Limitations in the Planning Ability of AlphaZero

Daisy Lin
Center for Neural Science
New York University
New York, NY 10003
xl1005@nyu.edu

Brenden Lake
Center for Data Science and Department of Psychology
New York University
New York, NY 10003
brenden@nyu.edu

Wei Ji Ma
Center for Neural Science and Department of Psychology
New York University
New York, NY 10003
weijima@nyu.edu

Abstract

AlphaZero, a deep reinforcement learning algorithm, has achieved superhuman performance in complex games like Chess and Go. However, its strategic planning ability beyond winning games remains unclear. We investigated this using 4-in-a-row, a game used to study human planning. We analyzed AlphaZero's feature learning and puzzle-solving abilities. Despite strong gameplay, AlphaZero exhibited a 93% failure rate in puzzles. Our feature analysis showed that its self-learned strategies during training lacked certain critical human-like features. We added human-inspired cognitive value function into its policy and value outputs, leading to a 15% improvement in puzzle-solving accuracy. Our findings highlight the potential for human insights to enhance AI's strategic planning beyond self-play.

1 Introduction

AlphaZero has demonstrated remarkable proficiency in mastering complex games such as Chess and Go, achieving superhuman performance through self-play [Silver et al., 2017, 2018]. Despite its success, an important question remains: what exactly does it learn through self-play, and are there limitations in its planning strategy? Planning, defined as the ability to anticipate and simulate future actions and states, is a cornerstone of human intelligence, allowing us to navigate dynamic and uncertain environments [Baker et al., 2017, Gershman et al., 2015]. However, many AI models, including Large Language Models, often struggle to plan when faced with tasks that humans can easily solve. [Valmeekam et al., 2022]. Similarly, traditional reinforcement learning models often overfit to the specific environments in which they are trained, which restricts their capacity to generalize their planning abilities across domains. Understanding the differences between human and AI planning is critical for advancing AI systems that can plan flexibly and adaptively in real-world scenarios. The full scope of these differences remains poorly understood, underscoring the need for deeper investigations into the cognitive mechanisms underpinning AI and human planning [Griffiths et al., 2019, Lake et al., 2017].

To bridge this gap and deepen our understanding of AI planning mechanisms, we turn to the game of 4-in-a-row, a well-established task for studying human planning [van Opheusden et al., 2023]. By training AlphaZero on 4-in-a-row and comparing its performance to that of a validated human cognitive model[van Opheusden et al., 2023], we aim to explore the underlying processes driving AlphaZero's feature learning and strategic planning ability.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

2 Methods

2.1 Task

4-in-a-row is a two-player game where players take turns dropping black and white pieces into a vertical grid. The objective is to connect four of one's own pieces horizontally, vertically, or diagonally before the opponent does. The intermediate complexity of 4-in-a-row, with a state space complexity of 1.2×10^{16}, provides an ideal middle ground between complexity and computational tractability. It is sufficiently difficult for AlphaZero to learn, yet manageable enough to facilitate meaningful comparisons between AlphaZero's planning and human planning.

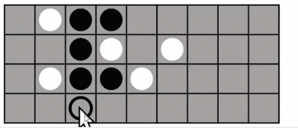


Figure 1: An example of a 4-in-a-row game board. The grid consists of six rows and seven columns.

2.2 AlphaZero Implementation

We implemented AlphaZero using a deep neural network architecture combined with Monte Carlo Tree Search (MCTS) [Silver et al., 2017, 2018]. All agents were trained using high-performance computing (HPC) resources. For each agent, we use an HPC node equipped with 64GB of memory and 8 CPUs for the training process. The models were trained over a period of 5 days.

The agents were trained through self-play, where it generated its own training data by playing 100 games per training iteration. Each training example contains a tuple of (board positions s, MCTS output $\pi(s, a)$, game outcome r). The DNN outputs $(\mathbf{p}, v)$ are trained to match the game result r under a mean-squared error loss and the action probabilities π under a cross-entropy loss, with L_2 weight regularization. The DNN parameters are optimized by the Adam Optimizer, using the training examples from the last 20 iterations, in mini-batches of 64 examples. During each training iteration, the DNN is trained for 10 epochs. The updated network will play 30 games against the current best network. If the updated network can win more games than it loses, it will be accepted and become the new current best network for data generation and network comparison. For details on training and hyperparameter choices, please refer to [Zheng et al., 2022].

2.3 Playing Strength

We established a human performance benchmark to assess AlphaZero's playing strength. We ran a tournament in which the strongest human player we could find played four games against eight selected agents. The top eight agents trained via self-play consistently surpassed the most skilled human player's Elo rating [Elo, 1978], with a mean Elo difference of 90.4 (SD = 17.7). This demonstrates AlphaZero's effectiveness in winning games against human-level opponents. For a comprehensive comparison between agent playing strength and human benchmarks, please refer to [Zheng et al., 2022].

