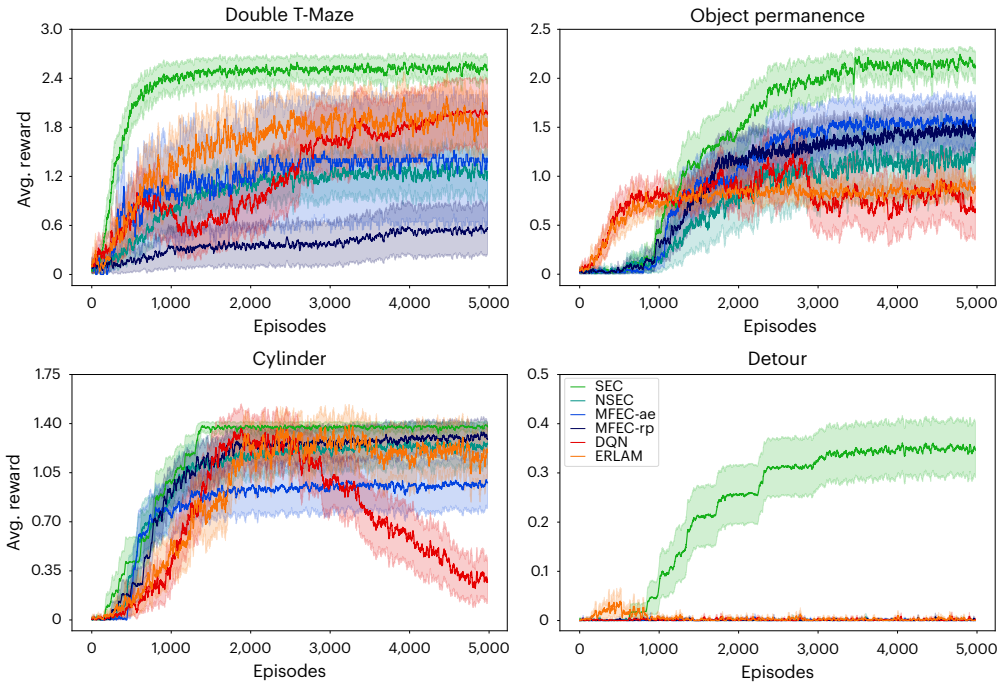


Fig. 4 | Comparative performance of SEC against benchmark algorithms DQN, MFEC, ERLAM and NSEC. The presented results encompass four distinct Animal-AI benchmarks: Double T-Maze (top-left), Cylinder (bottom-left), Object Permanence (top-right) and Detour (bottom-right) tasks. For clarity and statistical robustness, average performance metrics were calculated using a sliding window encompassing 20 episodes (20,000 frames). The error bars denote the standard error (s.e.) to provide a measure of the variability in the dataset. Avg., average.

asymptotic performance more quickly. The difference in convergence rates to fill the sequential memory illustrates the bootstrapping effect of behavioural feedback. This difference increases between the two versions of the model when the memory capacity limit is increased. In all cases, the performance plateaus shortly after reaching the memory capacity (as indicated by the vertical bars in Fig. 5). Importantly, the decrease in mean entropy shows that SEC stabilizes its policy, whereas NSEC does not, maintaining a high policy entropy in all conditions despite increasing reward acquisition. Thus, the sequential bias component of SEC allows the agent to achieve a higher level of accumulated rewards and also provides a behaviour stabilization mechanism.

The analysis of the policy entropy of the SEC model shows similar patterns to the reward accumulation dynamics, indicating that agents using the SEC model quickly converge to a stable policy. The entropy of the NSEC model, on the other hand, is much higher than SEC in all conditions. It is also important to note that the degree and speed of convergence are notably affected by memory capacity. In both episodic control models, there is a consistent reduction in entropy when the memory capacity is increased. This effect indicates an improvement in the stabilization of the policy that also is reflected in the increased reward acquisition.

Finally, we systematically compare the performance of each memory condition of NSEC against SEC to have a detailed estimate of how much performance increase is obtained using the sequential bias. The results shown in Extended Data Fig. 1 present a table containing this analysis. The data show that the SEC model can reach a similar level of performance to its non-sequential version, NSEC, with an order of magnitude lower memory capacity. In other words, the SEC model with a fixed memory of 125 units achieves 1.2 times the performance of NSEC with 1,000 memory units. Moreover, when both models are evaluated with equal memory limitations, the performance difference between SEC and NSEC increases as the memory capacity is reduced: SEC obtains 1.5 times the performance of NSEC with a limit of 1,000 memories, and it scales up to 2.5 times with 125 memory units. These results show the key advantage of SEC algorithms because the sample efficiency of SEC also translates into memory efficiency and, therefore, reduces computational costs. This demonstrates the favourable scaling properties of SEC and similar solutions.

Limited impact of valuation components

In this study, we perform several ablations on SEC's valuation function (equation (5)) to analyse the differential effect of its three components: the eligibility score $G_{i,j}$, the distance to the goal and the relative reward. The results of the ablation studies are depicted in two graphs in Extended Data Fig. 2, one for each set of studies: the Single Mechanism Ablation and the Double Mechanism Ablation (see also Supplementary Section 3 for a detailed statistical analysis).

In the Single Mechanism Ablation study, omitting the eligibility score $G_{i,j}$ (SEC-noGi) resulted in a slight decrease in average reward compared to the full SEC model, suggesting that although $G_{i,j}$ contributes to performance, its absence does not drastically impair the model. When the distance to the goal was not considered (SEC-noDist), we observed a more pronounced drop in the average reward, indicating that spatial considerations play a more prominent role in SEC's success. Finally, the removal of the relative reward component (SEC-noRR) had an intermediate impact, more than SEC-noGi but less than SEC-noDist, underscoring its importance but also the model's resilience to its absence. The entropy plots correlate with these findings, showing that the models with a single mechanism removed generally maintained similar uncertainty levels in their action selection.

The Double Mechanism Ablation results indicate that the SEC model retains competitive performance even when reduced to a single operational mechanism, be it the eligibility score $G_{i,j}$ (SEC-soloGi), distance to the goal (SEC-soloDist) or relative reward (SEC-soloRR). Throughout the ablation studies conducted in this experiment, it is noteworthy that although the modified versions did not achieve the same high level of results as the complete SEC model, they consistently

