Deciphering the contributions of episodic and working memories in increasingly complex decision tasks

Snigdha Dagar, Frédéric Alexandre, Nicolas P. Rougier

Deciphering the contributions of episodic and working memories in increasingly complex decision tasks

Snigdha Dagar
Inria Bordeaux Sud-Ouest
Univ. Bordeaux, CNRS
snigdha.dagar@inria.fr

Frederic Alexandre
Inria Bordeaux Sud-Ouest
Univ. Bordeaux, CNRS
Frederic.Alexandre@inria.fr

Nicolas Rougier
Inria Bordeaux Sud-Ouest
Univ. Bordeaux, CNRS
Nicolas.Rougier@inria.fr

***Abstract*—Augmenting the representation of the current state of the external world with internal states corresponding to working and episodic memories has been proposed as a bio-inspired solution to apply models of reinforcement learning to non-Markovian tasks. But, transposing these results to behavioral and experimental neuroscience, it is not completely clear how each of these memories can contribute to learning the augmented representations and when they must act in association for more complex tasks. Choosing an elementary implementation of these memories and experimental tasks of decision making in rodents, we explore these pivotal situations and make concrete the underlying mechanisms and criteria. We also specify cases where additional mechanisms must be envisaged.**

***Index Terms*—reinforcement learning, working memory, episodic memory**

I. Introduction

Learning and decision making are fundamental aspects of cognition and are closely linked. They have long been studied in animals through various levels of complexity in behavioral tasks.

Reinforcement Learning (RL) provides a theoretical framework for modeling tasks in which agents interact with their environment and learn rules by receiving reward signals upon taking actions, and has an undeniable biological basis [1]. It is nonetheless constrained by the Markovian property, stating that the decision can be directly made from the present state, not consistent with known characteristics of animal behavior in real world situations or in even cognitive tasks.

This class of partially observable Markov decision problems (POMDPs) has been solved by extending the present state (representing the state of the environment) with internal representations [2] that might correspond to memories built from previous experiences. A basic version of this principle has been proposed in [3] and related to biological basis, with the distinction between a working memory (associated with the prefrontal cortex) [4], where a given cue can be kept present in memory for some time, even if it disappears from the experienced world, and an episodic memory (associated with the hippocampus), where a previous episode (series of steps) can be recalled by similarity from the present state and manipulated as a virtual state.

On the computational and experimental neuroscience sides, a biological neural network framework was proposed [5] to explain how working memory representations in the PFC may be updated and maintained. This concept of gating models was also used to study rule acquisition in rats [6], by comparing the ability of two RL algorithms to replicate rat behavior. The ability and limitations of such a model in a common human behavioral task has also been demonstrated [7]. More recently, a simplification of this model has been extended [8] to include a bias that better explains the performance of rats in a spatial alternation task.

On the machine learning side, learning and exploiting these forms of memory have been adapted to RL [9], introducing complex representational and computational mechanisms that have also been compared in more details with human brain circuitry, thus introducing meta-RL [10] and episodic-RL [11] as new paradigms for addressing non Markovian problems. But this impressive level of performance is at the price of complex computations, requiring an often prohibitive training time (and correspondingly corpus size) and resulting in obscure computing phenomena, difficult to interpret and analyse in terms of functional contributions of the respective memory mechanisms.

What we propose in this paper is to design a study where a basic RL agent is extended with a minimal version of working memory and of episodic memory, that can be trained quickly and without hyper-parameters, and to define tasks where the usefulness of each memory can be analysed in details. Particularly, what we want to understand and share with our experimental neuroscientific colleagues are the conditions where one or the other kind of memory are needed and where they are not sufficient and should be complemented with more complex mechanisms. In other words, this explanatory study is the premise for predictions to be confirmed in forthcoming experimental studies in neuroscience and for precise specification to be implemented in more powerful learning algorithms for machine learning.

This work is partly supported by the project Ecomob, co-funded by Inria and by the french region Nouvelle Aquitaine.

