Attention Schema in Neural Agents

Dianbo Liu
Mila – Quebec AI Institute
National University of Signapore

Samuele Bolotta
Mila – Quebec AI Institute

Mike He Zhu
Mila – Quebec AI Institute
University of McGill

Zahra Sheikhbahaee
Mila – Quebec AI Institute
University of Montreal

Yoshua Bengio
Mila – Quebec AI Institute
CFAR AI Chair

Guillaume Dumas
Mila – Quebec AI Institute
University of Montreal

Abstract

Attention has become a common ingredient in deep learning architectures. It adds a dynamical selection of information on top of the static selection of information supported by weights. In the same way, we can imagine a higher-order informational filter built on top of attention: an *Attention Schema (AS)*, namely, a descriptive and predictive model of attention. In cognitive neuroscience, Attention Schema Theory (AST) supports this idea of distinguishing attention from AS. A strong prediction of this theory is that an agent can use its own AS to also infer the states of other agents' attention and consequently enhance coordination with other agents. As such, multi-agent reinforcement learning would be an ideal setting to experimentally test the validity of AST. We explore different ways in which attention and AS interact with each other. Our preliminary results indicate that agents that implement the AS as a *recurrent internal control* achieve the best performance. In general, these exploratory experiments suggest that equipping artificial agents with a model of attention can enhance their social intelligence.

1 Introduction

In deep learning, attention can be understood as a dynamical control of information flow Mittal et al. (2020). In the last decade, the scope of attention mechanisms has grown from the first implementation in RNNSearch Bahdanau et al. (2014) to the recent large-scale models currently dominating natural language processing and text-to-image generation Santana & Colombini (2021). Transformers have particularly demonstrated how *attention may be all we need*, from sequence learning Vaswani et al. (2017) to visual processing Dosovitskiy et al. (2020) and time series forecasting Lim et al. (2021).

In this regard, attention schema theory (AST) is a neuroscientific theory that postulates that the human brain, and possibly the brain of other animals, constructs a model of attention: an attention schema Graziano & Webb (2015b).Building a rich, flexible, and coherent model of attention is a complex task for AI, especially because to express all of its potential, language and higher-order cognition must have access to it Wilterson & Graziano (2021). In this paper, we are not trying to solve the entire problem but rather the foundation for future research on this topic. Specifically, we want to start tackling the following question: Can equipping agents with self-monitoring capabilities boost their ability to intelligently control and deploy their limited processing resources? To investigate this, we implement internal control as a recurrent network and attention as a key-value attention mechanism; we then compare five different possible relationships between attention and its internal control, as suggested by the cognitive science literature (Figure 1) Graziano & Webb (2015a). A strong prediction of AST is that an agent can use attention and its own internal control to also infer the states of the attention of other agents. As a consequence, this is expected to enhance coordination with other agents. Therefore, we test the five different hypotheses in multi-agent reinforcement learning

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

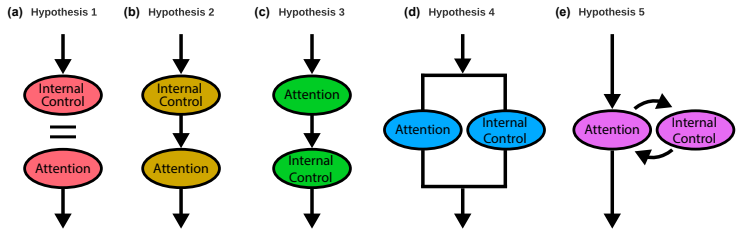

Figure 1: Hypotheses about the relationship between attention and internal control of attention. Adapted from Webb & Graziano (2015).

environments in which cooperation among agents is important. Our results suggest that higher performance is achieved by agents modeled according to the fifth hypothesis, which corresponds to AST and postulates the existence of an internal control serving as an inner regulation of the attention mechanism.