https://doi.org/10.1038/s42256-024-00950-3

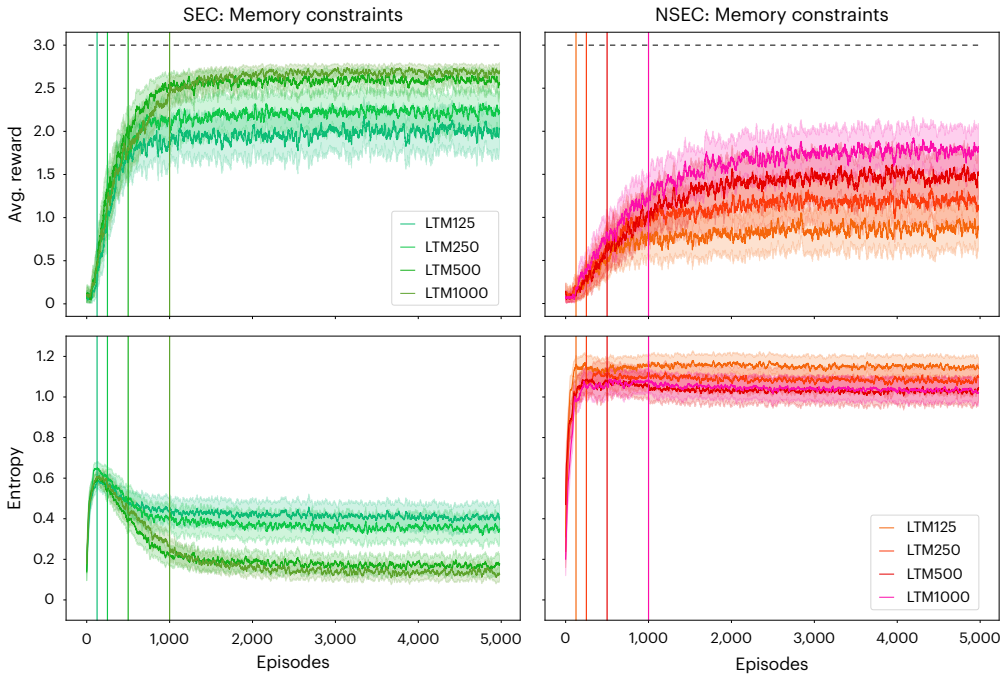


Fig. 5 | Effect of memory constraints on SEC and NSEC in the Double T-Maze. Top panels, mean reward per episode accumulated by SEC (left) and NSEC models (right). Bottom panels, mean entropy on the episodic policy, computed as the average of the entropies of the probability distributions derived from $\widehat{Q}^{\mathrm{LTM}}_{s,a}$ at every time step of the episode. Vertical bars represent the average episode around which the memory was filled. Average values were computed using a sliding window of 20 episodes. Error bars represent s.e. Avg., average.

outperformed the state-of-the-art models evaluated in Experiment 1 within the Double T-Maze task.

The results from these ablation studies suggest that SEC's components contribute to its performance synergistically. Removing individual or paired components did not significantly degrade SEC's performance (Supplementary Section 3), highlighting the robustness of the architecture across different configurations. However, it remains advantageous to maintain all components for optimal performance.

In contrast with these results, NSEC (the variant of SEC where the sequential bias is removed) obtained significantly inferior results compared to all the other ablated versions and the complete SEC model. This marked difference highlights the pivotal role of sequentiality within SEC. The sequential bias is evidently a critical factor that contributes to the SEC's optimal performance in complex tasks such as the Double T-Maze, and its absence is detrimental to the model's success. In essence, although the SEC can function above state-of-the-art standards without certain components, it is the integration of sequentiality that propels it to achieve the best results.

Forgetting mechanisms show small yet positive impact

The results from Experiment 4, depicted in Fig. 6, indicate that the integration of forgetting mechanisms has a nuanced impact on the performance of episodic control models (see also Supplementary Section 3 for a detailed statistical analysis).

For the SEC model, both forgetting mechanisms (SEC-fifo and SEC-rwd) show slightly enhanced performance over the original SEC. In the case of NSEC, the addition of forgetting (NSEC-fifo and NSEC-rwd) also results in improved performance, with the reward-based forgetting variant (NSEC-rwd) standing out with a higher average reward than NSEC-fifo.

The entropy plots reveal that the addition of forgetting mechanisms leads to reduced entropy in the policy distribution, suggesting more stable behaviour. This effect is more pronounced in the reward-based forgetting variants of both SEC and NSEC, indicating that prioritizing the forgetting of lower-reward experiences leads to a more focused and effective policy.

Overall, the performance enhancement from forgetting is evident but not as substantial as the improvement gained from sequential information processing, as seen when comparing the full SEC model to the NSEC variants with forgetting. This comparison underscores the critical role of sequentiality in the model's success. Nonetheless, the combined effects of forgetting and sequentiality contribute to the highest overall performance and the lowest entropy, as demonstrated by the SEC-rwd model. This suggests that the benefits of forgetting are additive when paired with the sequential chaining capability, resulting in an even more powerful episodic control model.

In summary, the findings indicate that introducing first-in, first-out (FIFO) or reward-based forgetting mechanisms in SEC and NSEC provides modest performance gains under specific memory constraints. However, the lack of statistical significance suggests that these forgetting mechanisms do not substantially alter the models' performance in the studied Double T-Maze task. However, the incorporation of sequentiality significantly does enhance the model's capabilities, with the combination of both features yielding the best results (Supplementary Section 3).

Discussion

Episodic control models seek to overcome the sample-inefficiency problem in RL by implementing a form of instance-based learning that capitalizes on the storage of previously successful experiences to quickly learn optimal strategies. Although they are generally inspired by the mammalian hippocampus, those models treat stored experiences as isolated units without taking into account the sequential nature in which those events unfolded in real time. In this Article, we show that the omission of this feature has implications in terms of both sample efficiency and memory efficiency for episodic RL. We present an episodic control algorithm, SEC, that addresses both issues by incorporating a more complete picture of the hippocampal function. Crucially, SEC stores complete behavioural sequences in its LTM and

https://doi.org/10.1038/s42256-024-00950-3

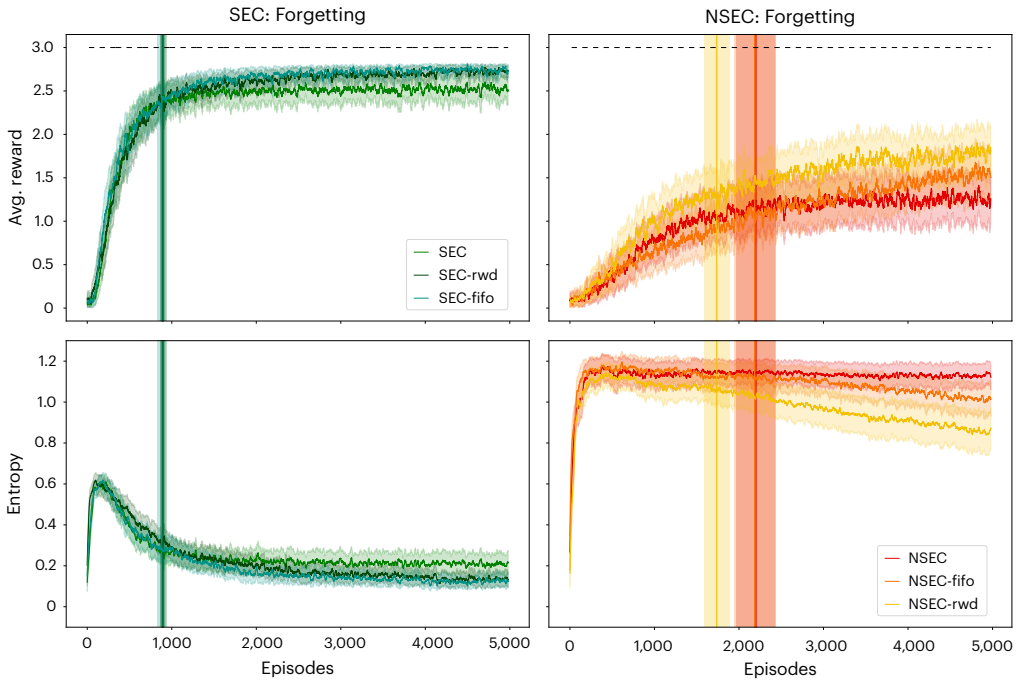


Fig. 6 | Forgetting enhances episodic control performance and policy stability in the Double T-Maze benchmark. Model comparison of episodic control models with forgetting (SEC-FIFO, dark green; NSEC-FIFO, orange), without forgetting (SEC, green; NSEC, red) against the MFEC (blue) benchmark. Left, average performance per episode. Right, entropy over the policy for SEC and NSEC models. Lower entropy values imply greater policy stability. Average values were computed using a sliding window of 20 episodes. Error bars represent s.e. Avg., average.

