FAST LEARNING VIA EPISODIC MEMORY: A PERSPECTIVE FROM ANIMAL DECISION-MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

A typical experiment to study cognitive function is to train animals to perform tasks, while the researcher records the electrical activity of the animals neurons. The main obstacle faced, when using this type of electrophysiological experiment to uncover the circuit mechanisms underlying complex behaviors, is our incomplete access to relevant circuits in the brain. One promising approach is to model neural circuits using an artificial neural network (ANN), which can provide complete access to the "neural circuits" responsible for a behavior. More recently, reinforcement learning models have been adopted to understand the functions of cortico-basal ganglia circuits as reward-based learning has been found in mammalian brain. In this paper, we propose a Biologically-plausible Actor-Critic with Episodic Memory (B-ACEM) framework to model a prefrontal cortex-basal ganglia-hippocampus (PFC-BG) circuit, which is verified to capture the behavioral findings from a well-known perceptual decision-making task, i.e., random dots motion discrimination. This B-ACEM framework links neural computation to behaviors, on which we can explore how episodic memory should be considered to govern future decision. Experiments are conducted using different settings of the episodic memory and results show that all patterns of episodic memories can speed up learning. In particular, salient events can be prioritized to propagate reward information and guide decisions. Our B-ACEM framework and the built-on experiments give inspirations to both designs for more standard decision-making models in biological system and a more biologically-plausible ANN.

1 INTRODUCTION

A hallmark of higher brain function is the ability to form decisions from sensory inputs to guide appropriate behavioral responses. Understanding the relationship between an animal's behavioral responses, and how this is encoded in the brain, is a major goal in neuroscience. Neurophysiologists have began to undertake studies on behavior training of nonhuman primates in a variety of decision tasks, such as perceptual discrimination (Shadlen & Newsome, 2001; Romo et al., 2004; Heekeren et al., 2008). These electrophysiological experiments have uncovered neural signals at the single-neuron level that are correlated with specific aspects of decision computation. However, in the mammalian brain, a decision is not made by single neuron, but by the collective dynamics of a neural circuit. Unfortunately, the animal experiment does not allow us to access to a complete record of all relevant neural circuits in the brain. Therefore, neural circuit modeling using ANN can provide a valuable tool to uncover circuit mechanisms underlying complex behaviors.

Reinforcement learning (RL) has greatly influenced the neuroscience study of recognitive function, which integrates computational modeling and empirical research in neuroscience. A wide array of evidence shows that the cortico-basal ganglia circuit appears to implement RL algorithm (O'Doherty et al., 2004; Sohal et al., 2009), which is driven by a reward prediction error (RPE). This RPE signal, conveyed by dopamine, is thought to gate Hebbian synaptic plasticity in the striatum (PR Montague & Sejnowski, 1996). Over the last decade, this approach has produced explicit models to understand the functions of dopamine and cortico-basal ganglia circuits (Cohen & Frank, 2009; Maia, 2009). Recent functional magnetic resonance imaging (fMRI) studies in humans revealed that the activation in the hippocampus, a central for storing episodic memory (Paller & Wagner, 2002)), is modulated by reward, which demonstrates a link between episodic memory and RL (Wittmann et al., 2005; Krebs et al., 2009). However, the existing RL models does not take into account the effect of episodic memory, which is necessary for those who want to explore cognitive functions using modeling circuits.

In this paper, we propose a Biologically-plausible Actor-Critic with Episodic Memory (B-ACEM) (Figure 1, right), based on biological anatomy and RL algorithms for artificial systems, to model cortico-basal ganglia-hippocampus circuits (Figure 1, left). B-ACEM is an actor critic-based framework modeled by recurrent neural network (RNN), which is a natural class of models to study mechanisms in neuroscience systems because they are both dynamical and computational (Mante et al., 2013). This framework was trained for a classical perceptual decision-making task, i.e. random dots motion discrimination, in which a monkey is asked to arbitrarily choose the direction (left or right) of a flow of moving dots. We show that an agent that accesses episodic memories reproduces qualitative results including (i) psychometric function, a tool for analyzing the relationship between accuracy and stimulus strength (Figure 3, top), and (ii) chronometric function, a tool for analyzing the relationship between response time and stimulus strength (Figure 3, bottom). **We need to emphasize that, unlike most machine learning applications, our aim in training B-ACEM framework is not simply to maximize its performance, but to train networks so that their performance matches that of behaving animals while the architecture is as close to biological systems as possible.** On the other hand, anatomical and electrophysiological studies in animals, including humans, suggest that the episodic memory in the hippocampus is critical for adaptive behavior. Yet existing theory fails to describe how the brain selects experiences, from many possible options, to govern the decisions that are made. To address this gap, we investigated which episodic memories should be accessed to enable the most rewarding future decisions using the validated, biologically plausible B-ACEM model. The results show that all patterns of episodic memories can speed up learning, where salient events in memory replay are prioritized to propagate reward information and guide decisions.

