# The Emergence of Abstract and Episodic Neurons in Episodic Meta-RL

**Badr AlKhamissi** *
Sony CSL
Tokyo, Japan
badr [at] khamissi.com

**Muhammad ElNokrashy**
Microsoft ATL
Cairo, Egypt
muelnokr [at] microsoft.com

**Michael Spranger**
Sony CSL
Tokyo, Japan
michael.spranger [at] sony.com

## Abstract

In this work, we analyze the reinstatement mechanism introduced by Ritter et al. (2018) to reveal two classes of neurons that emerge in the agent's working memory (an epLSTM cell) when trained using episodic meta-RL on an episodic variant of the Harlow visual fixation task. Specifically, *Abstract* neurons encode knowledge shared across tasks, while *Episodic* neurons carry information relevant for a specific episode's task.

## 1 Introduction

Starting as a method to study animal conditioning in psychology (Pavlov, 1927; Rescorla & Wagner, 1972), *Reinforcement Learning* (RL) has become an efficient way to train artificial agents in solving complex tasks such as Go or StarCraft (Silver et al., 2016; Vinyals et al., 2019). Despite such successes important problems remain. State-of-the-art RL algorithms require enormous amount of training data and do not easily adapt to new tasks.

One research strand trying to address these issues is *meta-reinforcement learning* (meta-RL) (Wang et al., 2016) - in which agents have to learn to deal with a number of different tasks. Typically, in this work recurrent neural networks - specifically LSTMs - are used to learn representations that encode an RL algorithm. Ritter et al. (2018) proposed to extend these LSTMs with neural memory - so the agents are able to remember and retrieve knowledge gained over past tasks when re-encountering them. The memory proposed by Ritter et al. (2018) relies on a gating mechanism that decides which memory activations are retrieved and reinstated into the LSTM. This gating mechanism is trained as part of the overall optimization problem and constitutes a key artifact of learning.

We study how the gating mechanism interacts with LSTM neurons, and show that they can be roughly categorized. We identify two classes of neurons: *episodic* and *abstract* neurons - that differ in their characteristics w.r.t how information is restored as well as their impact on the performance of the system. *Abstract neurons* encode structural task knowledge relevant across different episodes, while *Episodic neurons* carry episode-specific environmental information (potentially reoccurring in later episodes).

This paper proceeds as follows: (1) we introduce a simplified version of the Harlow Task (a standard Meta-RL environment) with episodic cues, (2) we introduce the model implementation, followed by

[*] Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(3) an analysis, definition and tests for abstract and episodic neurons. Lastly, we briefly discuss our results in the wider context of meta-RL.

## 2 Methods

### 2.1 Task Formulation

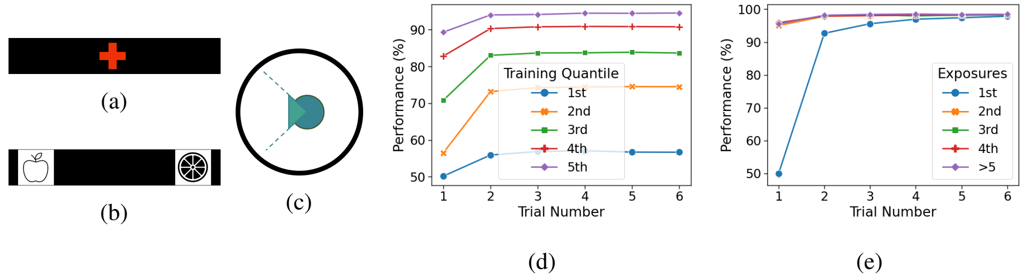


Figure 1: Illustration of the 1D Symbolic Harlow task and its training and testing performance. **(a)** Fixation cross at the center of agent's receptive field. **(b)** Objects placed in agent's receptive field upon fixation. **(c)** Top-down view of agent in the environment. **(d)** Average training performance at each trial number, per training quantile. **(e)** Testing performance at each trial number, per number of exposures to a specific task.

Let $M_i \in \mathcal{D}$ be a distribution of tasks each characterized as a Markov Decision Process (MDP): $M_i = (\mathcal{S}, \mathcal{A}, \mathcal{T}_i, \mathcal{R}_i)$. The agent learns over a sequence of MDPs by taking an action $a \in \mathcal{A}$ to transition from state $s$ to $s'$ (where $s, s' \in \mathcal{S}$) and receiving a scalar reward $r$, using some transition probability distribution $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. To introduce the concept of identifiably reoccurring tasks, Ritter et al. (2018) extend the previous formulation by associating a context $k_i$ with each $M_i$, sampled as $(M_i, k_i) \sim \mathcal{D}$ uniformly with replacement. With each new task, the agent can use the context $k_i$ to identify if the task had been seen before, and hence leverage previously discovered policies to avoid redundant exploration.