takes into account the sequential nature of its stored experiences for action selection. We show how SEC can reach sample-efficient performance compared to episodic controllers that store state–action pairs as independent elements in memory, such as MFEC. To address memory efficiency, we systematically study the effect of memory limitations and forgetting in episodic control. We demonstrate that constraining the capacity of the memory buffer substantially affects the performance and stability of episodic control algorithms. The results show how the SEC model outperforms an episodic controller that does not take into account the sequential structure of the stored memories and that this difference in performance is increased when greater memory limitations are imposed. Our results also indicate that forgetting generally improves performance and policy stabilization on episodic control, but sequentiality plays a major role in comparison. Taken together, this work seeks to contribute to the available proposed solutions to the sample-inefficiency problem in RL by taking inspiration from empirical and theoretical research in the cognitive sciences. This work also extends the possibilities of episodic control algorithms by showing how leveraging the sequential nature of the stored experiences improves sample and memory efficiency during learning.

In contrast with standard episodic control models, which treat their memory units as disconnected events, SEC emphasizes the learning of complete goal-oriented behavioural sequences. The incorporation of this hippocampus-based architectural bias bootstraps the formation of a behavioural feedback loop[50], where the agent will tend to retrieve and orderly follow previously successful behavioural sequences, thus resulting in the generation of similar behavioural outcomes, which in turn will lead to the acquisition of similar sequential memories. This memory-driven behavioural feedback loop allows SEC to rapidly transition from an initial exploration phase to an exploitation phase in which the behaviour of the agent stabilizes over time while maximizing reward acquisition.

Within the framework of episodic control models, a theoretical distinction aligns with the traditional dichotomy in RL between model-based and model-free algorithms[51]. In this case, the distinction is made based on the structure of the stored memories. According to this perspective, both ERLAM and SEC can be classified as 'episodic model-based' control systems because they preserve the sequential structure of memory, thereby maintaining a certain order of events. In contrast, pure MFEC controllers, such as NSEC, do not utilize the temporal structure of stored memories, or they discretize and store events as isolated instances, as in MFEC[13].

Although SEC and ERLAM can be both categorized as 'episodic model-based' systems due to their retention of structured event sequences, they diverge notably in application. The primary distinction is that SEC employs these sequences for direct control, whereas ERLAM utilizes the value estimates from its associative memory as a target signal for a DQN. Additionally, ERLAM actively engages in an associative process where a model is constructed from stored memories, creating a graph network over the visited states and effectively forming a world model akin to model-based RL methods[10]. A promising avenue for future research involves a comparative analysis of SEC with an episodic model-based algorithm that incorporates ERLAM's associative memory for direct action selection. Such a study could elucidate the distinct contributions and roles of episodic memory and associative mechanisms within the decision-making process, providing deeper insight into the interplay between these cognitive systems in guiding behaviour.

Conversely, although in this Article we have only shown the capacity of the SEC model for control, it could also be implemented like ERLAM's associative memory: that is, in combination with other learning algorithms to bootstrap the learning phase from a task-relevant set of successful samples, as shown in previous work[11,12]. Moreover, such an application of the combination of control and batch learning does not need to be mutually incompatible. Indeed, it could be possible to use episodic control models to drive an agent's behaviour during the initial stages of learning, by rapidly latching onto successful experiences, whereas a slow-learning algorithm such as a DQN could use the growing set of successful sequences obtained by the episodic controller to speed up the acquisition of an optimal policy through offline batch learning and replay[42,52].

Episodic control algorithms, by their inherent design, resonate deeply with the instance-based learning (IBL) theory from cognitive science[53,54]. The IBL theory postulates that decisions are made based on the recall of specific past episodes or instances, rather than by aggregating across them[53]. Similarly, episodic control models prioritize recent experiences, leveraging specific memories to make informed decisions[13,55]. These models can be seen as an emulation of the cognitive processes underlying episodic memory, where past instances are recalled to provide context and guide current decision-making[31,54].

In contrast with standard RL methods, the main driver of policy learning in episodic control is the constant acquisition and retrieval of new sequential memories. In other words, the behavioural policy of an episodic controller is implicitly updated by the memories it forms during its interaction with the environment. Precisely due to this factor, the performance of episodic control algorithms is sensitive to memory constraints (that is, limits on its memory capacity) and therefore can also benefit from the implementation of adequate forgetting mechanisms, as shown in this work and studied in ref. 15. Therefore, in the absence of forgetting, when an episodic controller fills its memory capacity, its learning and performance will tend to stagnate because it will no longer be able to store new memories.

Choosing when and what to forget are fundamental challenges for the generalization, robustness and long-term performance of episodic control algorithms. Of course, problems regarding memory limitations and efficiency might not be relevant for models using unbounded memory buffers. Nonetheless, such issues become fundamental in the development of embodied artificial agents. In robotics and embodied AI, the question of autonomy is central, and hence the need for fast learning is intertwined with the optimization of energetic and computational demands, making solutions like SEC a promising avenue for progress.

Episodic control algorithms like SEC, with their ability to rapidly exploit past successful strategies, can be particularly beneficial in multi-agent environments where agents must adapt to continually changing policies of other learning agents[56,57]. Recent work has further demonstrated the potential of shared episodic memories in enhancing collaborative tasks. High-fidelity social learning via shared episodic memories has been shown to substantially improve collaborative foraging performance among agents[58]. By sharing detailed behavioural sequences, agents can learn from each other's experiences more effectively, leading to more distributed and efficient resource collection.

Moreover, the use of sequential memory can aid in the formation of better internal models of other agents, a task that is challenging for traditional RL approaches[59–62]. By building a sequential memory of past social interactions, agents can support the virtualization of the 'other', which is essential for successful multi-agent interactions and has practical applications in human–robot collaboration[63,64].

A notable challenge faced by this work, as well as by episodic control models more broadly, lies in their so-far-limited domain of application[13]. Episodic control excels in environments with deterministic state transitions and rewards and in situations where an agent can benefit from memories of similar past experiences[13,14]. However, further work is required to assess how well this type of algorithm performs in more complex, non-stationary settings and how well it deals with generalization. One potential solution for managing non-stationary environments in episodic control is the development of more sophisticated forgetting mechanisms that can selectively prune outdated memories of events or states that are no longer relevant due to sudden changes in the environment. The capacity of episodic controllers to generalize their acquired knowledge to different tasks and environments depends on their ability to adequately recruit their memories for use in similar perceptual states. In this work, we show that SEC can perform very well in high-dimensional state spaces, where it can utilize its memories to generalize previously successful behaviours to similar states. SEC addresses state generalization by building internal state representations through a convolutional autoencoder and selecting memories based on the perceptual similarity between observed and stored states, but other options exist[15]. Future research into different methods for building such internal representations might play a key role in developing more efficient episodic control algorithms.