II. Methods and Tasks

A. Computational model and architecture

a) Basic RL agent: A tabular, actor critic temporal difference learning architecture with ϵ greedy exploration was used. The agent maintains a table of state values V, where $V(s)$ is the agent's current estimation of expected, temporally discounted reward that will follow state s. The agent also maintains a table of action values Q, where $Q(s, a)$ is the value of taking action a in state s. The *actor* part of the architecture follows a simple policy where the agent picks the action with the highest Q-Value, except with a probability $0 \leq \epsilon \leq 1$ the agent selects an action at random.
In the *critic* part of the architecture, the TD error δ is computed when the agent takes an action a in state s and transitions to state s' after receiving a scalar reward r :

$$\delta = r + \gamma V(s') - V(s) \tag{1}$$

where $0 \leq \gamma \leq 1$ is a temporal discounting factor. The old state value and action value are then updated as :

$$V(s) \leftarrow V(s) + \alpha\delta \tag{2}$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha\delta \tag{3}$$

where α is a learning rate parameter.

The agent acts on the state space $S = S_L$ where S_L are all the possible location states an agent can find itself in (numbered squares in a T-shape grid world), by taking motor actions $A = up, down, right, left$ which can result in a change of the location state.

b) RL agent with Working Memory: : To include a working memory representation, the state space was augmented with an extra memory element. A factored state space representation was used : $S : S_L X S_{WM}$ with each tuple of the form (S_L, S_{WM})(grid world location, working memory contents) having its own set of action values. The total number of possible WM states is one more than the number of location states (one extra for empty memory at the start of the episode) i.e. $|S_{WM}| = |S_L| + 1$. The memory element is "hidden" as the agent can only access this state using the update action - which sets the working memory state to the current sensory location state, until it is overwritten when the next memory action is taken.

c) RL agent with Episodic Memory: : We include an abstraction of the episodic memory system in the model, which is content addressable and temporally indexed. The model maintains a history of the agent's *n* most recently visited states, in order. After each action that changes the agent's location state, the previous state is added to the end of the list, and the oldest state in the list is removed (no state is removed for the first *n* steps). The model now has a 3 element tuple for state representation. The factored state space $S = S_L X S_{EP} X S_{WM}$. The episodic memory state S_{EP} can take either one state from the episodic memory list or an additional state representing "nothing recalled". To interact with this memory system, the agent can take two actions - "cue retrieval" to find the most recent instance of the agent's current state in the list (and set S_{EP} to that state) or "advance retrieval" to set S_{EP} to the state following its current state in the list. This kind of abstraction allows the model to retrieve a specific episode from its past and replay the memory from the retrieved point.

B. Tasks

a) Task A : Discrimination: In the first, tactile discrimination task, the agent receives a sensory stimulus or a "cue" at the starting state. In this version of the task, it was a tactile cue about the surface, which could be rough on the right or left (represented by different states, as in Figure 1.B) and was indicative of the rewarding arm i.e. if the surface was rough on the right, the reward was placed at the end of the right arm while if it was rough on the left, the reward was placed at the end of the left arm. Thus, the agent's choice depended on learning this contextual or sensory rule. Another version of this task could be one where instead of a tactile stimulus, the agent could receive an auditory or odor stimulus in the starting state [12]. The important point is that this kind of sensory cue allows the rat or agent to form a distinct representation of the state

b) Task B : Alternation: The second task is a spatial alternation task, in the environment as shown in Figure 1.A. This class of tasks is widely used to study hippocampal and working memory (PFC) functions [13]. The agent begins in the bottom of the central hallways (square marked with a black dot) and proceeds up to the choice point. On the first (sample) trial, the agent can either turn left or right and receives a positive reward at the end of the arm (squares marked highlighted in black). The agent then continues along the return arm and back to the starting point (where it is prevented from entering the side hallway by a barrier). Following the first trial, the agent receives a positive reward if it chooses the opposite direction as on the previous trial, otherwise it receives a negative reward.

c) Task 3 : Radial Maze: In this task, there are three task conditions or contexts as represented in Figure 2 (left). In each task condition, the agent has the option of choosing only between 2 arms, with the rest of the arms blocked (the visual barriers being the contextual or sensory cue). For each of the contexts (A,B,C or D), the agent has to learn the alternating rule. In the trial phase, each context is presented once - A, B then C. The agent is free to go into any of the 2 open arms, and receives a reward at the end of the arm. Following the trial phase, the contexts are presented at random and the agent only receives a reward if it chooses the arm that it had not picked in the previous trial of the same context.