2 Biologically-plausible Actor-Critic with Episodic Memory

2.1 Recurrent neural network

In this section, we present an RNN unit in biological contexts. Some neuroscientists have introduced RNNs into the field of neuroscience systems. Wilson & Cowan (1972) initially exploited a recurrent neural network to describe the average firing rate of neural populations in a biological context. A more modern and general definition of an RNN unit is given by (Sussillo, 2014):

$$\tau \frac{d\boldsymbol{x}}{d\boldsymbol{t}} = -\boldsymbol{x} + \boldsymbol{W}^{rec}\boldsymbol{r} + \boldsymbol{W}^{in}\boldsymbol{u} + \boldsymbol{b}, \tag{1}$$

where the ith component, x_i, of the vector $\boldsymbol{x}$, can be viewed as the the sum of the filtered synaptic currents at the soma of a biological neuron. The variable r_i is the instantaneous, positive firing rate, which is obtained by a threshold-linear activation function $[x]^+ = max(0, x)$. The vector $\boldsymbol{u}$ presents the external inputs to the network. Each unit in the network receives a bias, b_i. The time constant τ sets the timescale of the network. On the other hand, a parallel neural system allows biological agents to solve learning problems on different timescale. Theoretical and modelling studies indicate that learning with a multiple timescale can improve performance (Koutnik et al., 2014) and speed up learning (Neil et al., 2016). This multiplicity of timescales is also an important feature of gated recurrent units (GRUs) indicated by Roitman & Shadlen (2002), in which each unit learns to adaptively capture dependencies over different time scales. Therefore, in this work we use modified GRUs, which is modified through Equation (1) and combined with the standard GRUs. A continuous-time form of the modified GRUs is described in Appendix, which can be discretized to Euler form in time steps of size Δt as follows:

$$\boldsymbol{\alpha}_t = \sigma_g(\boldsymbol{W}_\alpha^{rec}\boldsymbol{r}_{t-1} + \boldsymbol{W}_\alpha^{in}\boldsymbol{u}_t + \boldsymbol{b}_\alpha), \tag{2}$$

$$\boldsymbol{\beta}_t = \sigma_g(\boldsymbol{W}_\beta^{rec}\boldsymbol{r}_{t-1} + \boldsymbol{W}_\beta^{in}\mathbf{u}_t + \boldsymbol{b}_\beta), \tag{3}$$

$$\boldsymbol{x}_t = (1 - \boldsymbol{\lambda}_t) \circ \boldsymbol{x}_{t-1} + \boldsymbol{\lambda}_t \circ f(\boldsymbol{W}^{rec}(\boldsymbol{\beta} \circ \boldsymbol{r}_{t-1}) + \boldsymbol{W}^{in}\boldsymbol{u}_t + \boldsymbol{b}_t + \sqrt{2\frac{\tau\sigma_{rec}^2}{\Delta t}}\boldsymbol{N}(0,1)), \tag{4}$$

$$\boldsymbol{r}_t = [\boldsymbol{x}_t]^+, \tag{5}$$

$$\boldsymbol{z}_t = \boldsymbol{W}^{out}\boldsymbol{r}_t, \tag{6}$$

where $\circ$ denotes the Hadamard product, σ_g is sigmoid function. The function $f(x) = x$ is a linear function. The vector $\boldsymbol{\lambda}_t = \frac{\Delta t}{\tau}\boldsymbol{\alpha}_t$, and the size of update gate $\boldsymbol{\alpha}_t$ is scaled by τ and Δt. $\mathbf{N}(0, 1)$ are