In nature, complex organisms make use of different learning, memory and decision systems adaptively[65,66]. A classical distinction in cognitive science is between deliberate and habitual behaviour. On the one hand, deliberate, model-based planning allows the assessment of multiple courses of action and their potential consequences. However, the computations involved in planning take a long time and might not be the best solution under time-pressure conditions. In addition, it requires that a model of the environment has been formed in the first place. On the other hand, habitual or model-free decision-making systems are much faster and therefore more suitable for contexts in which fast decisions need to be made. They do not rely on the acquisition of a model; however, they take much longer to be learned[13,67]. Besides these two modes of operation, some researchers have argued that episodic control represents a third type of system that could also play a role in generating adaptive behaviour[55]. An episodic controller operates by remembering the action that led to the best outcome in a given situation. It is computationally lighter than model-based algorithms and does not suffer from too much uncertainty or noise due to the complex calculations involved in forward search. Moreover, it takes much less time to acquire than a habit. Therefore, in situations when an animal is exploring a new environment, and no model, policy or habit has yet been formed, relying on the fast IBL provided by episodic control systems like SEC might be of critical importance.

In the framework proposed by ref. 55, each of these learning systems (model-based/deliberate, model-free/habitual and instance-based/episodic) has its trade-offs, derived from their intrinsic properties and inductive biases[55,68]. By their very nature, these control systems have an optimal performance at different stages of the learning process of the agent and could be operating at different timescales[8], as also demonstrated by the relation between the reactive, adaptive and contextual layers of the Distributed Adaptive Control (DAC) framework[69]. Understanding the regimes in which they optimally operate and how they combine to attain adaptive behaviour in physical and social environments is of paramount importance if we aim to build synthetic embodied cognitive systems able to exhibit distributed adaptive control in complex settings.

Methods

SEC

The SEC model is formulated in the context of the DAC theory of mind and brain[65]. DAC considers the brain a multilayer control system comprising reactive predefined behaviours, adaptive state-space encoding and memory-based action selection (that is, the reactive, adaptive and contextual layer, respectively; see Fig. 1). The SEC model is fundamentally built upon the idea that the hippocampus incorporates mechanisms for the encoding of both perceptual and action states and their further integration into goal-oriented sequences in memory, thus corresponding to the adaptive and contextual layers in the DAC framework[23]. Following this framework, SEC integrates the embedded states learned by its adaptive layer into the sequential memory system of the contextual layer. SEC's adaptive layer is composed of a convolutional autoencoder operating as an embedding function ϕ that builds compressed feature representations s_t of the observations o_t obtained from the environment, $s_t = \phi(o_t)$ (see Supplementary Section 2 for more details on the autoencoder architecture). SEC's contextual layer includes a short-term memory buffer, STM, and a long-term episodic memory system, LTM, combined with a memory-based action-selection algorithm (Algorithm 1). Finally, SEC incorporates a random exploration algorithm in its reactive layer to drive the initial exploration of the state space and acquire its first sequential memories.

Episodic control algorithms have two main functions: memory storage and memory retrieval (see Fig. 2 for a visual description). During the memory storage phase, the short-term memory buffer STM transiently stores the most recent sequence of state–action pairs (that is, events) encountered by the agent and is updated following a FIFO rule. Upon encountering a goal state (that is, reward), the sequence stored in the STM is then consolidated in the LTM along with its associated reward value (r_t).

During the memory retrieval phase, memories are selected based on their perceptual and behavioural relevance to the current state of the agent. Concretely, for a given state vector (s_t) with dimensionality D, an eligibility score is computed for each state–action pair (that is, memory) stored in LTM. The eligibility score $G_{i,j}$ of a memory is defined by its combined perceptual matching and sequential bias values, as expressed in equation (1). To calculate the perceptual matching, the perceived state (s_t) is compared to all the state representations stored in the LTM based on a similarity score—using a distance metric $d(s_t, s_{i,j})$ corresponding to the mean absolute error (equation (2)):

$$G_{i,j} = (1 - d(s_t, s_{i,j}))E, \quad \text{with } i,j \in \text{LTM}_{m\times n} \tag{1}$$

$$d(s_t, s_{i,j}) = \frac{1}{D}\sum_{k=1}^{D}|s_{t,k} - s_{i,j,k}| \tag{2}$$

The matrix E modulates the eligibility scores by keeping track of the recent history of retrieved state–action memories. At the beginning of each episode, E is initialized as a unit matrix with dimensions equal to the LTM ($m \times n$). Then, at every time step t, the value of every entry $E_{i,j}$ is increased by α only if the preceding state $s_{i,j-1}$ in the LTM was selected in time step $t-1$ (equation (3)). The selection is in turn indexed by a mask matrix M (explained below in equation (4)), with $M_{i,j} = 1$ for selected memories and $M_{i,j} = 0$ for the unselected ones. Then, unselected memories tend to slowly return to their initial unit values at a constant rate of decay β:

$$E_{i,j}(t) = E_{i,j}(t-1) + \alpha M_{i,j-1}(t-1) - \beta, \quad \text{with } E_{i,j}(t) = 1 \text{ if } E_{i,j}(t-1) < 1 \tag{3}$$

This mechanism implements a sequential inductive bias by enhancing the selection of consecutive memories within a sequence, thus favouring the retrieval of the entire sequence over time. Notably, this process reflects the sequential activation of hippocampal index neurons[70] during the phenomenon known as phase precession during hippocampal theta oscillations[38,71] and the consequential re-activation of cortical patterns representing the stored content of the events[22]. The sequential bias is further amplified through behavioural feedback[50], whereby the resulting bias in input sampling due to its memory-based action selection favours the storage of similar sequences of state–action pairs.

To complete the retrieval phase, all memories go through a selection process that determines M based on the eligibility scores $G_{i,j}$. Only those state–action pairs that surpass both absolute (θ_{abs}) and proportional (θ_{prop}) thresholds are retrieved and selected to contribute to action selection—see equation (4), where $H(x)$ is the Heaviside function and M is the resulting mask matrix introduced earlier. This procedure enforces a soft WTA mechanism akin to the competitive dynamics of the theta-gamma code in the hippocampus[37]:

$$M_{i,j} = H(G_{i,j} - \theta_{\text{abs}})H\left(\frac{G_{i,j}}{G_{\max}} - \theta_{\text{prop}}\right) \tag{4}$$

Lastly, in the action-selection phase, the action-value function $\hat{Q}_{a\in A}$ is computed following equation (5) given the selected memories, indexed by M from the memory retrieval phase. First, the value of each selected state–action pair $Q_{s,a}|_M$ is computed by using its eligibility score $G_{i,j}$ and relative discounted reward. The discounted reward is obtained by applying an exponential decay $e^{-d_{i,j}/\tau}$ to the relative reward of the corresponding memory sequence—that is, $r_i/r_{\max}$, where $r_{\max}$ is the maximum reward across all selected memories. In turn, the decay is based on a time constant τ and the distance $d_{i,j}$ from the corresponding state–action pair to the end of the sequence i in the LTM. This mechanism implements a relative reward valuation between the selected memories grounded on decision-making valuation processes in the brain[72–74]. Finally, the action-value function $\hat{Q}_{a\in A}$ is the result of summing across the state–action values of selected memories $Q_{s,a}|_M$ for every action a in the action space A:

$$\hat{Q}_{a\in A} = \sum_{i,j\in M} G_{i,j}\frac{r_i}{r_{\max}}e^{-d_{i,j}/\tau}\Big|_a \tag{5}$$

Then, the resulting action is selected by sampling the probability distribution generated by normalizing $\hat{Q}_{a\in A}$ so that Q-values across the action space sum up to 1. This method for computing the Q-values from relevant memory sequences favours the selection of actions that were taken in very similar states in the past while prioritizing those that are closer to potential high rewards.