III. Results

A. Need for episodic memory

We first show, individually through the discrimination and alternation tasks, that the simple RL agent is unable to

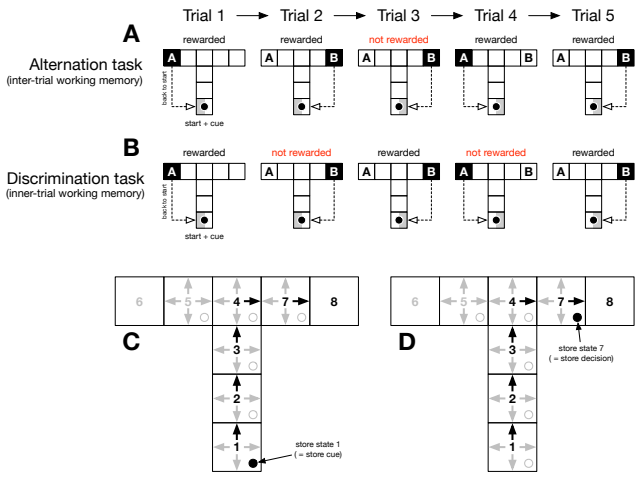


Fig. 1: **A** The alternation task requires for an agent to alternate its choice between two different options (A & B). In the displayed setup, the environment is a classical T-maze where the agent starts from the bottom location in order to reach A or B. The first choice is free and the reward is obtained after each alternation between A and B. After each trial, the agent restarts from the initial location. On this example, the 5 trials can be written as ABBAB and only transitions AB and BA are rewarded. This task implies for the agent to remember its previous choice across trials. **B** The discrimination task requires for an agent to learn which cue (out of two) is associated with a reward. The agent has to choose between the two different options (A & B) depending on a cue that is presented at the entrance of the T-Maze. This task implies for the agent to remember the initial cue until it reached the corresponding target during a single trial. **C** One example for the discrimination task where the agent has learned to memorize the location at the entrance. **D** One example for the alternation task where the agent has learned to memorize the location after its choice has been made.

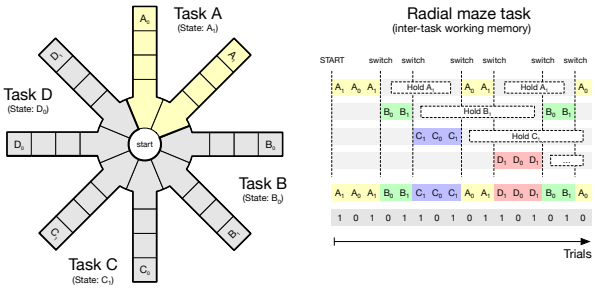


Fig. 2: The radial maze task corresponds to a contextualized alternation task with four different contexts (A, B, C, D) as illustrated on the left part of the figure. At any time, only one context X is open such that the agent has only to choose between X_0 or X_1. The difficulty however is to maintain a memory for a given context when the context is changed. To be able to successfully solve this global task, the agent has thus to maintain simultaneously four different memories corresponding to the four different states of the context.

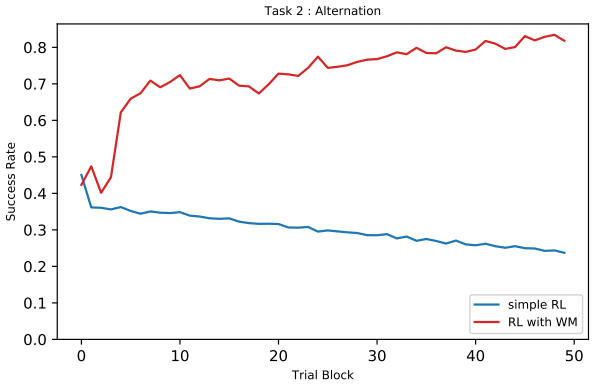
learn these tasks, remaining at chance performance for the discrimination task, and below chance for the alternation task. Performance in this task is plotted in Figure 3 (a) and (b). The RL agent with a working memory reaches an optimal performance by appropriately updating its internal memory at the right time. The agent learned different policies for each item that may be stored in the working memory. Accordingly, this can be interpreted as learning behaviors for a given context. In the discrimination task, the agent does this by buffering the sensory cue it received at the start state (or central arm), and maintaining it in memory until the choice point, thus disambiguating the trial type. In the alternation task, the agent had to buffer a location following the choice, and maintain this in memory until the choice point in the succeeding trial.