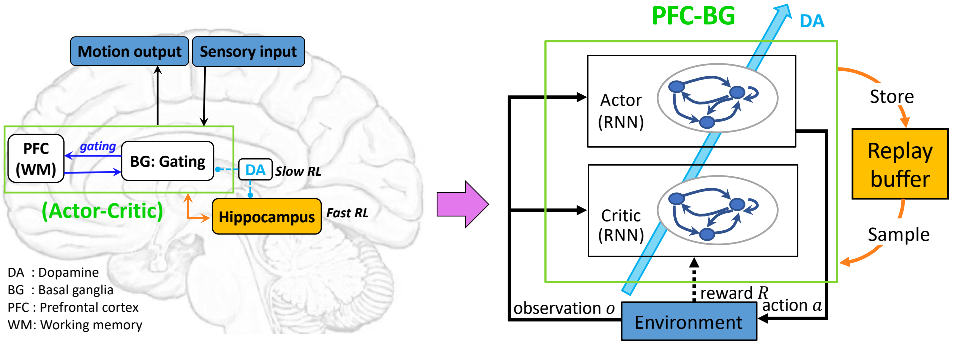

Figure 1: Biologically-plausible Actor-Critic with episodic memory framework. **(Left) Components of a B-ACEM framework**, based on biological connections and functions of the **PFC** (robust active maintenance of task-relevant information), **BG** (dynamic gating of PFC active maintenance), **DA** (training the BG gating), **Hippocampus** (storing episodic memory). Sensory inputs are processed by PFC-BG circuit and corresponding motor signals are sent out by Thalamus (not shown here). Working memory representations in PFC are updated via dynamic gating by the BG. These gating functions are learned by BG on the basis of modulatory input from dopaminergic neuron (blue dotted line), i.e., dopamine drives incremental learning (slow RL) in BG regions. Moreover, dopamine modulates episodic memories in the hippocampus, supporting adaptive behaviors (fast RL). The synaptic weights in the PFC-BG network are adjusted by an RL procedure, in which DA conveys a reward prediction error signal. **(Right) The computational framework of B-ACEM**. The PFC-BG circuits in the brain were mapped to the actor-critic framework (green box). At each time step, the actor receives an observation o and selects an action a based on the past experience (working memory stored in RNN). The reward R is given followed by the chosen action and the environment moves the next state. Replay buffers are used to store episodic memories, similar to the function of the hippocampus. The weights of the neural networks are updated through gradient descent, whose error is driven by DA.

standard, normal distributed random numbers, which are scaled by σ_{rec}. We note that in the case where $\Delta t = \tau$, i.e., $\lambda_t = \alpha_t$, the only difference from standard GRUs is that there is threshold-linear activation function $[\boldsymbol{x}_t]^+$ in the system.

2.2 Biologically-plausible framework

The framework we proposed is based on four assumptions listed below:

1. *Actor-critic architecture for RL in biological system.* This assumption states that a cortex-basal ganglia (PFC-BG) circuit can be modeled as an actor-critic architecture (Dayan & Balleine, 2002; Suri et al., 2001; Joel et al., 2002; O'Doherty et al., 2004; Haber, 2014). In this process the midbrain dopamine neurons play a central role, which code reinforcement prediction error. Frank (2005) has demonstrated that basal ganglia (BG) can perform dynamic gating through the modulatory mechanism of 'No-Go' pathway, facilitating maintenance of task-related information in the prefrontal cortex and suppressing other distracting information (Figure 1, left). The actor-critic view of action selection in the brain suggests that the dorsal striatum in PFC-BG is responsible for learning stimulus-response association, which can be thought of as the 'actor' in the actor-critic architecture. The ventral striatum in BG, together with cortex, is mainly used to learns state values, which is akin to the 'critic' in this framework (Maia, 2009; 2010).

2. *Recurrent neural networks reproduce neural population dynamics.* This assumption states that we can conceptualize a PFC-BG system using recurrent neural networks (RNNs), for both actor and critic. There are many essential similarities between RNNs and biological neural circuits: First, RNNs units are nonlinear and numerous. Second, the units have feedback connections, which allows them to generates temporal dynamic behavior within the circuit. Third, individual units are simple, so they need to work together in a parallel and distributed manner to implement complex computations. RNNs are very powerful. As long as the number of hidden units is sufficient, an optimized RNNs can approximate any dynamical system. Both dynamical and computational features of RNNs make it an ideal model for studying the mechanisms of system neuroscience

(Rajan et al., 2016; Sussillo, 2014; Mante et al., 2013). Since basal ganglia can perform dynamic gating via reinforcement learning mechanisms (Figure 1, left), here we consider more sophisticated units, i.e., gated recurrent units (GRUs) to implement this gating mechanism. At the computational level, our model is most closely related to GRUs, whose dynamic gating signals are trained by error backpropagation (Cho et al., 2014).

3. *Episodic memory contributes to decision making.* This assumption states that episodic memory, depending crucially on the hippocampus and surrounding medial temporal lobe (MTL) cortices, can be used as a complementary system for reinforcement learning to influence decisions. First, in addition to its role in remembering the past, the MTL also supports the ability to imagine specific episodes in the future (Hassabis et al., 2007), with direct implications for decision making (Peters & Büchel, 2010). Second, episodic memories are constructed in a way that allows relevant elements of a past event to guide future decision (Shohamy & Wagner, 2008).

4. *There are two different forms of learning in biological systems: slow learning and fast learning.* Many researchers believe that cortex-basal ganglia circuits appear to implement reinforcement learning Frank et al. (2004). Hence, we assumed that the synaptic weights of dopamine targets (striatum in BG) in the circuit, including the PFC network, can be modulated by a model-free RL procedure. As mentioned earlier, this method of incremental parameter adjustment makes it a slow form of learning. On the other hand, the hippocampus has been shown to support encoding of single events, and episodic memories stored in hippocampus impact reward-based learning to bias related behavior. Thus, the hippocampus can serve as a supplementary system to reinforcement learning. From this we assumed that episodic memories in replay buffer (a function similar to the hippocampus) can be used to estimate the value of actions and states to guide reward-based decision-making (Wimmer et al., 2014), which is a fast form of learning.

These assumptions are all based on existing research. For demonstration, we abstract the neural basis of RL in biological systems (Figure 1 left) into a simple computational model (Figure 1 right), an actor-critic equipped with episodic memory architecture, in which actor network leverages perceptual data provided by the environment to make a choice, while the critic network emits the value of the selected option. We exploit recent advances in deep RL, specifically the application of the policy gradient algorithm to RNN (Bakker, 2002), to update the weights of the network. Here the networks can be trained for a well-known perceptual decision-making task (Figure 2 left), which is characterized by an input-output mapping. On each trial, the value of GRU cells is updated based on the inputs and the last value of the GRU cells in the network, which enables the hidden layer to track relevant history information (refer to the working memory in biology). Through this training, the actor network learns to extract history experiences into the hidden state in the form of a working memory (WM). This working memory is thought to be facilitated by the prefrontal cortex, which instructs the actor system to select rewarding actions. The critic system learns a value function to train the actor network, which in turn furnishes a dynamic gating mechanism to control the updating of working memory.