Algorithm 1: Sequential Episodic Control.

```
STM: short-term memory
LTM: long-term memory
for each episode do
  Initialize empty STM
  t = 1
  while t < T and r_t = 0 do
    Receive observation o_t from environment
    Let s_t = φ(o_t)
    Retrieve relevant memories for state s_t via equation (1), equation (4)
    Estimate return for each action a via equation (5)
    Let a_t ← π(Q̂_{a∈A}(s_t))
    Take action a_t, receive reward r_{t+1}
    Append (s_t, a) to STM
    t ← t + 1
  end while
  if r_t > 0 then
    Append (STM, r_t) in LTM
  end if
end for
```

Experimental setup

This Article presents a sequence of four experiments aimed at evaluating the performance of the SEC model. Experiment 1 benchmarks the SEC model against established state-of-the-art algorithms across four challenging benchmarks from the Animal-AI testbed. Subsequent experiments focus on the Double T-Maze task to investigate the effects of memory capacity (Experiment 2), the contribution of individual components within the model's valuation function (Experiment 3) and the effect of forgetting mechanisms (Experiment 4).

Experiment 1 involves a comparative analysis of SEC against several state-of-the-art RL and episodic control models, including DQN[6], MFEC[13] and ERLAM[10]. To isolate the impact of sequentiality on performance, we introduced a control version of the model, NSEC, which omits the sequential bias matrix E in equation (1). In this version, the eligibility score $G_{i,j}$ is modified to exclude sequential information, resulting in the following simplified equation for NSEC:

$$G_{i,j} = 1 - d(s_t, s_{i,j}), \quad \text{with } i,j \in LTM_{m\times n} \tag{6}$$

Regarding the models' hyperparameters, a grid parameter search was performed to set the most-performing values for SEC. The same

values were also used for NSEC. A detailed account of the hyperparameters used by the SEC and NSEC models in this experiment can be found in Supplementary Section 1. For the MFEC algorithm, we reproduced the implementation described in ref. 13, with $k = 50$ giving the best performance. We implement both versions of MFEC: MFEC-rp, which uses random projections as the embedding function, and MFEC-ae, which uses an autoencoder for the embedding function. The SEC, NSEC and MFEC-ae models use the same autoencoder architecture. As in previous approaches[13], for each benchmark, we first train the autoencoder for 10,000 episodes (approximately 10 million frames) using random exploration; then weights were frozen for the experimental phase (see Supplementary Section 2 for more details). Regarding the DQN, we keep the standard setting for network architecture and hyperparameters as in ref. 6. Finally, for the ERLAM algorithm, we reproduce the algorithm following ref. 10 using the same reported hyperparameters (notably $\lambda = 0.3$).

To draw general performance comparisons, we test all the models (SEC, NSEC, DQN, MFEC-ae, MFEC-rp, ERLAM) in four benchmarks of the Animal-AI testbed[75,76]: the Double T-Maze, the Detour task, the Object Permanence task and the Cylinder task (Fig. 3):

- The Double T-Maze is a task with a sparse reward structure. This task requires the sequential encoding and retrieval of relevant visual cues to reach the reward location in a minimal amount of time. In each episode, the agent randomly starts at one of the corners, and it must reach the centre of the maze to obtain a reward. The reward at the centre is the only positive reward (+3) available in the environment. Due to the walls of the maze, the agent does not see the reward directly and needs to explore the maze first.
- The Object Permanence task involves food that moves out of sight that the agent needs to still attain. At the beginning of the episode, the agent can observe how a big reward (+3) falls into one of several holes of the maze until it is completely occluded. The agent needs to find the shortest path to reach the hidden reward while avoiding falling into the other holes.
- The Cylinder task includes either opaque or transparent cylinders. In this task, the agent needs to get inside the cylinder to reach a medium-sized reward (+2).
- The Detour task tests the ability to make a detour around an object to get food and assess the shortest path to the object. The wall is transparent but cannot be traversed, so the agent can perceive the reward from the other side. The small reward (+1) at the top corners is the only positive reward available in the environment.

In all these environments, agents receive at each step the first-person visual scene from the environment with a resolution of 84 by 84 pixels. The action space is composed of a two-dimensional vector of integers that allows values from 0 to 2. The first value of the vector represents the translation axis, and the second value is the left–right axis. We apply a standard frameskip of 4 to reduce computational requirements. At each step, the reward value decreases by 0.001 to promote efficient trajectories. An episode finishes when the agent gets the reward or 1,000 frames are reached.

Experiment 2 was designed to investigate the influence of limited memory capacity on the performance of the SEC algorithm. We compared the standard SEC with its ablated version, NSEC, under various LTM capacities in the Double T-Maze task. The memory capacities tested were fixed at 125, 250, 500 and 1,000 sequences, meaning that once these capacities were reached, no additional memories could be recorded. This setup allowed us to examine the efficiency and effectiveness of memory use within these algorithms.

In Experiment 3, we investigated the individual contribution of various mechanisms within SEC's valuation function to its overall performance (equation (5)). This was achieved through a series of ablation studies conducted in the Double T-Maze environment. We carried out two sets of ablation studies.

Single Mechanism Ablation: In the first set, we deactivated one mechanism at a time from SEC's decision process, specifically the eligibility score $G_{i,j}$, relative reward and distance to the goal. Each mechanism was inactivated by directly removing the corresponding term from the valuation equation (equation (5)), thereby eliminating its influence on action selection. The ablations were implemented as follows:

- SEC-noGi: The eligibility score term $G_{i,j}$ was removed from the valuation function, meaning actions were selected without accounting for the perceptual matching between the current state and the stored memories. The modified valuation function for this condition is

$$\hat{Q}_{a\in A} = \sum_{i,j\in M} \frac{r_i}{r_{\max}} \mathrm{e}^{-d_{i,j}/\tau}\Big|_a \tag{7}$$

- SEC-noRR: The relative reward term was removed, making actions independent of the reward values associated with the retrieved memories. The valuation function for this condition is

$$\hat{Q}_{a\in A} = \sum_{i,j\in M} G_{i,j} \mathrm{e}^{-d_{i,j}/\tau}\Big|_a \tag{8}$$

- SEC-noDist: The distance to the goal term was removed, eliminating the influence of spatial proximity on action selection. The modified valuation function for this ablation is

$$\hat{Q}_{a\in A} = \sum_{i,j\in M} G_{i,j} \frac{r_i}{r_{\max}}\Big|_a \tag{9}$$

Double Mechanism Ablation: In the second set of ablations, two mechanisms were inactivated simultaneously, allowing us to assess performance with only one mechanism active. This produced three variants: SEC-soloGi, SEC-soloRR and SEC-soloDist, each relying on a single mechanism from the action-selection equation.