2 Internal control and attention

Potentially, model-based reinforcement learning offers numerous benefits over model-free reinforcement learning. Real-world samples are expensive, but learning a model of the environment can improve data efficiency, prompt targeted exploration, and promote better asymptotic convergence, which is crucial for scaling reinforcement learning (RL) to real-world applications Moerland et al. (2021). This paper is well aligned with this intuition: instead of simply focusing attention on various elements of the environment and then reinforcing those actions in which a higher reward was obtained, a simple model of attention can be used to exert greater control on behavior. In other words, our goal is to leverage the relationship between external stimuli, their representation in the brain of the agent (coordinated by attention), and the successful control of behavior (coordinated by internal control). To unravel the details of such a relationship, we set out five hypotheses about the relationship between attention and its internal control, as illustrated in Figure 1.

2.1 Internal control and attention are the same

Attention and its internal control are the same. The purpose of this setup is to act as a baseline. In humans, there is evidence of cases where attention is unfolding without its internal control being active, but this seems to correlate with poorer performance Graziano & Webb (2015b).

2.2 Internal control precedes attention

The system can only focus its attention on what has been modeled. In other words, the external stimulus is not attended immediately. There are two problems with this scenario: a) attention seems to be necessary to bind together the different components of a representation, which makes it impossible for the representation of a stimulus to be bound to the information contained in its internal control Treisman & Gelade (1980) b) without attention, the representation of a stimulus has lower signal strength and is less likely to have an influence on the policy Graziano & Webb (2015b).

2.3 Attention precedes internal control

The system can only model what is under the focus of attention. This hypothesis is close to what AST proposes, except for one aspect: internal control is not modeling attention in a recurrent way. This makes the system less flexible, and therefore less efficient in keeping up with ever-changing environments.

2.4 Internal control and attention are independent processes

The system simultaneously models the input and focuses attention on some sub-parts of the input. Both processes occur in a parallel fashion. Considering how closely related these two components are in humans, it is unlikely that implementing them as independent processes can lead to good performance.

2.5 Internal control of attention: the attention schema

This hypothesis is the most accurate representation of AST. The system models attention and can control it. The internal control corresponds to the attention schema: a coherent set of information that represents the basic properties of and changes in the state of attention in a dynamical fashion. This setup should offer the maximal potential for coordination between agents.

3 Background and notation

The whole Multi-Agent Reinforcement Learning (MARL) game is defined as $(N, S, A, P, R, O, \gamma)$. $N = \{1, ..., N\}$ denotes the set of N > 1 agents. $S_t = \{s_{1,t}, s_{2,t}, ..., s_{N,t}\}$ denotes the state of each agent at time step t. $O_t = \{o_{1,t}, o_{2,t}, ..., o_{N,t}\}$ denotes the partial observation of the environment that each agent receives as input at time step t. O_t could be different from S_t. A_{j_t} denotes the action of agent j at time step t . $P : S \times A \to \Delta(S)$ denotes the probability of state transition. $R^j : S \times A \times S \to R$ is the reward function that determines the immediate reward received by agent j for a transition from $(s; a)$ to s' and γ is the discount factor.

4 Methods

The five hypotheses were implemented using PyTorch Paszke et al. (2019), and tested on two tasks from two different MARL benchmarks: the GhostRun environment Jiang (2019) and the Multi-Agent Particle environment Mordatch & Abbeel (2017) (for more details, see Supplementary information).

4.1 Architectures

The attention module, the internal control module, and their interactions are implemented according to hypotheses 1-5 as described in Section 2. Attention modules are implemented as multi-head attention layers similar to the kind implemented by Transformers Vaswani et al. (2017); they dynamically select information from the observation space or, in the case of hypotheses 2 and 4, from the output of the internal control module. The internal control modules are implemented as recurrent neural networks (RNNs) with gated recurrent units (GRU), which take as input the time series signal or, in the case of hypotheses 2 and 4, the observation space; to make a fair comparison, the implementation of different hypotheses uses the same architectures of internal control of attention and other components that are shared by different hypotheses but without sharing parameters. All architectures are trained with proximal policy optimization (PPO). The formal definition of the architecture used to test each hypothesis follows.