3 The role of memory in RL

3.1 Memories contribute to RL: a picture in biological system

Memory is essential to make decisions, enabling organisms to draw on past events to predicting possible future outcomes. Working memory, a temporary storage in the brain, has been shown to guide choices by maintaining and manipulating task-relevant information (Krebs et al., 2009). Oberauer (2009) suggested that the working memory model can be divided into two largely independent subsections: One subsection represents procedural memory (implicit memory) and the other subsection represents declarative memory (explicit memory). RL theory has recognized the relationship between RL, and procedural memory and declarative memory. Procedural memory, created by repeating complex activities over and over again, is responsible for generating model-free policies or action values in reinforcement learning. Declarative memory refers to general world knowledge of the world that is independent of personal experience and can be clearly articulated. Furthermore, declarative memory contains information about reinforcement learning models or environmental mappings. Yet, early work demonstrates that working memory capacity is limited, resulting in decisions that are often made with finite information. Therefore, with a transient characteristic caused by the limited capacity and fast decay rate of working memory, it is not an ideal memory system

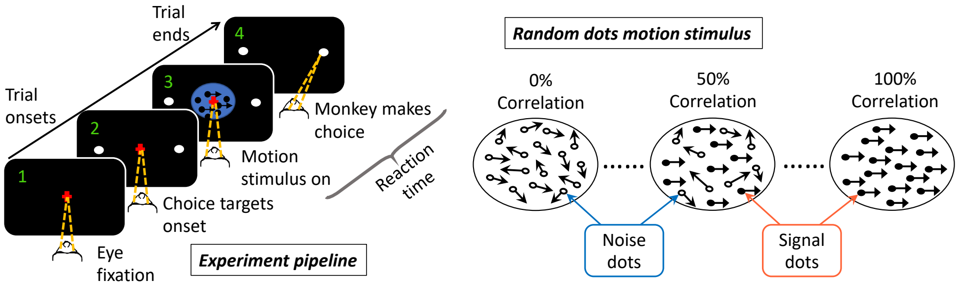

Figure 2: **(Left)** Pipeline for Random Dots Motion Discrimination (RDMD) task. Monkeys are trained to discriminate the direction of motion in a random dot stimulus that contained horizontal coherent motion. After fixation (screen 1), the two choice targets appeared in the periphery (screen 2). After a variable delay period, dynamic random dots appeared in a 5° diameter aperture (screen 3). The monkey was allowed to make a saccadic eye movement to a choice target at any time after onset of random-dot motion to indicate the direction of perceived motion (screen 4). Reaction time (RT) is defined as the elapsed time from motion onset to the initiation of the saccade, which was controlled by the monkeys and could be measured. **(Right)** Examples of random-dot motion stimulus of variable motion coherence. Stimulus strength is varied by changing the proportion of dots moving coherently in a single direction, which determines the difficulty of the task. The lower (higher) the coherence levels, the more difficult (easier) the task is. Coherently moving dots are the "signal", and randomly moving dots are the "noise".

to support decision-making independently. For this reason the brain needs other complementary systems, such episodic memory, to support reinforcement learning along with working memory.

Episodic memory, the type of memory that we mainly consider here, has also been implemented in decision-making. Psychologically, episodic memory refers to the capacity for consciously recollecting an autobiographical memory of events that occurred in a particular times and places. Computationally, we often mainly emphasize the notion of one-time episodes (like one-trial learning in a task). Earlier research suggests that episodic memory could be used to store the specific rewarding sequence of state-action pairs then later, attempt to mimic the sequence in a process termed episodic control (Lengyel & Dayan, 2008). Based on this idea, we propose a different computational principle, in which episodic memory is used to optimize the policy rather than directly extract policy.

3.2 Memory in B-ACEM framework

The link between reward-guided choice and episodic memory brings us to two areas of research, which have a well-developed bodies of computational theory and widespread impact on cognition. This opens the door for utilizing many other features of episodic memory to make decisions. However, it seems reasonable to suspect that decision makers can have access to episodic information to estimate value. It is necessary to further study how the episodic characteristics of memory samples affect their sampling process. For this reason, we leverage the B-ACEM framework to explore how episodic memory should be sampled in order to better facilitate agents to learn a task. We store the agent's experiences (a sequence of observations, actions and rewards) at each time-step, pooled over many episodes (an episodic memory is a trial) into a replay buffer as shown in Figure 1. Unlike the classical method of replay experience, described in Mnih et al. (2015), a complete episodic memory (such as a complete trial) that meets certain conditions is stored here. The agent will learn behavior strategy with episodic memory to maximize the future accumulation of rewards.