In the simulations, we could often observe that the agent could find the corresponding strategy and could learn to update and maintain the memory with the previous choice (for alternation) or the cue (for discrimination), whether left or right (accordingly, we call this strategy "remember both") (Figure 5 Right, Figure 6 Right). Interestingly enough, we could also observe sometimes the elaboration of another strategy, which is valid even if based on a side effect. In this "remember one" strategy, it is sufficient for the agent to remember only one choice (for alternation) or one cue (for discrimination) (Figure 5 Left, Figure 6 Left) and to simply label the other case by clearing the working memory. What is important at the end is to learn to associate the good action with a non ambiguous encoding of the state. The strategy adopted by the agent as a percentage over 100 runs is shown in Figure 4.

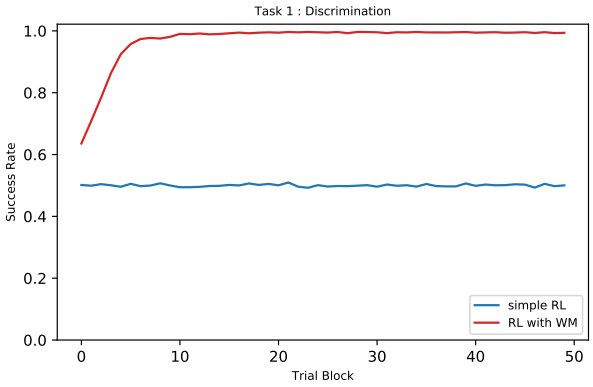
Next, we demonstrate the agent's behavior when the task is made more complex, and the agent has to learn the same behavioral rule of alternating between it's choices, but under 3 different contexts. We use the radial maze environment and show that a simple augmentation of working memory is not enough to learn this task. While in a task with only two contexts, the RL agent with one working memory element may be able to perform the task at slightly above chance performance, increasing the number of tasks quickly deteriorates the learning. A separate memory mechanism, or the episodic memory is needed to learn the presented task, as shown in Figure 3 (c). To perform this task correctly, the agent depends on its episodic memory to retrieve the last presentation of the current context, advance one step in the memory to discover which direction it had previously chosen, and then choose the opposite direction. It may be possible to solve this task by increasing the number of memory elements, but this would exponentially scale the state space, making value learning increasingly difficult. Learning to use working memory is an implicit kind of learning, which evolves over time as knowing *what* and knowing *when* to update and maintain while making the use of episodic memory is an explicit learning of recall.

B. Need for performance monitoring / Transfer of learning

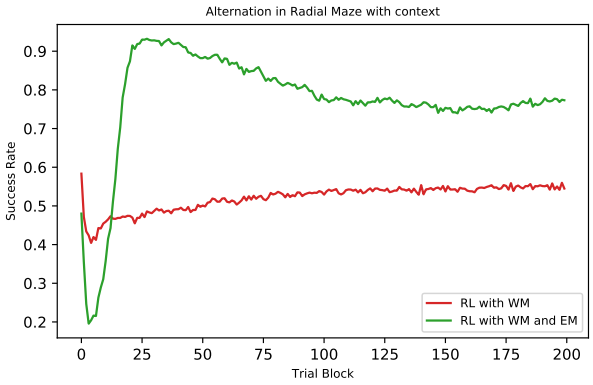
We analyzed the ability of the agent to adapt in a non-stationary environment. We tested if the agent is able to acquire the two distinct rules of task A (discrimination) and



(a)



(b)



(c)

Fig. 3: Performance of the agent in the **(a)** alternation task - a simple RL agents performs worse than chance as repeated visits to the same arm result in a negative reward **(b)** discrimination task **(c)** radial maze task. Each plot corresponds to the mean performance of 100 individual agents.

task B (alternation) by switching the underlying rule from one to the other, starting with task A, in the order ABAB. We show that after learning the first task, the agent is able to learn the second task (alternation), but not to an optimal level of performance (as shown in Figure 7). This is due to contextual interference from the first task, in which the

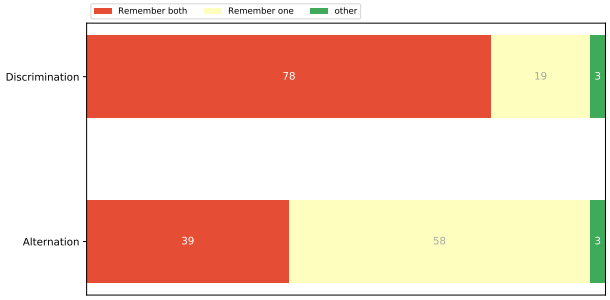


Fig. 4: Percentage of strategy types learned in the Alternation and Discrimination tasks. The proportion of trials in the final block in which the previous choice (alternation) or the cue (discrimination) was in memory at the choice point was calculated. A threshold of 2/3. was used, for example, **(a)** the number of trials in which cue1 was in memory / number of trials with cue1 was greater than threshold, strategy was ”remember cue1” (similarly for cue2) and **(b)** number of trials is which left was in memory / number of trials when previous choice was left was greater than threshold, strategy was ”remember left” (similarly for right); if proportions for both were above threshold, the strategy was ”remember both”; and if both were below threshold, it was ”remember other”.