Internal control module. The internal control module is implemented as an RNN. The input varies between hypotheses; for each element of the input sequence $x_{j,t}$, each layer computes the following function:

$$
\begin{aligned}
r_t &= \sigma(\mathbf{W}x_{j,t} + b + \mathbf{W}'h_{(t-1)} + b') \\
z_t &= \sigma(\hat{\mathbf{W}}x_{j,t} + \hat{b} + \hat{\mathbf{W}}'h_{(t-1)} + \hat{b}') \\
n_t &= \tanh(\mathbf{W}x_{j,t} + b + r_t * (\mathbf{W}'h_{(t-1)} + b')) \\
h_t &= (1 - z_t) * n_t + z_t * h_{(t-1)}
\end{aligned}
$$

Where h_t is the hidden state at time t, $x_{j,t}$ is the input at time t for agent j, $h_{(t-1)}$ is the hidden state of the layer at time t-1 and r_t, z_t and n_t are the reset, update, and new gates, and $\mathbf{W}\hat{\mathbf{W}}$, $\mathbf{W}$ are the corresponding weights respectively. σ is the sigmoid function, and $*$ is the Hadamard product.

Attention module

The keys (K), values (V), and queries(Q) vary across the hypotheses. Attention is computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where d_k is the dimension of the vectors **Hypothesis 1** (Figure 1a and Supplementary figures) postulates that attention and internal control are the same. Therefore, no internal control module is implemented. The partial visual observation of an agent j at time step t is rearranged into different patches as matrix $o^{pa}_{j,t}$, which has shape (number of patches $\times$ number of pixels per patch). The attention key is $K_j = \mathbf{W^k_j} o^{pa}_j$; the attention value is $V_j = \mathbf{W^v_j} o^{pa}_j$; the attention query q_j is the vector o^{pa}_j after taking the mean across the first dimension (number of patches). This design is

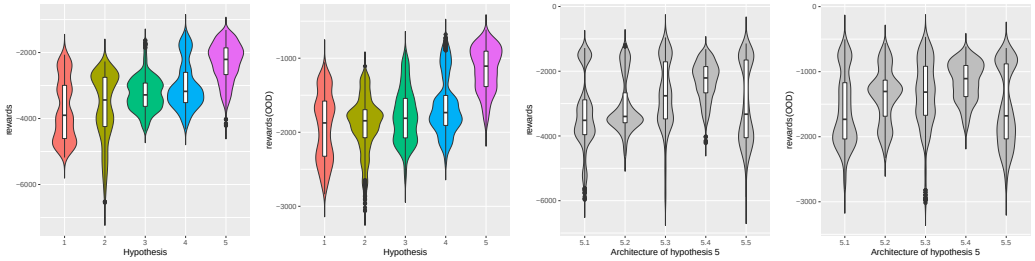

Figure 2: Comparison among the different hypotheses in "GhostRun" cooperative multi-agent reinforcement learning environment (preliminary results). We compared their rewards in the final 100 episodes (900-1000) in a testing environment that is the same as the training environment (IID) and a testing environment that is different from the training environment (OOD). The results are from 10 different random seeds. From left to right: (a) IID test rewards of main five hypotheses, (b) OOD test rewards for the five main hypotheses OOD. (c) IID test rewards for five alternative architectures for Hypothesis 5 and (d) OOD test rewards for five alternative architectures for Hypothesis 5.

motivated by the idea that an agent will decide where to focus based on the whole view it observes. Attention mechanisms in all other hypotheses follow the same strategy. The output vector of the attention module is $h1_{j,t}$, which is a weigthed sum of values by attention scores, and is directly fed into the policy network in Hypothesis 1.