How does memory, including working memory and episodic memory, in this B-ACEM architecture works in practice? On the one hand, working memory as represented by hidden states keeps track of the reward probabilities. To evaluate these quantities, working memory must maintain information about its past behavior and states. On the other hand, when the last step of the trial is reached, the agent will save the activation of the current working memory as long-term memory in the form of episodic memory. These episodic memories will be retrieved with a certain probability used to evaluate these quantities. As mentioned in section 2.2, there are two different forms of learning in biological systems: slow learning (implemented by working memory) and fast learning (implemented by episodic memory). Here, working memory encodes the outcome state and rewards as a short-term memory, which focus primarily on current and relevant information, while ignoring the remaining information (a form of “executive” attention). Once the information is stored as long-term memory, the agent can utilize this important episodic information to make model-free evaluations of trial information retrieved from long-term memory.

4 Experiments

4.1 Random Dots Motion Discrimination Task

We begin with a general description of the RDMD task (reaction-time version) as shown in Figure 2, in which a monkey chooses between two visual targets. First, the monkey was required to fixate a central point until the random dot motion appears on the screen. Then, the monkey indicated its decision in the direction of dots, by making a saccadic eye movement to the target of choice. In the standard reinforcement learning model, an RL agent learns by interacting with its environment and receiving rewards for performing actions. Accordingly, in the RDMD task, the actual direction of moving dots can be considered to be a state of the environment, which is partially observable, because the monkey does not know the precise direction of coherent motion. Therefore, the monkey needs to integrate the noisy sensory stimuli to figure out the direction. A positive reward, such as a fruit juice, is given for choosing correct target after the fixation cue turns off, such as a juice reward. Either breaking fixation too early or not making a choice during the stimulus period leads to negative reward in the form of timeouts. During the stimulation, the incorrect response was neither rewarded by the juice nor timeouts, with a corresponding reward of zero. Given the reward schedule, the policy could be modeled and optimized by the method of policy gradient.

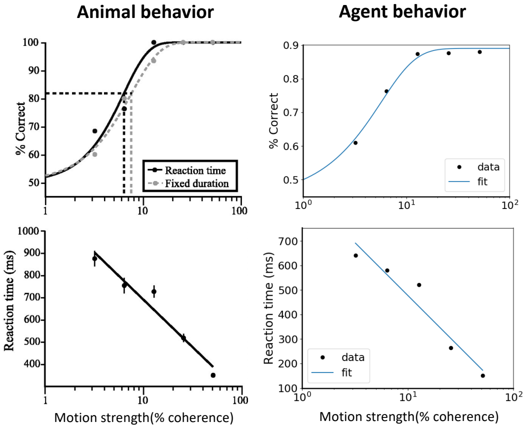

Figure 3: Behavior comparison of the animal and the agent during training in the RDMD task reflected in psychometric functions (top) and chronimetric fucntions (bottom). The **left** is animal behavioral data from one experience (reproduced from Roitman & Shadlen (2002)). The **right** is our agent behavioral data. **(Top)** Psychometric functions from reaction time version of the RDMD. The probability of a correct direction judgment is plotted as a function of motion strength and fitted by sigmiod functions. **(Bottom)** Effect of motion strength on reaction time (average reaction time of correct trials). The relationship between the log scaled motion strength and the reaction time fits a linear function.

4.2 Biological plausibility of ACER framework

We first investigated whether the B-ACEM framework we built, captures behavioral characteristics of animals in cognitive experiments. To empirically test this, we trained B-ACEM to perform a reaction-time version of RDMD task and then compared its behavior to the animal observed by Roitman & Shadlen (2002). The results are consistent with behavioral findings from the RDMD experiment, which are mainly reflected in psychometric function and chronometric function as shown in Figure 3. Performance accuracy on the RDMD task depends on the strength of sensory input, and the psychometric functions is a good tool for analyzing such relationship. The percentage of a correct direction judgment is plotted as a function of motion strength (is measured by the proportion of coherently moving dot). The Figure 3 shows that accuracy is high during strong motion, while it showed less accuracy as with more chance and a weaker motion, suggesting that the agent in our B-ACEM framework captures this important behavioral feature. Moreover, the theory of chronometric functions has a constraint on the relationship between response time and accuracy. When the task is difficult (weaker stimuli strength), it requires the agent to take more time to make a decision (Figure 3). This means that the additional viewing time for difficult trials was devoted to integrating sensory information. Thus the appropriate speed-accuracy trade-off is learned by this B-ACEM framework. We need to emphasize that, unlike the usual machine learning goals, *our objective is* *not to achieve "perfect" performance, but rather to train agents to match the smooth psychometric characteristics and chronometric characteristics observed in behaving monkeys.*