**The (One-dimensional) Symbolic Episodic Harlow Task** To study and analyze episodic meta-RL agents, we develop a simplified symbolic version with exact parallels to the task structure of the Harlow visual fixation task found in the PsychLab environment (Leibo et al., 2018) but which factors out the visual and spatial modeling of the environment. Further details can be found in Appendix A.

### 2.2 Model Description

The agent is trained[2] using the Advantage Actor-Critic (A2C) RL algorithm on a single thread (Mnih et al., 2015). The architecture follows that of the LSTM A3C model from Wang et al. (2016) but uses an epLSTM instead. The encoder is a stack of 2 affine layers with 64 and 128 units, respectively, and a ReLU non-linearity in-between. The epLSTM takes a concatenation of: **(a)** The encoding of the receptive field, **(b)** the reward at $t-1$, and **(c)** the action at $t-1$. The epLSTM is a one layer LSTM with $256$[3] hidden units plus the reinstatement mechanism. The memory module maps the context associated with the current task $k_i$ (as key) to the cell state $\mathbf{c}_T$ (as value) at the end of each episode (time $T$). This memory is updated at the same key each time the task $M_i$ reoccurs. The experiment is repeated $50$ times with different initializations. We analyze the top $30$ models (filtered by a threshold on the reward calculated on $100$ randomly generated episodes). Each instance is trained for $25,000$ episodes and tested for $1,000$ episodes (with different objects). The code is made open-source[4].

### 2.3 Reinstatement Mechanism

We use the reinstatement mechanism from Ritter et al. (2018)—defined as:

[2] All experiments were done on a single Nvidia GeForce RTX 2080Ti.

[3] Smaller models showed the same results as will follow, but took longer to converge.

[4] https://github.com/bkhmsi/emrl-neuron-emergence

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_{rx}\mathbf{x}_t + \mathbf{W}_{rh}\mathbf{h}_{t-1} + \mathbf{b}_r\right) \tag{1}$$

where $\mathbf{r}_t$ controls the flow of information from the retrieved memory $\mathbf{m}_t$ into the epLSTM cell state:

$$\mathbf{c}_t = \mathbf{i}_t \odot \tilde{\mathbf{c}}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1} + \underline{\mathbf{r}_t \odot \tanh\left(\mathbf{m}_t\right)} \tag{2}$$

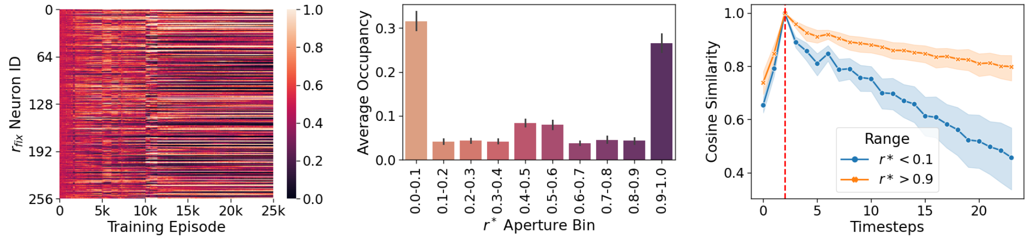
The vector $\mathbf{c}_t$ can be seen as encoding the agent's working memory state up to time $t$. Therefore, at the end of each episode the agent commits the accumulated knowledge learned about the current MDP $M_i$ to an external long-term memory module with the associated context vector $k_i$ as key. In a later episode where $M_i$ reoccurs, the context $k_i$ is used to query the memory to retrieve the corresponding cell state as $\mathbf{m}_t$ (equals the zero vector if $M_i$ is novel). The retrieval occurs when the agent first observes the pair of objects (after fixation at the first trial). The vector $\mathbf{m}_t$ is then interpolated into the current working memory using the reinstatement gate $\mathbf{r}_t$.

# 3 Analysis

## 3.1 Task Performance

Figure 1d shows the performance over successive stages of training as a function of the trial number. The performance on the first trial improves and is not stuck at random as in the classical Harlow experiment because the agent is able to reinstate relevant information when it re-encounters a specific task. Figure 1e shows the testing performance as a function of the number of times an agent is exposed to a particular task. It can be seen that when a task reoccurs the agent immediately identifies it and is able to solve it from the first trial.