In Experiment 4, we extended our investigation of the SEC models to evaluate the effects of memory constraints and forgetting mechanisms on performance. Specifically, we explored how these factors influence the models' behaviour in the Double T-Maze task, an environment that requires sophisticated memory management for optimal navigation and decision-making.

We introduced two types of forgetting mechanisms into the LTM of the models:

- FIFO forgetting (fifo): We applied a FIFO rule to the LTM, akin to the update mechanism of SEC's STM.
- Prioritized forgetting (rwd): In this new condition, less rewarding memories are more likely to be forgotten, prioritizing the retention of high-reward experiences.

The models tested under these conditions included the original SEC and its non-sequential counterpart, NSEC, as well as their respective forgetting variants: SEC-fifo, NSEC-fifo, SEC-rwd and NSEC-rwd.

Each reported experiment involved 20 simulations per model, with each simulation running for 5,000 episodes (5 million frames) to ensure statistical reliability. These controlled experiments are designed to dissect the operational parameters of the SEC model, providing insights into its functionality under different conditions and contributing to the broader understanding of episodic control in artificial intelligence.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data sets supporting the findings of this study are available via Zenodo at https://doi.org/10.5281/zenodo.11506323 (ref. 77).

Code availability

The implementation of the SEC model used in this study is available in the GitHub repository at https://github.com/IsmaelTito/SEC. The specific version used to generate the results is also archived on Zenodo at https://doi.org/10.5281/zenodo.14014111 (ref. 78).

References

1. Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018).
2. Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
3. Berner, C. et al. Dota 2 with large scale deep reinforcement learning. Preprint at http://arxiv.org/abs/1912.06680 (2019).
4. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, 1–58 (2017).
5. Marcus, G. Deep learning: a critical appraisal. Preprint at http://arxiv.org/abs/1801.00631 (2018).
6. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
7. Baker, B. et al. Emergent tool use from multi-agent autocurricula. *International Conference on Learning Representations* (ICLR, 2020).
8. Botvinick, M. et al. Reinforcement learning fast and slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
9. Hansen, S., Pritzel, A., Sprechmann, P., Barreto, A. & Blundell, C. Fast deep reinforcement learning using online adjustments from the past. In *Adv. Neural Information Processing Systems* (eds. Bengio, S. et al.) 10567–10577 (Curran Associates, 2018).
10. Zhu, G., Lin, Z., Yang, G. & Zhang, C. Episodic reinforcement learning with associative memory. In *International Conference on Learning Representations* (eds Zhu, G, Lin, Z., Yang G. & Zhang, C.) 370–384 (Curran Associates, 2019).
11. Lin, Z., Zhao, T., Yang, G. & Zhang, L. Episodic memory deep q-networks. In *Proc. IJCAI International Joint Conference on Artificial Intelligence* (ed. Lang, J.) 2433–2439 (IJCAI, 2018).
12. Lee, S. Y., Sungik, C. & Chung, S. Y. Sample-efficient deep reinforcement learning via episodic backward update. In *Advances in Neural Information Processing Systems* (eds Wallach, H. et al.) 2112–2121 (Curran Associates, 2019).
13. Blundell, C. et al. Model-free episodic control. Preprint at http://arxiv.org/abs/1606.04460 (2016).
14. Pritzel, A. et al. Neural episodic control. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Yeh, Y. W.) 2827–2836 (ACM, 2017).
15. Yalnizyan-Carson, A. & Richards, B. A. Forgetting enhances episodic control with structured memories. *Front. Comput. Neurosci.* **16**, 757244 (2022).
16. Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63**, 497–507 (2009).
17. Voegtlin, T. & Verschure, P. F. What can robots tell us about brains? A synthetic approach towards the study of learning and problem solving. *Rev. Neurosci.* **10**, 291–310 (1999).
18. Lisman, J. E. & Idiart, M. A. Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science* **267**, 1512–1515 (1995).
19. Jensen, O. & Lisman, J. E. Dual oscillations as the physiological basis for capacity limits. *Behav. Brain Sci.* **24**, 126 (2001).
20. Ramani, D. A short survey on memory based reinforcement learning. Preprint at http://arxiv.org/abs/1904.06736 (2019).
21. Buzsáki, G. & Tingley, D. Space and time: the hippocampus as a sequence generator. *Trends Cogn. Sci.* **22**, 853–869 (2018).
22. Lisman, J. & Redish, A. D. Prediction, sequences and the hippocampus. *Philos. Trans. R. Soc. B* **364**, 1193–1201 (2009).
23. Verschure, P. F., Pennartz, C. M. & Pezzulo, G. The why, what, where, when and how of goal-directed choice: neuronal and computational principles. *Philos. Trans. R. Soc. B* **369**, 20130483 (2014).
24. Merleau-Ponty, M. et al. *The Primacy of Perception: And Other Essays on Phenomenological Psychology, the Philosophy of Art, Hhistory, and Politics* (Northwestern Univ. Press, 1964).
25. Bornstein, A. M. & Norman, K. A. Reinstated episodic context guides sampling-based decisions for reward. *Nat. Neurosci.* **20**, 997–1003 (2017).
26. Wimmer, G. E. & Shohamy, D. Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science* **338**, 270–273 (2012).
27. Wu, C. M., Schulz, E. & Gershman, S. J. Inference and search on graph-structured spaces. *Comput. Brain Behav.* **4**, 125–147 (2021).
28. Johnson, A. & Redish, A. D. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12176–12189 (2007).
29. Ludvig, E. A., Madan, C. R. & Spetch, M. L. Priming memories of past wins induces risk seeking. *J. Exp. Psychol. Gen.* **144**, 24 (2015).
30. Wang, S., Feng, S. F. & Bornstein, A. M. Mixing memory and desire: How memory reactivation supports deliberative decision-making. *Wiley Interdiscip. Rev. Cogn. Sci.* **13**, e1581 (2022).
31. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
32. Santos-Pata, D. et al. Epistemic autonomy: self-supervised learning in the mammalian hippocampus. *Trends Cogn. Sci.* **25**, 582–595 (2021).
33. Santos-Pata, D. et al. Entorhinal mismatch: a model of self-supervised learning in the hippocampus. *iScience* **24**, 102364 (2021).
34. Amil, A. F., Freire, I. T. & Verschure, P. F. Discretization of continuous input spaces in the hippocampal autoencoder. Preprint at http://arxiv.org/abs/2405.14600 (2024).
35. Rennó-Costa, C., Lisman, J. E. & Verschure, P. F. The mechanism of rate remapping in the dentate gyrus. *Neuron* **68**, 1051–1058 (2010).
36. Estefan, D. P. et al. Coordinated representational reinstatement in the human hippocampus and lateral temporal cortex during episodic memory retrieval. *Nat. Commun.* **10**, 1–13 (2019).
37. de Almeida, L., Idiart, M. & Lisman, J. E. A second function of gamma frequency oscillations: an E%-max winner-take-all mechanism selects which cells fire. *J. Neurosci.* **29**, 7497–7503 (2009).
38. Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172 (1996).
39. Redish, A. D. Vicarious trial and error. *Nat. Rev. Neurosci.* **17**, 147–159 (2016).
40. Clayton, N. S. & Dickinson, A. Episodic-like memory during cache recovery by scrub jays. *Nature* **395**, 272–274 (1998).
41. Foster, D. J. & Knierim, J. J. Sequence learning and the role of the hippocampus in rodent navigation. *Curr. Opin. Neurobiol.* **22**, 294–300 (2012).
42. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
43. Eichenbaum, H. Memory: organization and control. *Annu. Rev. Psychol.* **68**, 19–45 (2017).