4.3 Episodic memory speeds up learning

Episodic memory making representations of past experiences plays a critical role in enabling us to act appropriately in the world. One of the unanswered questions in cognitive neuroscience is what types of samples of episodic memory should be selected. That is, how is the priority function defined to select samples from episodic memory? In the terms of non-parametric statistical or kernel-based model, this function corresponds to the kernel, which can be modifiable by past experience. Here, we combine reaction time task with B-ACEM model to investigate the impact of memories of individual trials on state-value estimation, so as to clarify computational nature of interactions between episodic memory and reinforcement learning. We first need to verify whether B-ACEM's episodic memory is performing effectively. For this purpose, we assessed the B-ACEM's performance in the case that all types of individual memories are stored, which act as as a baseline to measure the impact of a single, selected experience on decision-making. The blue lines in Figure 4 (left) show the learning curve of agents with memory for RDMD task (averaged return in 2000 trial samples). For comparison, an agent without episodic memory was trained and tested in exactly the same way, and the training curve is also shown in blue line. We can see that, the agent with episodic memory perform significantly faster in RDMD tasks than the one without episodic memory, although both models eventually reached the same baseline performance. These results are consistent with some recent research showing that animal decisions can indeed be guided by samples of individual past choices (Murty et al., 2016; Jang et al., 2019).

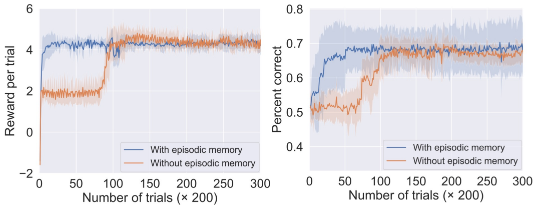

Figure 4: Learning curves of the agent with episodic memory (blue line) and without episodic memory (orange line) on the RDMD task. **(Left)** Average reward per trial. **(Right)** Percent correct, for trials on which the network made a decision.

The correct percent of trial is shown in Figure 4 (right), which is calculated by N_{right}/N_{choice}. The term N_{choice} represents number of trials in which the monkey made a choice (right or wrong) in 2000 trials and N_{right} is the number of correct choice. We can see that, at the beginning of the trial, the correct percentage of agents who cannot extract episodic memory from the replay buffer is maintained at around 50% (orange line), that is, the motion direction of the dot is randomly selected. It is only after substantial training (about $18,000$ trials) that the agent can achieve the baseline accuracy rate. Whilst the common GRU agent solves the task relatively successfully (as long as the training sample is sufficient), the agent being equipped with episodic memory shows better execution efficiency.

Like the monkeys in the RDMD trial, the network we built also converges on an optimal strategy: the agent not only successfully maintains gaze at the beginning of each trial, but also always chooses the direction that was rewarded starting on the second trial of each episode, regardless of whether the direction was left or right in the first trial. This reflects the agent's implicit understanding of task structure: After observing one trial results, the agent binds an unfamiliar environment to a specific task role.

4.4 Episodic memory for salient event

In the previous section, we have verified that episodic memory can indeed speed up the agent learning tasks. In this section, we will examine the questions raised at the beginning, regarding the types of events that should be stored as episodic memories to guide choice. The relationship between events is often clear only when they are reviewed. For example, when something positive happens, we want to know how to repeat this event. However, when an event occurs before the reward is given, how do we know what causes it? This is the 'temporal credit assignment problem' mentioned early, which can be solved by saving all potential determinants, such as rewards, of behaviorally relevant events into working memory. We proposed the question: How does episodic memory balance the need to represent these potential determinants of reward outcomes to deal with credit assignment? One solution may be to enhance episodic memory for notable events we refer to as ‘salient memory’, which are potential determinants of reward.

Both expectancy violations and expectancy conformance can be considered salient events to be stored in memory buffer. Because such long-term memories are potentially predictive of reward outcomes, it will provide a computationally feasible way to obtain future rewards. In the RDMD task, salient events included trials in which the right choice was made (rewarded with a glass of juice; expectancy conformance) or the fixation was broken (punished; expectancy violations). In a gaze-breaking trial, the agents policy can not be optimized due to insufficient interaction with the environment, thus we only choose expectancy conformance as a salient event. In the third type of trial, the monkeys make a response before a trial was over, but the choice was wrong. The incorrect response was neither rewarded by the juice nor punished. Such trial can be considered as a common event, because it’s not a particular event for monkeys.

To investigate if salient events in memory buffer can bias reward-guided choice more effectively than common events, we plot learning curve of B-ACEM with different types of episodic memory. Comparing the orange (salient events) and green (common events) curves, in Figure 5, it is apparent that B-ACEM with salient event outperforms B-ACEM with trivial event. As in figure 5 (left), if the agent extracts salient memories from the memory pool at a certain probability to train the network, it can make the network converge to the baseline policy faster (orange line). After fluctuating around the baseline (about 4.2) for a period of time, the agent’s episodic reward continues to increase, and finally is significantly higher than the baseline level. However, when agent extracts common event from memory pool (green line), the rewards obtained by agents will always be lower than the baseline level. In this case, the correct percentage of agent also always maintained at around 50% (random selection), suggesting that episodic memory about common events did not help the monkeys determine the direction of movement of random dots. In contrast, salient episodic memory is able to improve the final accuracy (orange line in Figure 5(right)) using exactly the same architecture and learning parameters as our baseline.