agent learned to ’pay attention’ (i.e store in memory) to the presented cues. Nonetheless, due to the variation in the type of strategy used by the agent, it is able to reach a sub-optimal, but above chance level of performance. In the two rules we consider, the state spaces are only partially overlapping, and the difference in the policies is about *when* to update, thus the agent’s performance doesn’t drastically drop after the second switch. However in a situation where the rules are reversed [6] (for example, for the discrimination task - initially if cue1 rewards the right choice and cue2 rewards the left choice, a rule reversal would mean cue1 rewards the left choice and cue2 rewards the right choice), using the presented model would show such a sudden drop. This is because *(a)* the same set of Q-values are learned and updated continuously as the rules change, and *(b)* the model uses the same agent for learning motor and memory policies (as opposed to some multi-agent RL models). In any case, this model has the limitation of being unable to recall a previously learned rule. Hence, the best it can do is to identify a change in the environment, ’forget’ its currently held policy, or adapt its exploitation/exploration, and relearn its action preferences.

IV. Discussion

The current study explored the ability and constraints of a RL model, augmented with working and episodic memories, to model rule learning under increasing levels of task complexity. Our focus for this work was to identify the limits of a minimal RL model, and increasingly complement it with biologically plausible mechanisms to explain learning behavior.

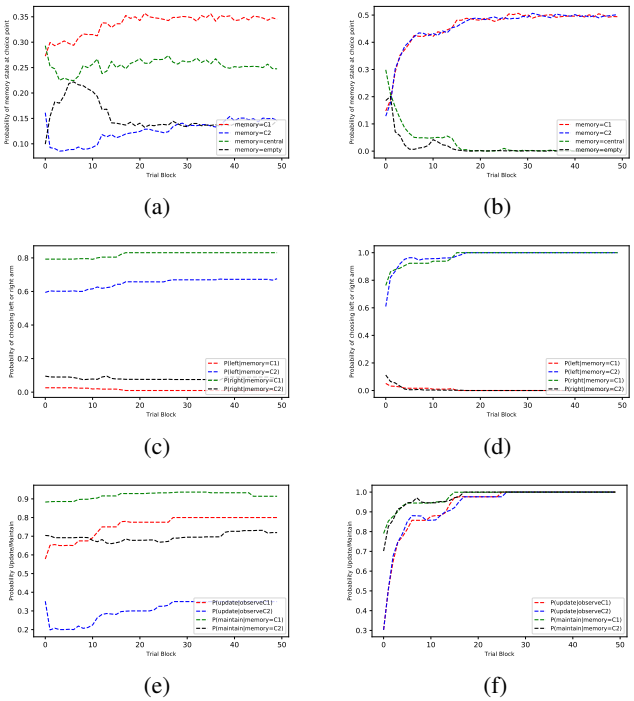


Fig. 5: **Examples of "remember one"(left) and "remember both"(right) strategies learned by the model in the discrimination task**. **(a) and (b)** Probability over blocks of different possible memory contents (cue1/2, central arm or empty) at the choice point. Since on average, half of the trials begin with an observation of cue1 and the other with an observation of cue2, the maximum proportion of trials that either of these can be in memory is around 50% **(c) and (d)** Probability over blocks of choosing to turn right or left as a function of different possible memory contents. Probabilities are derived from Q-values at the end of each block. **(e) and (f)** Probability over blocks of updating or maintaining memory contents conditional on (1) observing the cue (at the start state) or (2) having it already in memory (before the choice point). Probabilities are derived from Q-values at the end of each block