## 3.2 Recurrent Neurons and the Reinstatement Gate



(a)

(b)

(c)

Figure 2: **(a)** The values of $\mathbf{r}_{\text{fix}}[j]$ for each neuron $j$ across training episodes. **(b)** Average percentage of neurons in $\mathbf{r}^*$ that appear within a certain bin of "openness" during testing across 30 different seeds. **(c)** Average vector similarity between $\mathbf{c}_{\text{fix}}$ at the first fixation in an episode (vertical line), and $\mathbf{c}_t$ at every other step.

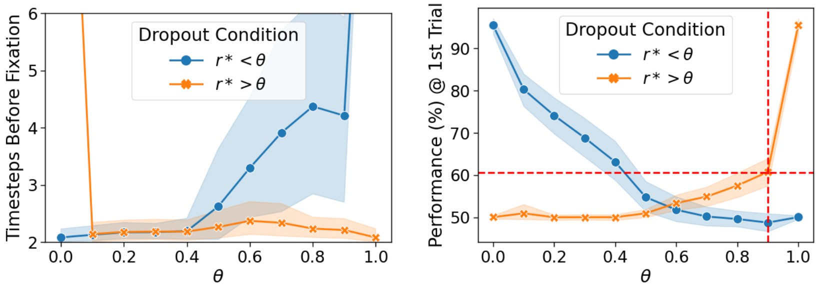
Noting that $\mathbf{r}_t[j] \in (0, 1)$ (by the $\sigma$ function) modulates the reinstatement of individual neurons from $\mathbf{m}_t$, we may interpret $\mathbf{r}_t[j]$ as the importance of neuron $j$ for recurring episodic information.

Figure 2a shows the values of $\mathbf{r}[j]$ at fixation (as heat) for each neuron $j$ across training episodes. Notice that the $\mathbf{r}[j]$ activations converge to stable values as training progresses, independent of changes in input or hidden state among the different episodes. We calculate from the last 1000 training episodes a static value for $\mathbf{r}_{\text{fix}}$ ($\mathbf{r}_t$ at $t$ = fixation) and use it for all further gate analysis. Formally ($e$ is the episode index and $N_e$ is the training episode count):

$$\mathbf{r}^* = \underset{e}{\text{mean}}\left\{\mathbf{r}^e_{\text{fix}} : N_e - 1000 < e \leqslant N_e\right\} \tag{3}$$

Figure 2b is a histogram of $\mathbf{r}^*[j]$ values. About 25% of the neurons have become biased to be open ($\mathbf{r}^*[j] \geqslant 0.9$), while ∼30% are biased to be closed ($\mathbf{r}^*[j] < 0.1$). Figure 2c shows the cosine similarity of certain regions (indicated by hue) of $\mathbf{c}_t$ to $\mathbf{c}_{\text{fix}}$ (averaged over 1000 testing episodes). This suggests that some specific neurons consistently hold the information needed across episodes to identify the winning object, and so change values less often within each episode (see Figure 3).

### 3.3 Testing for Abstract and Episodic Neurons



(a) Time steps

(b) Performance

Figure 3: **(a)** Average time-steps before fixating when dropping *episodic* ($\mathbf{r}^* \geqslant \theta$) or *abstract* ($\mathbf{r}^* < \theta$) neurons at $\theta$. **(b)** Average first trial performance at thresholds $\theta$. Note the $30\%$ regression when dropping *episodic* neurons at $\theta = 0.9$ (dashed lines).

To clarify the roles of individual neurons in $\mathbf{c}_t$, we test while gradually masking out the neurons $\mathbf{c}_t[j]$ based on $\{j : \mathbf{r}^*[j] \lesseqgtr \theta\}$ then analyze the behavioral change in two signals: **(a)** number of steps before fixation, and **(b)** first trial performance.

Figure 3a shows the average time-steps before fixation when zeroing out $\mathbf{c}_t$ based on the upper/*episodic* (orange) or lower/*abstract* (blue) regions of $\mathbf{r}^*$. Dropping more of the “abstract" region leads to worse performance (more steps to fixation). Performance remains largely stable as we drop “episodic" neurons first, up until before the extreme where all neurons are dropped ($\theta = 0.0$).

Figure 3b shows objective performance (choosing the rewarding object) at the first trial during testing where the MDP $M_i$ has occurred before. Dropping neurons from the “abstract" region (blue curve) shows a smooth drop in performance, while dropping “episodic" neurons shows a steep drop from as early as $\mathbf{r}^* > 0.9$, suggesting a strong correlation between the neurons where $\mathbf{r}^*[j] > \theta$ (for some reasonable $\theta$) and task performance on object re-occurrence. This seems to indicate that this region holds most of the object-based reward information.