Hypothesis 2 (Figure 1b and Supplementary figures) postulates that internal control precedes attention. The output vector $h2_{j,t}$ of the internal control module is used as an input for attention, which outputs $h1_{j,t}$ and feeds it directly into the policy network.

Hypothesis 3 (Figure 1c and Supplementary figures) postulates that attention precedes internal control. The output vector $h1_{j,t}$ of the attention module is used as input for the internal control module, which outputs $h2_{j,t}$ and feeds it directly into the policy network.

Hypothesis 4 (Figure 1d and Supplementary figures) postulates that internal control and attention are independent processes. Attention is calculated in the observation space and outputs the vector $h1_{j,t}$, which is directly fed into the policy network. In parallel, the internal module is applied to each element in the input sequence $o^{pa}_{j,t}$ and outputs the vector $h2_{j,t}$, which is directly fed into the policy network.

Hypothesis 5 (Figure 1e and Supplementary figures) postulates that the internal control corresponds to the attention schema. The internal control module is implemented as a recurrent neural network (RNN). However, the output is not used as an input to the policy. Instead, the internal control module learns to predict the results of attention via contrastive loss. In the diagram, $h1_{j,t}$ is the attention output vector and $h2_{j,t}$ is the RNN output from internal control. In this hypothesis $h2_{j,t}$ is used to predict the final attention output $h1_m$ which is used as input into the policy network via a constrastive loss. See supplementary figures for details.

$$ContrastiveLoss = (h1'_m - h1_m)^2$$

where $h1'_m$ is the vector predicted from $h2_{j,t}$. In addition, $h2$ is used to generate a binary mask that is applied on the attention scores in the attention mechanism resulting in: $AttScore' = AttScore \circ M_{att}$ where $M_{att} \in [0,1]^d_{att}$ is the binary mask and $\circ$ refers to element-wise multiplication. The mask is generated from a pair of MLPs inspired by a neural activator and suppressor and followed by a Gumbel softmax to binarize the value pairs Jang et al. (2017), where the activator and suppressor work together to generate a binary mask on attention scores in which a "1" indicates that the corresponding attention score is allowed to be active and a "0" means that the attention is suppressed.

Next, we ablated and rearranged different components in the architecture of hypothesis five. While hypothesis 5.4 has already been described previously, these are the other four versions of hypothesis five (see Appendix):

Architecture 5.4 is the one that best reflects what is proposed by attention schema theory because it is the only one in which control is exerted on attention. Our exploratory experiments suggest that hypotheses 5.4 and 5.5 show the best results (Figure 2).

5 Discussion and Future Work

In this paper, we compared five different possible relationships between attention and its internal control, as suggested by the cognitive science literature (Figure 1) Graziano & Webb (2015a). We have shown, through a sequence of experiments performed in MARL environments, that equipping agents with an internal control that models their attention helps coordination across agents to achieve a common goal.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Graziano, M. S. and Webb, T. W. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in psychology*, pp. 500, 2015a.

Graziano, M. S. A. and Webb, T. W. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500, 2015b. doi: 10.3389/fpsyg.2015.00500. URL https://www.frontiersin.org/article/10.3389/fpsyg.2015.00500.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv:1611.01144 [cs, stat]*, Aug 2017. URL http://arxiv.org/abs/1611.01144. arXiv: 1611.01144.

Jiang, S. Multi agent reinforcement learning environments compilation, 2019.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021.

Mittal, S., Lamb, A., Goyal, A., Voleti, V., Shanahan, M., Lajoie, G., Mozer, M., and Bengio, Y. Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. In *International Conference on Machine Learning*, pp. 6972–6986. PMLR, 2020.

Moerland, T. M., Broekens, J., and Jonker, C. M. Model-based reinforcement learning: A survey. *arXiv:2006.16712 [cs, stat]*, Feb 2021. URL http://arxiv.org/abs/2006.16712. arXiv: 2006.16712.