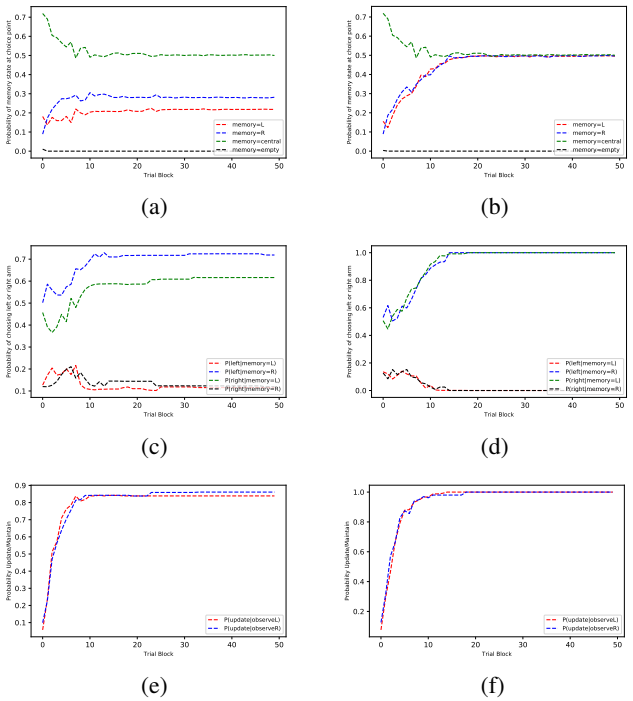


Fig. 6: **Examples of "remember one"(left) and "remember both"(right) strategies learned by the model in the alternation task.** Probabilities plotted are the same as described in Figure 5, with observations of cue1 and cue2 replaced with observations of turning left or right in the previous trial

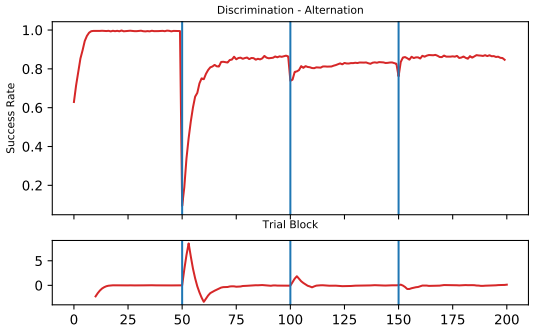


Fig. 7: The agent learns task A (discrimination) for the first 50 episodes, then task B (alternation) for the next 50 episodes, followed by a switch to task A and then back to task B. The presentation of the cues interferes with the learning of the task. A simple method of performance monitoring to identify a rule change can be maintaining a running mean of the rewards obtained over the last n episodes

Even though the three tasks considered here are not really ecological, they are indeed experimented with rodents by our neuroscientific colleagues and could be used to replicate the present findings and explore possible predictions. Particularly, they allow us to illustrate the necessity of having multiple memory systems (working memory and episodic memory) in this context. For the simple discrimination task, which shares a number of similarities with a regular delay match to sample (DMTS) task, we explained that the Markovian Property forbids a basic RL agent to solve the task and an additional hidden state (working memory) is necessary to reach the optimal behavior. The alternating task shares a lot of features with the discrimination task but the main difference concerning the working memory is not "what to store" but "when to store". We also showed that it is necessary for the model to store the state after the decision has been taken, that is, when the agent is close to one of the two end states. We also noticed that the model can efficiently learn to update and maintain working memory representations which can be intra-trial, inter-trial and inter-task.

Finally, the radial maze setup implies (in our

implementation) a different kind of memory (episodic memory) such as to be able to "upload" the right context to the working memory. An alternative implementation could have been to have four different working memories, one for each context. But even in such case, the core "routing" problem remains the same: depending on the context, the agent needs to store the memory at the right place.

If we now go back to behavior, we think that the radial maze task is actually quite representative of our day to day behavior. An agent already engaged in a given task may "pause" it in favor of another task provided it perceives some cue indicating that a new task can be advanced or completed. This kind of behavior requires in fact the temporal organization of behavior. Furthermore, even though the task is already quite complex, we nonetheless keep it simple by explicitly cuing the agent with the unambiguous identification of the context, contrarily to tasks such as the Wisconsin Card Sorting Task (WCST) [14] that require an effective cognitive effort to monitor performance to decide if the context has changed or not. It thus comes as no surprise that our human brains possess dedicated areas in the frontal cortex to monitor, select and instantiate such high level rules.