## 4 Discussion

In this work, we have shown the existence of two classes of neurons that emerge in the memory reinstatement-based episodic meta-RL paradigm. Each neuron may belong to a class that encodes episodic information, or a class that encodes abstract knowledge that is shared across episodes. This finding implies that one does not need to store the whole cell-state when committing it to the long-term memory module since only a fraction of the activations are actually going to be reinstated. Therefore, one optimization method is to store a sparse representation of $c_T$ while storing the required indices only once. In the case of the experiments conducted in this paper, this method can save up to $75\%$ of the storage cost for the memory module while maintaining close to optimal performance after deployment once $\mathbf{r}^*$ is computed.

Wang et al. (2018) had shown that the meta-RL framework has direct connections with structures and functions in the brain. Specifically, they conceptualize the prefrontal cortex (PFC) along with the subcortical structures to which it connects as forming a homogeneous recurrent neural network that is trained using striatal dopamine reward prediction error signals. Inline with this theory and the work presented in this paper, previous work have shown that the PFC contain single neurons that encodes abstract rules (Wallis et al., 2001). Future work may extend the analysis to different episodic tasks, and utilize the findings for incorporating stronger inductive biases. We hope this work meaningfully furthers the sharing of insights between the neuroscience and machine learning fields.

**Acknowledgments**

We offer special thanks to Sam Ritter for many helpful suggestions and comments on the paper.

## References

Joel Z. Leibo, Cyprien de Masson d'Autume, Daniel Zoran, David Amos, Charles Beattie, Keith Anderson, Antonio García Castañeda, Manuel Sanchez, Simon Green, Audrunas Gruslys, Shane Legg, Demis Hassabis, and Matthew Botvinick. Psychlab: A psychology laboratory for deep reinforcement learning agents. *CoRR*, abs/1801.08116, 2018. URL http://arxiv.org/abs/1801.08116.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

I. P. Pavlov. *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford University Press, 1927.

RA Rescorla and Allan Wagner. *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, volume Vol. 2. 01 1972.

Samuel Ritter, Jane X. Wang, Zeb Kurth-Nelson, Siddhant M. Jayakumar, Charles Blundell, Razvan Pascanu, and Matthew M Botvinick. Been there, done that: Meta-learning with episodic recall. In *ICML*, 2018.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.

Oriol Vinyals, Igor Babuschkin, Wojciech Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John Agapiou, Max Jaderberg, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575, 11 2019. doi: 10.1038/s41586-019-1724-z.

Jonathan Wallis, Kathleen Anderson, and Earl Mller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411:953–6, 07 2001. doi: 10.1038/35082081.

Jane X. Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Z. Leibo, Dhruva Tirumala, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matt M. Botvinick. Learning to reinforcement learn. *ArXiv*, abs/1611.05763, 2016.

Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *bioRxiv*, 2018. doi: 10.1101/295964. URL https://www.biorxiv.org/content/early/2018/04/06/295964.

## A The Symbolic Episodic Harlow Task

The task consists of a one-dimensional circular state-space with 16 discrete cells, of which the agent can observe 8 cells at any time step (see Figure 1 for an illustration of the task). It starts with a central fixation cross placed in the initial observable space (i.e. receptive field) of the agent; similar to the PsychLab version. The agent can then select one of two actions $|\mathcal{A}| = 2$: move one cell to the *left* or to the *right*. After the fixation cross appears in the center of the agent's receptive field, it is removed and two objects are introduced to the left and right of the center, one of which is randomly assigned to be the rewarding object throughout the episode. The objects are uniformly sampled from a distribution of $n = 100$ objects split into 80 for training and 20 for testing, resulting in $n(n-1) = 9,900$ possible combinations of object-reward pairs, because in each task either of the object pair may be rewarding. The agent must then choose one of the objects by orienting it towards the center of its receptive field. Following that, the fixation cross reappears initiating the next trial.

Similar to the PsychLab version, one episode consists of $6$ trials, but here it is terminated after a maximum of $120$ total time steps. We use the same reward values as used in Wang et al. (2018): $1$ and $-1$ for the rewarding and non-rewarding objects, respectively, and $0.2$ for arriving at the fixation cross. The episodic structure comes from the fact that the same objects with their associated rewards can be sampled more than once. In order for the agent to identify the current task, a context vector is randomly generated the first time the agent encounters a particular task. This creates a unique mapping between each possible MDP and its corresponding context.