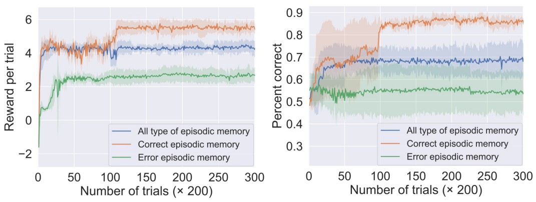

Figure 5: Learning curves of the agent on the RDMD task for different types of episodic memory, such as salient memory (orange line), common episodic memory (green line) and all types of episodic memory (blue). **(Left)** Average reward per trial. **(Right)** Percent correct.

Salient episodic memory is essential to future goal-directed behavior, which allows past relevant experience to improve choice and actions. Much research in humans has investigated how the hippocampus builds adaptive memory for past events. Our result suggesting that memory encoding was stronger for trials that involved salient events. The monkey would make the salient episodic memory in hippocampus more likely to be sampled during the ensuing choice.

5 Future Work

In this work, we developed a B-ACEM framework for the hippocampus, the prefrontal cortex, and basal ganglia based on anatomy and psychology. This framework supports reward-based learning in the brain, in which one circuit directly computes the policy to be followed, with one component computing the expected future reward to guide learning and one component acting as a complementary system for reinforcement learning. Although we have performed a few analyses using our B-ACEM, it is still an open problem of the detailed mechanisms by which episodic sampling can speed up learning and perform better than incremental learning alone. Many interesting and challenging questions remain. For example, training of spike neurons and reinforcement learning with spike-timing-dependent plasticity (STDP), a biological process that adjusts the strength of connections between neurons in the brain, is promising and future work will include these advances. Other physiologically relevant phenomena such as bursting, adaptation, and oscillations are currently not captured by our B-ACEM framework, which will be incorporated in the future.

REFERENCES

Bram Bakker. Reinforcement learning with long short-term memory. In *NIPS*, 2002.

Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014.

Michael X Cohen and Michael J Frank. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural brain research*, 2009.

Peter Dayan and Bernard W Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36 (2):285–298, 2002.

Michael J Frank. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of cognitive neuroscience*, 17(1):51–72, 2005.

Michael J Frank, Lauren C Seeberger, and Randall C O'reilly. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, 306(5703):1940–1943, 2004.

SUZANNE N Haber. The place of dopamine in the cortico-basal ganglia circuit. *Neuroscience*, 282: 248–257, 2014.

Demis Hassabis, Dharshan Kumaran, Seralynne D Vann, and Eleanor A Maguire. Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences*, pp. 1726–1731, 2007.

Hauke R Heekeren, Sean Marrett, and Leslie G Ungerleider. The neural systems that mediate human perceptual decision making. *Nature reviews neuroscience*, 2008.

Anthony I Jang, Matthew R Nassar, Daniel G Dillon, and Michael J Frank. Positive reward prediction errors during decision-making strengthen memory encoding. *Nature human behaviour*, 2019.

Daphna Joel, Yael Niv, and Eytan Ruppin. Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, 15(4-6):535–547, 2002.

Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. A clockwork rnn. In *ICML*, 2014.

Ruth M Krebs, Bjorn H Schott, Hartmut Schutze, and Emrah Duzel. The novelty exploration bonus and its attentional modulation. *Neuropsychologia*, 47(11):2272–2281, 2009.

Mt Lengyel and Peter Dayan. Hippocampal contributions to control: the third way. In *NIPS*, pp. 889–896, 2008.

Tiago V Maia. Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, Behavioral Neuroscience*, 2009.

Tiago V Maia. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning and behavior*, 2010.

Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Vishnu P Murty, Oriel FeldmanHall, Lindsay E Hunter, Elizabeth A Phelps, and Lila Davachi. Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, 2016.

Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. Phased lstm: Accelerating recurrent network training for long or event-based sequences. In *NIPS*, 2016.

Klaus Oberauer. Design for a working memory. *Psychology of learning and motivation*, 2009.

John O'Doherty, Peter Dayan, Johannes Schultz, Ralf Deichmann, Karl Friston, and Raymond J Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *science*, 2004.

Ken A Paller and Anthony D Wagner. Observing the transformation of experience into memory. *Trends in cognitive sciences*, 2002.

Jan Peters and Christian Büchel. Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron*, 2010.

P Dayan PR Montague and TJ Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 1996.

Kanaka Rajan, Christopher D Harvey, and David W Tank. Recurrent network models of sequence generation and memory. *Neuron*, 2016.

Jamie D Roitman and Michael N Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience*, 22(21): 9475–9489, 2002.

Ranulfo Romo, Adrián Hernández, and Antonio Zainos. Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron*, 2004.

Michael N Shadlen and William T Newsome. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. *Journal of neurophysiology*, 2001.

Daphna Shohamy and Anthony D Wagner. Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, 2008.

Vikaas S Sohal, Feng Zhang, Ofer Yizhar, and Karl Deisseroth. Parvalbumin neurons and gamma rhythms enhance cortical circuit performance. *Nature*, 2009.

Roland E Suri, J Bargas, and MA Arbib. Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103(1):65–85, 2001.

