Uncovering brain-wide planning strategies with deep RL: Lessons from the Tower of Hanoi

A.T.D. Andrews (austin.andrews@reuben.ox.ac.uk), **M. Garibbo (michele.garibbo@dpag.ox.ac.uk)**

J. Achterberg (jascha.achterberg@dpag.ox.ac.uk), **R.P. Costa (rui.costa@dpag.ox.ac.uk)**
Department of Physiology, Anatomy and Genetics, 13 Mansfield Road
Oxford, The United Kingdom

Abstract

Human planning involves generating and executing action sequences under environmental constraints (Mattar & Lengyel, 2022). Experimental studies have identified that areas such as the prefrontal cortex (PFC) , hippocampus and cerebellum play important roles during planning (Grafman et al., 1992; Goel & Grafman, 1995). We propose that the architectures of deep reinforcement learning agents capable of solving human-level planning tasks can offer a normative framework for understanding the involvement of different brain regions in planning. To demonstrate this, we use MuZero (Schrittwieser et al., 2020) and a widely used task to study goal-directed planning and behavior, Tower-of-Hanoi (ToH). We evaluate the performance of MuZero on the ToH under targeted network ablations to simulate brain region-specific lesion studies. Ablating the value network reproduces the behavior observed in patients with PFC damage, while ablating the policy network mimics cerebellar damage. Our preliminary results suggest that deep RL architectures may provide a brain-wide account of human planning.

Keywords: Deep Reinforcement Learning; Planning; Tower of Hanoi; Prefrontal Cortex; Cerebellum

Introduction

The Tower of Hanoi (ToH) is a problem-solving task used to assess planning and executive function. It involves three pegs and several discs of increasing size stacked in order on one peg (Fig.1A). The goal is to move the stack to a target peg, following two rules: only one disc may be moved at a time, and larger discs cannot be placed on smaller ones. Solving the puzzle in the minimum number of moves ($2^n - 1$ for n discs) requires anticipating future states and managing subgoals without violating the rules, making ToH a useful tool for studying planning in humans and artifical systems.

Previous literature has examined ToH performance in patients with prefrontal cortex (PFC) lesions (Goel & Grafman, 1995) and cerebellar atrophy (Grafman et al., 1992) Fig.1C. Patients with PFC lesions performed comparable to control subjects in problems they successfully solved; however, they completed fewer complex problems overall (Goel & Grafman, 1995). This impairment has been suggested to be due to the conflict between the goal and the subgoal in more complex instances of the task. Patients with cerebellar atrophy were found to take longer than controls to solve the problems and also solved fewer problems in total (Grafman et al., 1992).

Deep RL for planning

MuZero is a model-based reinforcement learning agent (Schrittwieser et al., 2020) that integrates planning through Monte Carlo Tree Search (MCTS) and learning through four artificial neural networks (Fig.1B):

The Representation network encodes a sequence of past observations into a latent state. This latent state serves as the initial root for planning. It does not try to reconstruct the environment but rather abstracts features into those useful for further planning.

The dynamics network models the change in the world state given a particular action. Given a latent state s_{k-1} and an action a_k the network predicts the next latent state s_k and an immediate reward r_k. This allows MuZero to predict the trajectory of its actions without actually changing the environment.

The Policy network provides a fast heuristic for driving action selection. Given a latent state s_k it returns a probability distribution over all possible actions p_k, reflecting which actions may be optimal from the current (latent) state.

The value network evaluates the quality of each latent state. Given a latent state s_k it returns the predicted long-term return from that state v_k (a value estimate).

The Monte Carlo Tree Search (MCTS) is used to plan every move. Each node in the tree corresponds to a latent state. The root node is generated by passing the observation history through the representation function. The prediction function is then used to estimate the initial policy and value. The tree is traversed by selecting actions that maximize an upper confidence bound. When a leaf node is reached, the dynamics function is used to simulate an action's outcome and the prediction function is applied to the resulting state. The new node is added to the tree and the value and reward from the new node are backed-up along the tree path. These steps are repeated for a fixed number of simulations, at which point the action at the root with the highest number of visits is selected.

Deep reinforcement learning systems like MuZero offer a promising model of human planning. They perform well on complex tasks (e.g., Go, Chess, Atari), and their modular architecture may parallel the brain's division of function and communication. MuZero also blends fast, intuitive decisions with slower, deliberative planning, mirroring human decision-making dynamics as it shifts from search to learned policies over time (Moskovitz, Miller, Sahani, & Botvinick, 2022).