Mordatch, I. and Abbeel, P. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Santana, A. and Colombini, E. Neural attention models in deep learning: Survey and taxonomy. *arXiv preprint arXiv:2112.05909*, 2021.

Treisman, A. M. and Gelade, G. A feature-integration theory of attention. 12:97–136, Jan 1980. doi: 10.1016/0010-0285(80)90005-5. URL https://www.sciencedirect.com/science/article/pii/0010028580900055.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Webb, T. W. and Graziano, M. S. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in psychology*, 6:500, 2015. URL https://doi.org/10.3389/fpsyg.2015.00500.

Wilterson, A. I. and Graziano, M. S. The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences*, 118(33), 2021.

A Supplementary information

A.1 Tasks

The experiments aim at investigating which combination of attention and its internal control achieves the best performance in a multi-agent cooperative context. Therefore, we tested the five hypotheses in Figure 1 in multi-agent reinforcement learning (MARL) tasks where coordination is required to achieve a cooperative goal.

We used two different environments where agents have a visual observation of their surroundings (RGB matrix of observed pixel values) and a discrete action space of movement (up/down/left/right).

GhostRun Environment. The GhostRun environment consists of multiple agents, each with a partial view of the environment. The space consists of ghosts, represented by red dots, trees, represented by green dots, and obstacles, represented by black dots. The ghosts move around randomly, whereas the trees and obstacles are stationary. The task for the team of cooperative agents is to escape from ghosts, or more quantitatively, to minimize the number of ghosts in each agent's partial observation of the environment. The reward received by each agent at each time step is the negative of the total number of ghosts in the view of all agents and a step cost of -1 for each step taken (Figure 3).

MazeCleaners Environment. In MazeCleaners, agents are moving in a maze and need to clean it cooperatively as quickly as possible. Agents receive a collective reward for each part of the maze cleaned up (Figure 4). Agents cannot walk through walls. The world size is 13 by 13, and the current scenario focuses on 2 agents. For OOD performance, the spawning location of the agents is random.

B Experimental design and preliminary results

The experiments were designed to explore which combination of attention and its internal control achieves the best performance in a multi-agent cooperative context. Therefore, we tested the five hypotheses in Figure 1 in multi-agent reinforcement learning (MARL) tasks where coordination is required to achieve a cooperative goal. Our preliminary results (see Appendix) suggest that hypothesis 5 shows marginal advantages over other hypotheses.

C Different version of hypothesis 5

- Hypothesis 5.1: There is no binary mask. In other words, there is no control.

- Hypothesis 5.2: The binary mask is applied to the action distribution predicted by the policy network. In other words, the control is on action and not on attention.

- Hypothesis 5.3: A binary mask is applied to the output of attention $h1_m$ rather than directly interfering with the attention mechanism. In other words, the control is on the output of attention and not on the attention itself.

- Hypothesis 5.5: It uses the prediction $h1'_m$ made by the internal control module as input into the policy network.

D Experiments

Experiments were conducted to investigate which of the five hypotheses show an advantage over others in MARL environments.

D.1 Comparison between the five main hypotheses

In the first set of experiments, we compared the performance of the five architectures in the "GhostRun" MARL environments. The models obtained from training were tested both in an environment that is the same as the training setting (Independent and Identically Distributed or IID testing) and in an environment with distributional shifts (Out-Of-Distribution or OOD testing). Our