Much work has been done on dealing with non-stationary environments, in model-free algorithms [15] in multi-arm bandits, or model-based RL frameworks such as that proposed by [16] but the neural underpinnings for these are not yet clear. On the other hand, models for rule-learning and rule-switching have been proposed that rely on bayesian inference [17] [18], and while they provide a unifying theory for these concepts, their complexity makes them infeasible for guiding and testing fundamental hypothesis in experiments. Coming back to biological inspiration and building upon working memory and episodic memory a biologically inspired cognitive control [19] [20] could be an interesting way to define a more flexible decision making agent, able to quickly adapt in an uncertain and changing world, as we will study in the near future.

REFERENCES

[1] D. Lee, H. Seo, and M. W. Jung, "Neural basis of reinforcement learning and decision making," *Annual review of neuroscience*, vol. 35, pp. 287–308, 2012.

[2] L. Peshkin, N. Meuleau, and L. Kaelbling, "Learning policies with external memory," *arXiv preprint cs/0103003*, 2001.

[3] E. A. Zilli and M. E. Hasselmo, "Modeling the role of working memory and episodic memory in behavioral tasks," *Hippocampus*, vol. 18, no. 2, pp. 193–209, 2008.

[4] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv, "Orbitofrontal cortex as a cognitive map of task space," *Neuron*, vol. 81, no. 2, pp. 267–279, 2014.

[5] R. C. O'Reilly and M. J. Frank, "Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia," *Neural computation*, vol. 18, no. 2, pp. 283–328, 2006.

[6] K. Lloyd, N. Becker, M. W. Jones, and R. Bogacz, "Learning to use working memory: a reinforcement learning gating model of rule acquisition in rats," *Frontiers in computational neuroscience*, vol. 6, p. 87, 2012.

[7] M. Todd, Y. Niv, and J. D. Cohen, "Learning to use working memory in partially observable environments through dopaminergic reinforcement," *Advances in neural information processing systems*, vol. 21, pp. 1689–1696, 2008.

[8] D. B. Kastner, A. K. Gillespie, P. Dayan, and L. M. Frank, "Memory alone does not account for the way rats learn a simple spatial alternation task," *Journal of Neuroscience*, vol. 40, no. 38, pp. 7311–7317, 2020.

[9] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, "Reinforcement learning, fast and slow," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 408–422, 2019.

[10] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick, "Learning to reinforcement learn," *arXiv preprint arXiv:1611.05763*, 2016.

[11] S. Ritter, J. X. Wang, Z. Kurth-Nelson, and M. Botvinick, "Episodic control as meta-reinforcement learning," *bioRxiv*, p. 360537, 2018.

[12] N. J. Broadbent, L. R. Squire, and R. E. Clark, "Rats depend on habit memory for discrimination learning and retention," *Learning & Memory*, vol. 14, no. 3, pp. 145–151, 2007.

[13] L. M. Frank, E. N. Brown, and M. Wilson, "Trajectory encoding in the hippocampus and entorhinal cortex," *Neuron*, vol. 27, no. 1, pp. 169–178, 2000.

[14] D. Stuss, B. Levine, M. Alexander, J. Hong, C. Palumbo, L. Hamer, K. Murphy, and D. Izukawa, "Wisconsin card sorting test performance in patients with focal frontal and posterior brain damage: effects of lesion location and test structure on separable cognitive processes," *Neuropsychologia*, vol. 38, no. 4, pp. 388–402, 2000.

[15] B. C. Da Silva, E. W. Basso, A. L. Bazzan, and P. M. Engel, "Dealing with non-stationary environments using context detection," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 217–224.

[16] K. Doya, K. Samejima, K.-i. Katagiri, and M. Kawato, "Multiple model-based reinforcement learning," *Neural computation*, vol. 14, no. 6, pp. 1347–1369, 2002.

[17] K. Lloyd and D. S. Leslie, "Context-dependent decision-making: a simple bayesian model," *Journal of The Royal Society Interface*, vol. 10, no. 82, p. 20130069, 2013.

[18] A. Collins and E. Koechlin, "Reasoning, learning, and creativity: frontal lobe function and human decision-making," *PLoS Biol*, vol. 10, no. 3, p. e1001293, 2012.

[19] H. Eichenbaum, "Memory: organization and control," *Annual review of psychology*, vol. 68, pp. 19–45, 2017.

[20] A. Shenhav, J. D. Cohen, and M. M. Botvinick, "Dorsal anterior cingulate cortex and the value of control," *Nature neuroscience*, vol. 19, no. 10, pp. 1286–1291, 2016.