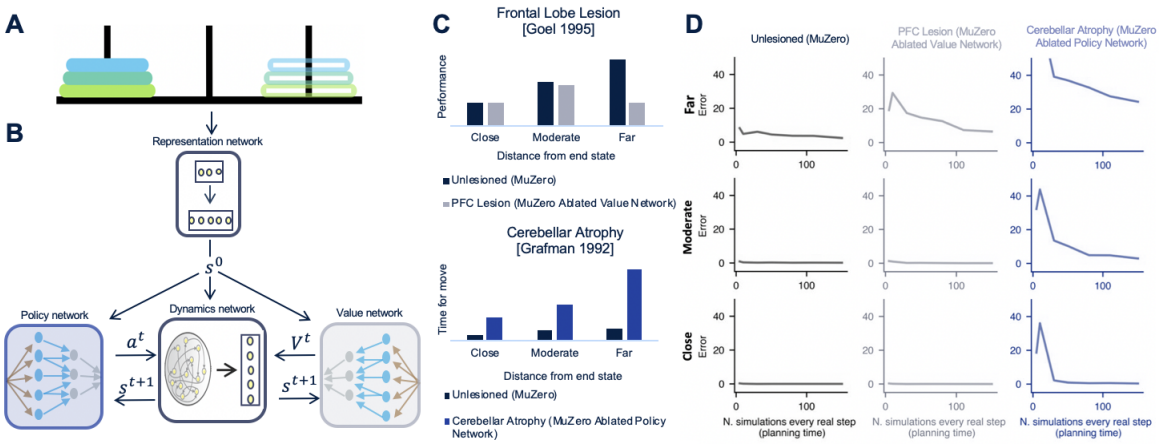

Figure 1: **MuZero accounts for PFC and cerebellar deficits in Tower-of-Hanoi.** **A**: A potential starting position for a ToH puzzle where the solid colors indicate the starting position and the striped the goal. **B**: Schematic of MuZero's multiple networks. **C**: Summary of human PFC and cerebellar lesion impact on the tower of hanoi. **D**: Performance comparison between full MuZero (black), MuZero with the value network ablated (red) and MuZero with the policy network ablated (blue). Note the different y-scales.

Results

We trained an implementation of MuZero on vector encoded representations of the ToH task. The available moves at any point are presented as a one dimensional array. ToH problems are either close, moderate or far from the goal in terms of the number of moves needed to complete the puzzle. The model is first trained and then evaluated on each category of problems with a different number of simulations allowed. This process is repeated, loading the same trained model, resetting the value network and then the policy network to random weights in order to ablate them.

Value network and PFC lesions: The ablation of MuZero's value network resulted in similar performance on problems of low to moderate complexity (Close, Moderate), but with an order of magnitude more errors on the most complex problems (Far) (Fig.1D). This pattern is similar to the pattern observed in patients with PFC lesions which showed the same performance on problems that were solved but much worse performance on more complex problems. This is indicated by having to allow Muzero to run a much higher number of planning simulations to achieve the same level of performance on complex problems as when the Value network is not ablated.

Policy network and cerebellar lesions: The ablation of policy network dramatically reduced performance with more illegal or unproductive moves in all classes of problems with an over ten times performance decrease (Fig.1D). The performance eventually reaches the same level as Muzero with the unablated policy network when given enough time to run additional simulations. This is similar to cerebellar atrophy patients requiring more time for the task in order to be successful.

Discussion and future work

MuZero offers a novel framework for modeling the involvement of specific brain regions in planning. Our ablation experiments show parallels with profiles observed in clinical populations. Patients with PFC lesion show reduced performance in the most complex problems, which has been linked to goal–subgoal conflict (Goel & Grafman, 1995). Our ablation of the value network supports this interpretation, as it disables the model's ability to estimate long-term reward. Without long-term value estimation, actions that appear immediately counterproductive (e.g., backward moves) become more difficult to justify or select. Cerebellar atrophy patients display increased completion times for successful tasks (Grafman et al., 1992). Our ablation of the policy network yields a similar pattern: with efficient action selection impaired, the model must consider more options and simulate more extensively to achieve comparable results.

These results support the idea that MuZero captures meaningful aspects of human cognitive-neuroscience function. Specifically, its value network appears analogous to executive planning functions in the prefrontal cortex, while its policy network mirrors cerebellar contributions to motor coordination and action efficiency. Additionally, the number of simulations used in MCTS offers a computational analogue to human lookahead depth in pre-action deliberation.

Though MuZero aligns with human planning in the ToH, it is unclear if this extends to other tasks. Future research should test it on tasks like the Tower of London or more complex tasks to further validate Muzero as a normative model of human planning.

Acknowledgements

ATDA is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1).

References

Goel, V., & Grafman, J. (1995, May). Are the frontal lobes implicated in “planning” functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia*, *33*(5), 623–642. doi: 10.1016/0028-3932(95)90866-P

Grafman, J., Litvan, I., Massaquoi, S., Stewart, M., Sirigu, A., & Hallett, M. (1992, August). Cognitive planning deficit in patients with cerebellar atrophy. *Neurology*, *42*(8), 1493–1493. (Publisher: Wolters Kluwer) doi: 10.1212/WNL.42.8.1493

Mattar, M. G., & Lengyel, M. (2022). Planning in the brain. *Neuron*, *110*(6), 914–934.

Moskovitz, T., Miller, K., Sahani, M., & Botvinick, M. M. (2022). A unified theory of dual-process control. *arXiv preprint arXiv:2211.07036*.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... Silver, D. (2020, December). Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, *588*(7839), 604–609. (arXiv:1911.08265 [cs]) doi: 10.1038/s41586-020-03051-4