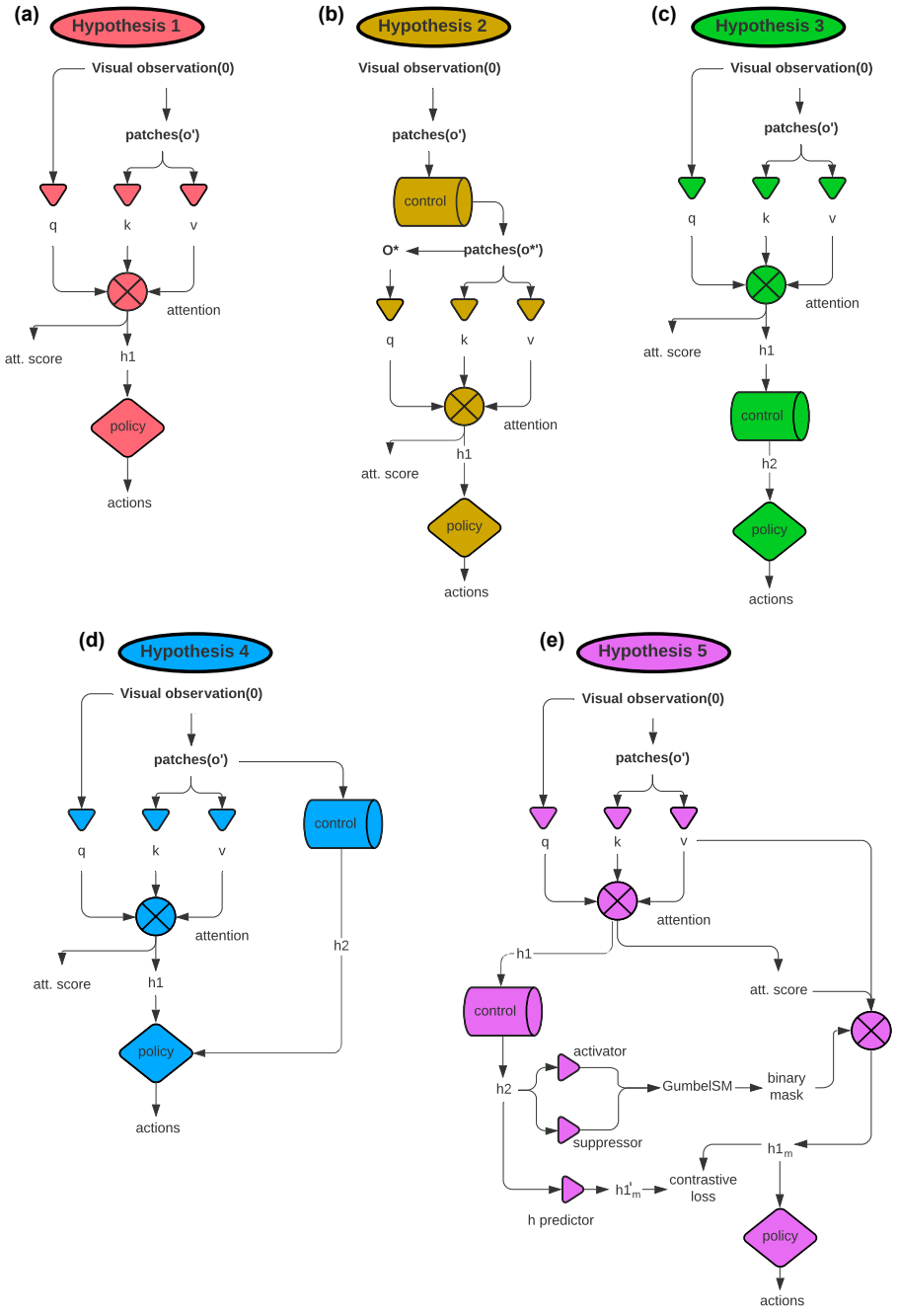

Figure 3: Architecture of the five hypotheses. "q" denotes the attention query; "k" denotes the attention key; "v" denotes the attention value; "h1" denotes the output of attention; "h2" denotes the output of internal control.

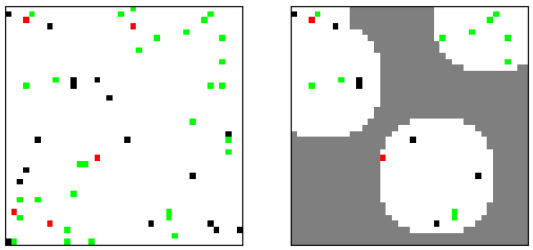

Figure 4: GhostRun environment: The agents need to keep away from ghosts (red pixels) as much as possible. The environment also contains trees (green pixels) and walls (black pixels). The right panel shows only the parts of the environment visible to the agents.

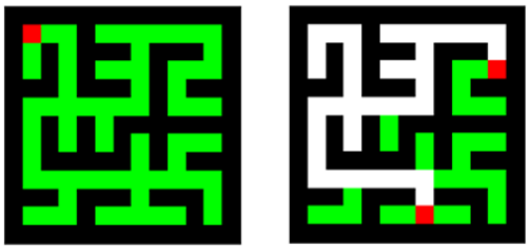

Figure 5: MazeCleaners environment: The agents (red pixels) need to clean the green parts of the maze as fast as possible while navigating between the walls (black pixels). The right side shows the maze on the left side after 30 epochs.

results suggest that hypothesis five, corresponding to the attention schema theory, achieves the best performance (Figure 5). The results are from 10 different random seeds.

D.2 Ability to generalize in a continuous learning setting

We compared the performance of the different architectures in a continuous learning setting, in which the difficulty of the GhostRun task increases every 50 episodes — with the addition of an additional ghost (Figure 6).

D.3 Replication in different task and environment

We replicated the main results of "GhostRun" in the "MazeCleaners" environment (Figure 7). Here, Hypothesis 5.4 also gave the best performance among all architectures.

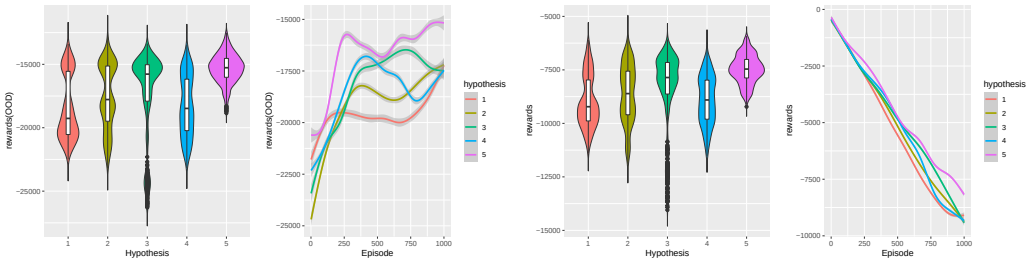

Figure 6: Preliminary results of performance of models in "GhostRun" corresponding to different hypotheses in a continual learning setting. From left to right: (a) reward distribution during learning, (b) Timecourse of the learning reward across episodes, (c) reward during OOD evaluation, and (d) timecourse of the OOD test reward across episodes. Notice the common decrease in reward since the number of ghosts to escape from is increasing with episodes.

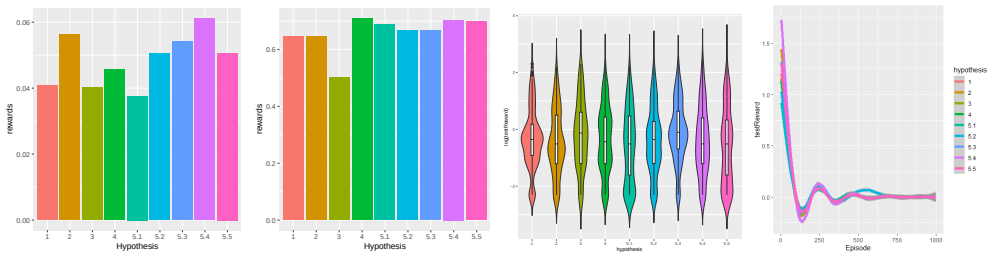

Figure 7: Preliminary results of performance of models in "MazeCleaners" corresponding to all the different hypotheses. Leftmost panel: reward during learning. Second from left panel: reward during OOD evaluation. Third from left panel: Boxplot of test rewards. Right panel: Timecourse of the test reward across episodes.