A Computational Model of the Hippocampus Supports Exploratory Behaviour in Reinforcement Learning Agents

Sabrina Du[1], Adel Halawa[1,2], Aleksei Efremov[1,2], Adrien Peyrache[1], Daniel Levenstein[3], Blake Richards[1,2,3,4,5]

[1]Montreal Neurological Institute, McGill University, Montreal, QC, Canada
[2]Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada
[3]Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA
[4]School of Computer Science, McGill University, Montreal, QC, Canada
[5]Learning in Machines and Brains Program, CIFAR, Toronto, ON, Canada

Keywords: Cognitive maps, Intrinsic reinforcement learning, Curiosity-driven exploration

Corresponding author email: sabrina.du@mail.mcgill.ca

Published: 05/05/2026
https://doi.org/10.26443/msurj.v21i2.437

The hippocampus supports spatial navigation, memory, and planning through the formation of cognitive maps—structured environmental representations reflected in neural activity. These neural dynamics can be modeled computationally using recurrent neural networks (RNNs) to provide insights into how cognitive maps guide behavior. However, these RNNs are typically trained via reinforcement learning (RL) using external rewards, failing to capture the intrinsic drive of freely exploring animals in the absence of external rewards. Instead, reward-free RL models, which rely on internal environmental representations, are better candidates to study novelty-seeking and exploratory behaviour. This study aims to investigate whether an RNN exhibiting hippocampal-like activity builds spatial representations sufficient to support exploratory behavior in reward-free RL agents. We leveraged an existing RNN trained for sensory sequence prediction, which exhibits hippocampal-like activity patterns, and used its prediction error as the intrinsic reward to train an Actor-Critic agent. Performance was evaluated using a Novel Object Recognition task to quantify its preference for novel versus familiar stimuli. The RL agent occupied the region of interest (defined as a 3-unit radius around the novel object) significantly more often than a random agent across multiple episodes and novel object locations. The RL agent's performance was also measured in a multi-room environment, where its visitation frequency to novel rooms was significantly higher than a random control. This work demonstrates that hippocampal-like representations can support autonomous exploratory behaviour, and provides a framework for investigating how cognitive maps guide exploration and navigation.