David Sussillo. Neural circuits as computational dynamical systems. *Current opinion in neurobiology*, 2014.

Hugh R Wilson and Jack D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, 1972.

G Elliott Wimmer, Erin Kendall Braun, Nathaniel D Daw, and Daphna Shohamy. Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *Journal of Neuroscience*, 34(45):14901–14912, 2014.

Bianca C Wittmann, Björn H Schott, Sebastian Guderian, Julietta U Frey, Hans-Jochen Heinze, and Emrah Düzel. Reward-related fmri activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron*, 45(3):459–467, 2005.

A Continuous-time form of RNNs in B-ACEM

$$\boldsymbol{\alpha} = \sigma_g(\boldsymbol{W}_{\alpha}^{rec}\boldsymbol{r} + \boldsymbol{W}_{\alpha}^{in}\boldsymbol{u} + \boldsymbol{b}_{\alpha}), \tag{7}$$

$$\boldsymbol{\beta} = \sigma_g(\boldsymbol{W}_{\beta}^{rec}\boldsymbol{r} + \boldsymbol{W}_{\beta}^{in}\boldsymbol{u} + \boldsymbol{b}_{\beta}), \tag{8}$$

$$\tau\frac{d\boldsymbol{x}}{d\boldsymbol{t}} = -\boldsymbol{\alpha}\circ\boldsymbol{x} + \boldsymbol{\alpha}\circ(\boldsymbol{W}^{rec}(\boldsymbol{\beta}\circ\boldsymbol{r}) + \boldsymbol{W}^{in}\boldsymbol{u} + \boldsymbol{b} + \sqrt{2\tau\sigma_{rec}^2}\boldsymbol{\xi}), \tag{9}$$

$$\boldsymbol{r} = [\boldsymbol{x}]^+, \tag{10}$$

$$\boldsymbol{z} = \boldsymbol{W}^{out}\boldsymbol{r}, \tag{11}$$

where the vector $\boldsymbol{\xi}$ are $N = 256$ (the number of recurrent unit) independent Gaussian white noise process with unit variance and present noise intrinsic to the RNN, which are scaled by σ_{rec}. Threshold-linear activation function $[x]^+$ guarantees that Equation (7)-(11) is a nonlinear dynamic system. These leaky threshold-linear units are modulated by the time constant τ, with an update gate $\boldsymbol{\alpha}$ and reset gate $\boldsymbol{\beta}$.

B Discretization process of RNNs in B-ACEM

We mainly give the derivation from equation (9) to (4):

According to Euler's Method:

$$\frac{d\boldsymbol{x}}{dt} \approx \frac{\boldsymbol{x}_t - \boldsymbol{x}_{t-1}}{\Delta t} \tag{12}$$

The equation (7) can be written as

$$\boldsymbol{x}_t - \boldsymbol{x}_{t-1} = \frac{\Delta t}{\tau}[-\boldsymbol{\alpha}_t\circ\boldsymbol{x}_{t-1} + \boldsymbol{\alpha}_t\circ(\boldsymbol{W}^{rec}(\boldsymbol{\beta}_t\circ\boldsymbol{r}_t) + \boldsymbol{W}^{in}\boldsymbol{u}_t + \boldsymbol{b}_t + \sqrt{2\tau\sigma_{rec}^2}\boldsymbol{\xi}_t)], \tag{13}$$

Then we have

$$\boldsymbol{x}_t = \boldsymbol{x}_{t-1} + \frac{\Delta t}{\tau}\boldsymbol{\alpha}_t[-\boldsymbol{x}_{t-1} + (\boldsymbol{W}^{rec}(\boldsymbol{\beta}_t\circ\boldsymbol{r}_t) + \boldsymbol{W}^{in}\boldsymbol{u}_t + \boldsymbol{b}_t + \sqrt{2\tau\sigma_{rec}^2}\boldsymbol{\xi}_t)], \tag{14}$$

i.e.,

$$\boldsymbol{x}_t = (1-\boldsymbol{\lambda}_t)\circ\boldsymbol{x}_{t-1} + \boldsymbol{\lambda}_t\circ f(\boldsymbol{W}^{rec}(\boldsymbol{\beta}\circ\boldsymbol{r}_{t-1}) + \boldsymbol{W}^{in}\boldsymbol{u}_t + \boldsymbol{b}_t + \sqrt{2\frac{\tau\sigma_{rec}^2}{\Delta t}}\boldsymbol{N}(0,1)), \tag{15}$$

where $\boldsymbol{\lambda}_t = \frac{\Delta t}{\tau}\boldsymbol{\alpha}_t$, $\boldsymbol{\xi}_t = N(0,1)$, $f(x) = x$.

C Implementation Details

Table 1. Parameters for B-ACEM framework training.

Parameters List			
Parameter	Default value	Parameter	Default value
Learnig rate	0.001	Dropout	0.01
Size of actor/critic	256	Discount factor	0.99
Time step (Δt)	10 ms	max-gradient-norm	40
time constant (τ)	50ms	Recurrent noise(σ_{rec}^2)	0.01