2.4 Puzzles

To evaluate AlphaZero's problem-solving ability, we designed a set of 30 puzzles derived from 4-in-a-row game states. Each puzzle presents a scenario where there is a forced win for the current player within five moves. Solving these puzzles requires constructing sequential threats and anticipating the opponent's responses, thus testing the agent's strategic planning and sequential reasoning.

3 Results

We investigated AlphaZero's planning ability in 4-in-a-row, focusing on two key questions: (1) Can AlphaZero acquire features similar to humans? (2) Can we leverage human insights to improve its performance?

3.0.1 Probing for Human-Used Features

To understand how AlphaZero became proficient at winning games, we employed feature probing techniques similar to concept activation vectors [Kim et al., 2018]. This approach allowed us to detect features used by human players, such as “3-in-a-row” and “2-in-a-row” configurations, identified by van Opheusden et al. [2023]. We trained classifiers using activations from specific layers of the neural network during training to predict the presence of these human-used features.

Our analysis revealed that the network acquired the crucial “3-in-a-row” feature in both the value head and intermediate layers, even without exposure to human-generated data (Figure 2). However, the “2-in-a-row” feature was not prominently represented in the network. This suggests potential limitations in AlphaZero’s ability to learn the full spectrum of strategic features used by humans.

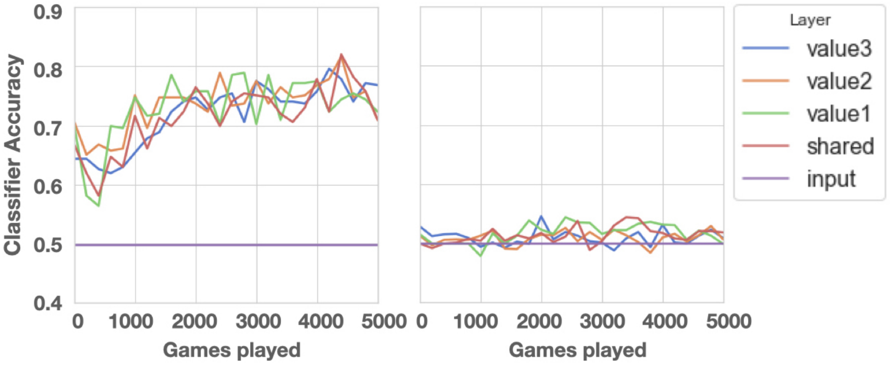


Figure 2: Feature Probing Analysis: Detection of “3-in-a-row” (left) versus “2-in-a-row” (right). Activations from the value head and an intermediate layer show learning of the “3-in-a-row” feature. Control inputs are included for reference.

3.0.2 Unsupervised Feature Representation

To further explore what AlphaZero learned through self-play without predefined concepts, we applied a well-established method, Nonnegative Matrix Factorization (NMF), to extract and visualize latent features from hidden layers [Lee & Seung, 2000, McGrath et al., 2022]. We concatenated activations from 14,907 random game states into a matrix and approximated it as the product of weight and feature matrices, minimizing reconstruction error. The resulting factors provided insights into the network’s understanding of the game by highlighting important activation patterns.

NMF analysis revealed interpretable factors in the network’s intermediate layers, even though AlphaZero was never exposed to human data. (Figure 3). These factors captured diagonal, vertical, and horizontal patterns, suggesting AlphaZero’s ability to represent various game-relevant features that are interpretable to humans.

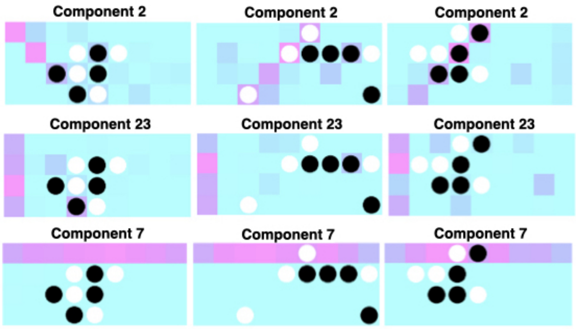


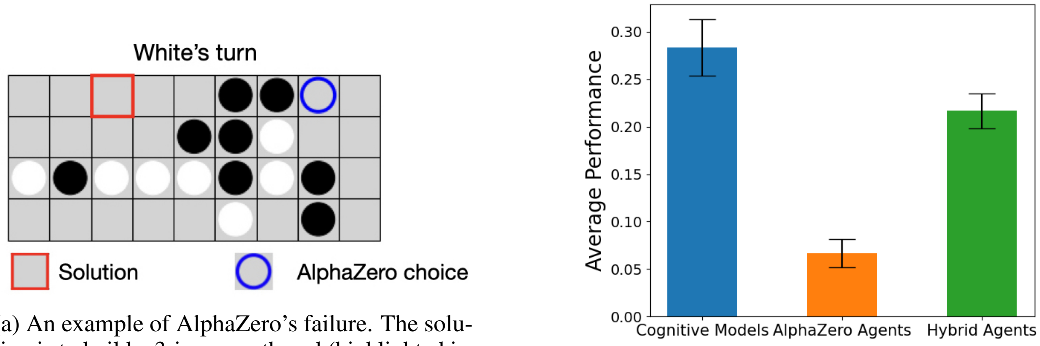
Figure 3: Visualization of NMF for selected factors. Panels show features captured by different residual blocks: diagonals, verticals, and horizontals.

3.1 Puzzle Testing

Despite its strong playing strength, AlphaZero showed a surprising 93% failure rate in finding the best move to solve the puzzles. These puzzles required constructing a logical sequence of moves to force a win within a limited number of turns. In some instances, the agent displayed overly defensive play, neglecting opportunities to build offensive threats (Figure 4a). This observation suggests a gap between AlphaZero's learned strategies and the specific reasoning path used by humans in planning ahead.

3.2 Incorporating Human-Inspired Features

We hypothesized that incorporating a cognitive value function, as described by [van Opheusden et al., 2023], with a linear combination of human-used features, could enhance AlphaZero's puzzle-solving performance. Specifically, we added this cognitive value function output to both the policy and value outputs of AlphaZero network, leveraging features not typically captured in AlphaZero's self-learned heuristics and strategy, such as the "2-in-a-row" and "unconnected-2-in-a-row" configurations. This augmentation of the value and policy predictions led to a 15% improvement in puzzle-solving accuracy (Figure 4b). We observed that Cognitive Models demonstrated the highest performance with an accuracy of 0.28 ± 0.03. In contrast, AlphaZero Agents exhibited substantially lower accuracy, achieving 0.08 ± 0.01. When the cognitive function was incorporated into the Hybrid Agents, their performance improved to 0.21 ± 0.03. This finding highlights the potential of incorporating human cognitive insights to augment AI performance in tasks requiring specific strategic reasoning path, such as solving puzzles optimally.



(a) An example of AlphaZero's failure. The solution is to build a 3-in-a-row thread (highlighted in blue).

(b) Puzzle-solving accuracy comparison.

Figure 4: (a) Unsupervised feature representation and (b) AlphaZero's failure in puzzle solving.

4 Discussion

This study investigated AlphaZero's strategic planning in 4-in-a-row, offering insights into its potential limitations. Our analysis revealed a duality: while AlphaZero successfully learned certain human-interpretable features, such as 3-in-a-row patterns, it struggled to fully capture the breadth of features employed by humans. Despite having very strong heuristics for winning games, AlphaZero's feature learning appears incomplete when compared tothat of humans, as evidenced by the absence of features like 2-in-a-row in its learned representations.

Despite achieving superhuman playing strength, AlphaZero struggled with puzzles requiring a logical sequence of reasoning [Steingrimsson, 2021].These results point to a fundamental limitation in AlphaZero's self-play training regime: it excels at winning games but falls short in tasks requiring strategic, human-like planning. Notably, AlphaZero is not optimized for winning games in the shortest way possible. As long as it secures a victory, the efficiency of the path taken is of little consequence. Therefore, it is perhaps less surprising that AlphaZero's performance declines in scenarios where the shortest, most logical sequence is crucial.

By incorporating human-inspired features into AlphaZero's policy and value estimations, we observed an improvement in puzzle-solving accuracy, demonstrating that human cognitive insights can be leveraged to enhance AI performance in tasks requiring sophisticated strategic reasoning. This improvement underscores the value of blending human intuition with AI learning models, suggesting that hybrid approaches could help address some of the gaps in AI planning capabilities.

However, there are limitations to this approach that warrant further investigation. While introducing human-inspired value function led to measurable gains in performance, it raises questions about the generalizability of these improvements. AlphaZero's enhanced performance may be specific to the puzzles tested, and those puzzles represent a narrow set of highly structured scenarios.

5 Conclusion

Our work demonstrates that while AlphaZero excels in gameplay through self-play, it may not fully acquire the nuanced of strategic planning used by humans. Incorporating human-inspired value function can bridge this gap, improving the agent's ability to solve complex puzzles requiring logical sequence reasoning. This approach underscores the value of combining human cognitive insights with AI learning methods to enhance performance in tasks that mirror human strategic thinking.

References

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80:2668–2677.

Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., Kim, B., Paquet, U., & Kramnik, V. (2022). Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences of the United States of America*, 119(47), e2206625119.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.

Steingrimsson, H. (2021). Chess fortresses, a causal test for state of the art symbolic [neuro] architectures. *2021 IEEE Conference on Games (CoG)*, 1–8.

Valmeekam, K., Narayan-Chen, A., Cheung, B., Reddy, S., & Srivastava, S. (2022). Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, 618(7967), 1000–1005.

Baker, C., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 0064.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253. `doi:10.1017/S0140525X16001837` PMID: 27881212.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, 29, 24–30. ISSN 2352-1546. `doi:10.1016/j.cobeha.2019.01.005`

Zheng, Z., Lin, X., Topping, J., & Ma, W. (2022). Comparing Machine and Human Learning in a Planning Task of Intermediate Complexity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44.

Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., Saldyt, L., & Murthy, A. (2024). LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. *arXiv preprint*.