44. Estefan, D. P. et al. Volitional learning promotes theta phase coding in the human hippocampus. *Proc. Natl Acad. Sci. USA* **118**, e2021238118 (2021).
45. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018); https://doi.org/10.1109/tnn.2004.842673
46. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
47. Kubie, J. L. & Fenton, A. A. Heading-vector navigation based on head-direction cells and path integration. *Hippocampus* **19**, 456–479 (2009).
48. Mathews Z. et al. Insect-like mapless navigation based on head direction cells and contextual learning using chemo-visual sensors. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* 2243–2250 (IEEE, 2009).
49. Amil, A. F. & Verschure, P. F. Supercritical dynamics at the edge-of-chaos underlies optimal decision-making. *J. Phys. Complex.* **2**, 045017 (2021).
50. Verschure, P. F., Voegtlin, T. & Douglas, R. J. Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature* **425**, 620–624 (2003).
51. Vikbladh, O., Shohamy, D. & Daw, N. Episodic contributions to model-based reinforcement learning. In *Annual Conference on Cognitive Computational Neuroscience* (CCN, 2017).
52. Cazé, R., Khamassi, M., Aubin, L. & Girard, B. Hippocampal replays under the scrutiny of reinforcement learning models. *J. Neurophysiol.* **120**, 2877–2896 (2018).
53. Gonzalez, C., Lerch, J. F. & Lebiere, C. Instance-based learning in dynamic decision making. *Cogn. Sci.* **27**, 591–635 (2003).
54. Gonzalez, C. & Dutt, V. Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological Rev.* **118**, 523 (2011).
55. Lengyel, M. & Dayan, P. Hippocampal contributions to control: the third way. In *Proc. Advances in Neural Information Processing Systems* (eds. Platt, J. et al.) 889–896 (Curran, 2008).
56. Freire, I. T., Moulin-Frier, C., Sanchez-Fibla, M., Arsiwalla, X. D. & Verschure, P. F. Modeling the formation of social conventions from embodied real-time interactions. *PLoS ONE* **15**, e0234434 (2020).
57. Papoudakis, G., Christianos, F., Rahman, A. & Albrecht, S. V. Dealing with non-stationarity in multi-agent deep reinforcement learning. Preprint at http://arxiv.org/abs/1906.04737 (2019).
58. Freire, I. & Verschure, P. High-fidelity social learning via shared episodic memories can improve collaborative foraging. Paper presented at Intrinsically Motivated Open-Ended Learning Workshop@NeurIPS 2023 (2023).
59. Albrecht, S. V. & Stone, P. Autonomous agents modelling other agents: a comprehensive survey and open problems. *Artif. Intell.* **258**, 66–95 (2018).
60. Freire, I. T., Arsiwalla, X. D., Puigbò, J.-Y. & Verschure, P. F. Limits of multi-agent predictive models in the formation of social conventions. In *Proc. Artificial Intelligence Research and Development* (eds Falomir, Z. et al.) 297–301 (IOS, 2018).
61. Freire, I. T., Puigbò, J.-Y., Arsiwalla, X. D. & Verschure, P. F. Modeling the opponent's action using control-based reinforcement learning. In *Proc. Conference on Biomimetic and Biohybrid Systems* (eds Vouloutsi, V. et al.) 179–186 (Springer, 2018).
62. Freire, I. T., Arsiwalla, X. D., Puigbò, J.-Y. & Verschure, P. Modeling theory of mind in dyadic games using adaptive feedback control. *Information* **14**, 441 (2023).
63. Kahali, S. et al. Distributed adaptive control for virtual cyborgs: a case study for personalized rehabilitation. In *Proc. Conference on Biomimetic and Biohybrid Systems* (eds Meder, F. et al.) 16–32 (Springer, 2023).
64. Freire, I. T., Guerrero-Rosado, O., Amil, A. F. & Verschure, P. F. Socially adaptive cognitive architecture for human-robot collaboration in industrial settings. *Front. Robot. AI* **11**, 1248646 (2024).
65. Verschure, P. F. Distributed adaptive control: a theory of the mind, brain, body nexus. *BICA* **1**, 55–72 (2012).
66. Rosado, O. G., Amil, A. F., Freire, I. T. & Verschure, P. F. Drive competition underlies effective allostatic orchestration. *Front. Robot. AI* **9**, 1052998 (2022).
67. Daw, N. D. Are we of two minds? *Nat. Neurosci.* **21**, 1497–1499 (2018).
68. Freire, I. T., Urikh, D., Arsiwalla, X. D. & Verschure, P. F. Machine morality: from harm-avoidance to human-robot cooperation. In *Proc. Conference on Biomimetic and Biohybrid Systems* (eds Vouloutsi, V. et al.) 116–127 (Springer, 2020).
69. Verschure, P. F. Synthetic consciousness: the distributed adaptive control perspective. *Philos. Trans. R. Soc. B* **371**, 20150448 (2016).
70. Goode, T. D., Tanaka, K. Z., Sahay, A. & McHugh, T. J. An integrated index: engrams, place cells, and hippocampal memory. *Neuron* **107**, 805–820 (2020).
71. Amil, A F., Albesa-González, A. & Verschure, P. F. M. J. Theta oscillations optimize a speed-precision trade-off in phase coding neurons. *PLOS Comp. Biol.* **20.12**, e1012628 (2024).
72. Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398**, 704–708 (1999).
73. Cromwell, H. C., Hassani, O. K. & Schultz, W. Relative reward processing in primate striatum. *Exp. Brain Res.* **162**, 520–525 (2005).
74. Soldati, F., Burman, O. H., John, E. A., Pike, T. W. & Wilkinson, A. Long-term memory of relative reward values. *Biol. Lett.* **13**, 20160853 (2017).
75. Beyret, B. et al. The Animal-AI environment: training and testing animal-like artificial cognition. Preprint at http://arxiv.org/abs/1909.07483 (2019).
76. Crosby, M., Beyret, B. & Halina, M. The Animal-AI olympics. *Nat. Mach. Intell.* **1**, 257 (2019).
77. Freire, I. T. Dataset for 'Sequential memory improves sample and memory efficiency in episodic control'. *Zenodo* https://doi.org/10.5281/zenodo.11506323 (2024).
78. Freire, I. T. IsmaelTito/SEC: SEC v.1.0 release (v.1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.14014111 (2024).

Acknowledgements

This study was funded by the Counterfactual Assessment and Valuation for Awareness Architecture (CAVAA) project (European Innovation Council's Horizon programme, grant no. 101071178) awarded to P.F.M.J.V.

Author contributions

Conceptualization: all authors. Methodology: I.T.F. and A.F.A. Software: I.T.F. and A.F.A. Data curation: I.T.F. Formal analysis: I.T.F. Resources: P.F.M.J.V. Writing—original draft: I.T.F. Writing—review and editing: all authors. Visualization: I.T.F. Supervision: P.F.M.J.V. Project administration: P.F.M.J.V. Funding acquisition: P.F.M.J.V. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

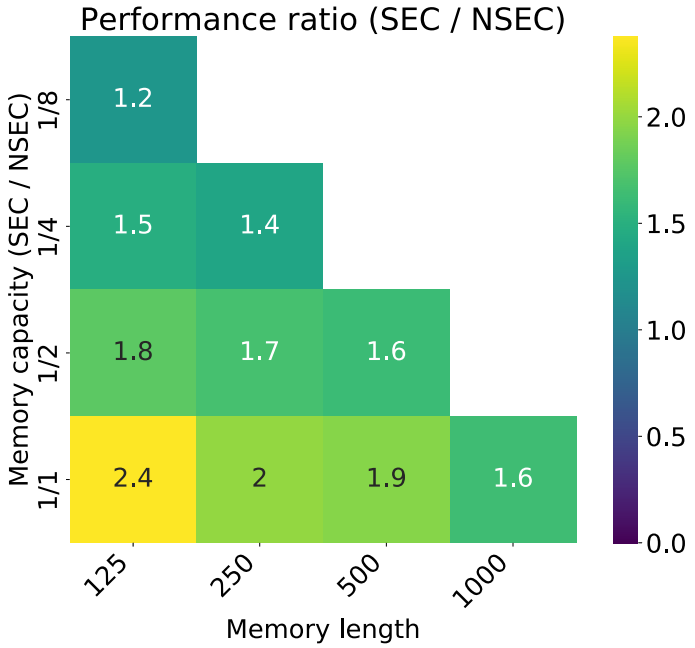
Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00950-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00950-3.

Correspondence and requests for materials should be addressed to Ismael T. Freire, Adrián F. Amil or Paul F. M. J. Verschure.

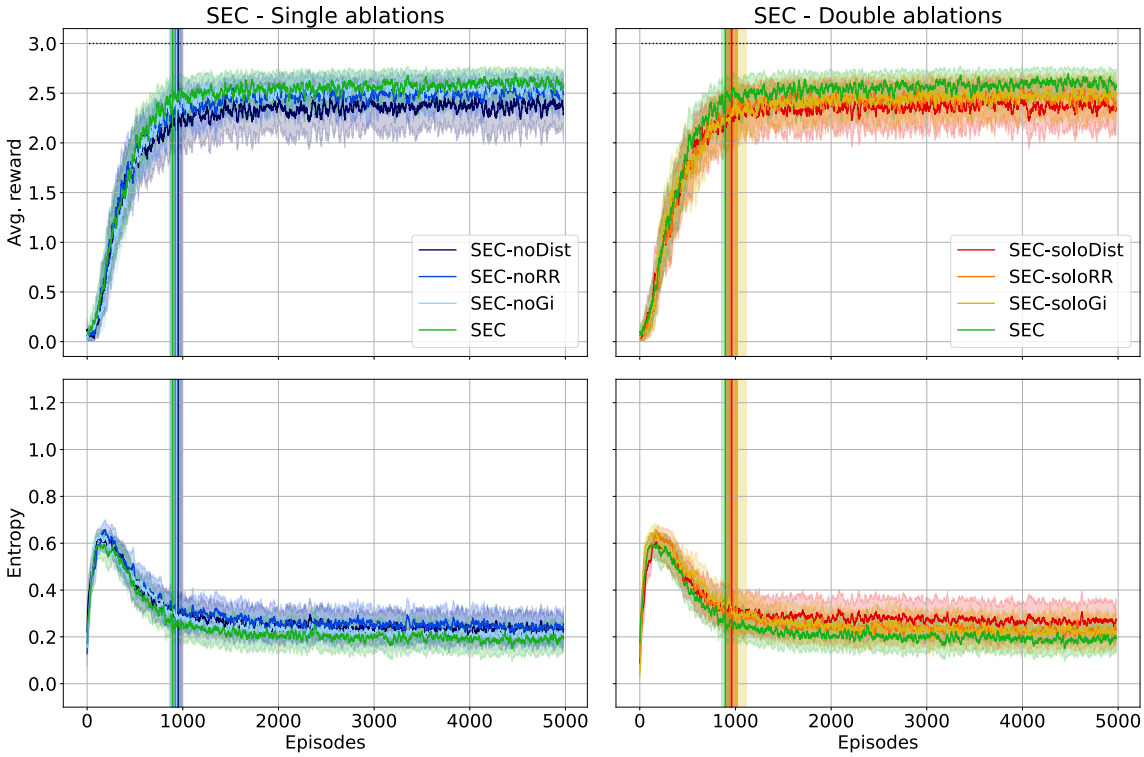
Peer review information *Nature Machine Intelligence* thanks Mehdi Khamassi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-024-00950-3



Extended Data Fig. 1 | Performance increase of SEC over NSEC across different memory-limit conditions in the Double T-maze benchmark. Units reported are the total mean perfor- mance of SEC over NSEC. Each column shows SEC's performance with a limited memory capacity of 125, 250, 500, and 1000 sequences respectively. Each row shows the memory ratio between SEC and NSEC, ranging from 1/1 (equal memory limit) to 1/8 (SEC memory limit is 8 times smaller than NSEC).

https://doi.org/10.1038/s42256-024-00950-3



Extended Data Fig. 2 | Ablation studies of the Sequential Episodic Control (SEC) algorithm in the Double T-Maze task. The left panel presents the single mechanism ablations with average reward (top) and entropy (bottom). The right panel shows the double mechanism ablations under the same metrics. In the single ablations, 'SEC- noDist' lacks the distance to the goal component, 'SEC-noRR' lacks the relative reward component, and 'SEC-noGi' lacks the eligibility score component. In the double ablations, 'SEC- soloDist', 'SEC-soloRR', and 'SEC-soloGi' operate with only the distance to the goal, relative reward, and eligibility score components, respectively. The full SEC model is included as a benchmark in both panels. These graphs demonstrate the comparative impact of individual and combined components of the SEC algorithm on its performance and decision-making uncertainty. Vertical bars represent the average episode around which the memory was filled. Average values were computed using a sliding window of 20 episodes. Error bars represent SE.

nature portfolio

Corresponding author(s): Ismael T. Freire

Last updated by author(s): Oct 29, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The custom code use to generate the findings of this study are available in a code repository with the identifier(s): https://github.com/IsmaelTito/SEC, https://zenodo.org/records/14014111
Data analysis	The custom code use to generate the findings of this study are available in a code repository with the identifier(s): https://github.com/IsmaelTito/SEC, https://zenodo.org/records/14014111

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are available in Zenodo with the identifier(s): https://zenodo.org/records/11506323

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In this study, a sample size of 20 runs per computational model was used to capture variability in model performance without relying on a predetermined statistical calculation for sample size. This choice aligns with common practices in machine learning research, where multiple simulations per model are conducted to demonstrate trends in model behavior and stability across trials. Given the high dimensionality and variability in performance metrics typical of reinforcement learning tasks, this sample size was deemed sufficient to capture meaningful trends, allowing for qualitative insights into model performance without relying on strict significance testing.
Data exclusions	No data has been excluded.
Replication	Code for the computational models used in this study, along with the datasets generated and the data analysis scripts are available.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging