Comparing Machine and Human Learning in a Planning Task of Intermediate Complexity

Zheyang (Sam) Zheng* (zz737@nyu.edu)

Xinlei (Daisy) Lin* (xl1005@nyu.edu)

Jake Topping* (email@jaketopping.co.uk)

Wei Ji Ma (weijima@nyu.edu)

* These authors contributed equally to this work

Abstract

Deep reinforcement learning agents such as AlphaZero have achieved superhuman strength in complex combinatorial games. By contrast, the cognitive science of planning has mostly focused on simple tasks for experimental and computational tractability. Using a board game that strikes a balance between complexity and tractability, we find that AlphaZero agents improve in value function quality and planning depth through learning, similar to human in previous modeling work. In addition, these metrics reflect causal contributions to AlphaZero's playing strength. Yet the strongest contributor is the policy quality. The decrease in policy entropy also drives the increase in planning depth. The contribution of planning depth to performance is lessened in late training. These results contribute to a joint understanding of machine and human planning, providing an interpretable way of understanding the learning and strength of AlphaZero, while generating novel hypothesis on human planning.

Keywords: Human-DNN comparison; Planning; Learning; Interpretable Machine Learning;

Introduction

There has long been a positive mutual influence between research on artificial intelligence (AI) and research on human intelligence (Turing, 2009; Lake, Ullman, Tenenbaum, & Gershman, 2017). The recent success of deep learning has inspired a plethora of work comparing deep neural networks (DNN) and biological neural networks, on both neuronal and behavioral levels, using such comparison to further our understanding of both sides. For instance, DNNs have been shown to be good models of neuronal activities in the human ventral visual cortex (Yamins & DiCarlo, 2016). Shape bias in object categorization similar to that of human children has been found in a DNN model of one-shot learning (Ritter, Barrett, Santoro, & Botvinick, 2017).

Despite the fruitful comparisons in the field of vision, they have been less common in the field of planning, defined as a cognitive process in which the decision-maker mentally simulates future states, actions or outcomes in a decision tree. One obstacle comes from the lack of alignment of the tasks: artificial intelligence research has focused on solving complex tasks like Chess (Silver et al., 2018) and Go (Silver et al., 2017), while cognitive science and psychology have favored detailed modeling using simple tasks like the two-step task in (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). When comparisons do happen, the tasks used still lie on either extreme. Wang et al. (2018) proposed a meta-reinforcement learning (RL) model performed by a recurrent neural network that explained well diverse findings regarding the role of dopamine and pre-frontal cortex in reward-based learning. They compared human behavior and neural activities with the model on the two-step task, the simplest task that allows for planning (Daw et al., 2011). The model generated reward prediction error related activities similar to those in the human ventral striatum, yet the behavior of the model is fully model-based, in contrast to human's mixed strategy. On the other hand, in addition to using human players as baselines, AI research on Chess and Go often examines moves made under specific situations by the AIs to see if and when they have learned human concepts about the game (Tian et al., 2019; McGrath et al., 2021), or have developed non-standard strategies beyond the scope of traditional human knowledge (Silver et al., 2017; Dou, Ma, Nguyen, & Nguyen, 2020).

We aim to complement the above line of comparative work by using a task that is challenging enough for both humans and artificial agents while being computationally tractable. Therefore, we use 4-in-a-row, a variant of tic-tac-toe in which two players alternate placing pieces on a 4-by-9 board, aiming to get four pieces in a row horizontally, vertically, or diagonally. van Opheusden et al. (2021) developed a computational model for human planning in 4-in-a-row. They showed that the playing strength of human subjects increased with experience. More specifically, the quality of their value functions and planning depth increased.

In the present study, we compare how AI and human learn 4-in-a-row. We adapt our agents from the AlphaZero family, the state-of-the-art deep RL algorithm that learns to play board games at superhuman levels, which includes AlphaGo Zero (Silver et al., 2017), AlphaZero (Silver et al., 2018), and MuZero (Schrittwieser et al., 2020). For simplicity, we will refer to our agent as AlphaZero. We characterize AlphaZero's performance using three derived metrics: planning depth, value function quality and policy quality. Note that we are not using AlphaZero as a model for human. The goal is simply to dissect AlphaZero learning and playing strength using the similar planning metrics as in the human studies (with one additional metric for AlphaZero). We then com-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

pare how these metrics change during learning in AlphaZero and human and study how they contribute to AlphaZero performance.

Similar to human, the agent improves in all the metrics with training. Using a combination of observation and causal manipulations, we show that planning metric improvements are not epiphenomena of training, but mediate the increase in playing strength. We also find phenomena that have not been reported in human study. Policy quality contributes the most to performance. The increase in planning depth is mediated through a decrease in the entropy of the policy, reflecting a more concentrated search process. The contribution of planning depth to playing strength diminishes at later stages of learning. These discrepancies might either reflect mechanistic differences between machine and human planning, point to phenomena only observable in the more extreme end of the spectrum in human expertise, or highlight alternative hypotheses about human planning. Our result not only disentangles contributions to the playing strength and learning of AlphaZero, but also inspires future research directions in human planning.

Results

We trained AlphaZero type deep RL agents to play 4-in-a-row. AlphaZero uses a DNN to guide its Monte-Carlo tree search (MCTS). Given the input (a board and optionally the player's color), the DNN returns an immediate value v of a board, indicating how likely the current player will win/lose from this board. The DNN also returns a policy $\mathbf{p}$ (or $p(a|s)$ for a board s and move a), a prior probability of selecting an action before doing any tree search. MCTS simulates future actions and states in a way that balances exploration and exploitation. It then integrates those results to inform current decision. The training involves using self-play to generate training data. Stochastic gradient descent (SGD) based optimization is used to train the DNN to predict past game results and its own post-MCTS action probabilities. After ten training epochs, the updated model plays against the current best agent to determine its acceptance/rejection. The accepted model will become the new data generator. If the updated model is rejected, it is either kept to be trained again (same as in AlphaGo Zero), or reverted back to the previous best model. One training iteration consists of a whole cycle of data generation, DNN parameter update, and evaluative games.

Using thirteen sets of hyperparameters, we produce thirteen Networks. We call models sharing the same hyperparameter set and initialization during training a "Network", to distinguish this concept from an individual iteration saved during training, which we call an "agent" or a "model". Thirteen networks (497 agents) show diverse playing strength and planning metrics (Figure 1). The choice of the hyperparameters is not systematic. The goal is simply to add variations in playing strength and planning metrics. A round-robin tournament is held among all the models. The results are used to derive Elo ratings, a standard way to measure play strength in board games (Coulom, 2008).

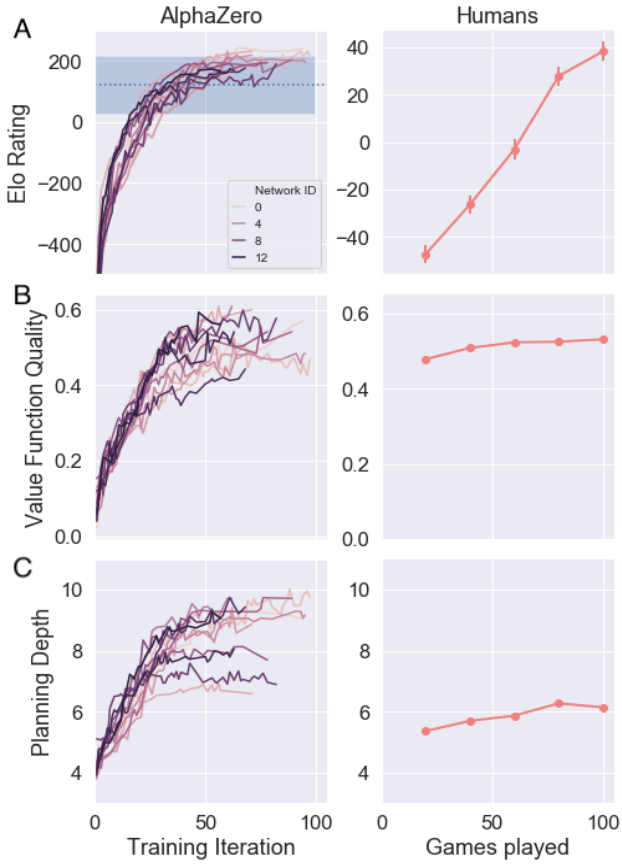

Figure 1: **Elo and planning metric comparison between AlphaZero and human**. Playing strength (Elo rating, A), value function quality (B) and planning depth (C) of both AlphaZero and human increase with training. Solid lines represent Networks. Dotted line in (A) represents the Elo of a strong human player, and the shade reflects the 95-confidence interval of this Elo estimate. Human results are reproduced from data in van Opheusden et al. (2021). The scale of Elo ratings are different between AlphaZero (left) and human (right) and the numbers are not directly comparable because there is no tournament between human players from prior study with our AlphaZero agents.

Playing strength increases

AlphaZero's playing strength increases over training across all Networks (Figure 1A; left). To obtain a human benchmark, we have the strongest human player we could find play 4 games each against 8 selected agents, with Elo ratings ranging from 140 to 242 (the best). Agents at middle training iterations already start to surpass the human bench mark, with later agents lying above the 95-confidence interval. Human learning curve from the previous study is recreated here (1A; right). Our question is what aspects of the agents' capacity have improved to enable such an improvement in playing strength.

Evidence for smarter trees

Prior human modeling studies in 4-in-a-row used planning metrics to explain human learning and playing strength. Value function quality measures how closely people's heuristic evaluation of a board aligns with the game-theoretic value of a board (see Methods). Planning depth reflects how many steps into the future can one look ahead. Feature dropping rate reflects how often people ignore features on the board (van Opheusden et al., 2021). The study showed a learning effect on value function quality when the initial quality is not too high, on planning depth, and on feature dropping rate. Since AlphaZero agents don't have human-like attentional lapses, nor are their values directly computed from on explicit features (like controlling a 3-in-a-row on the board), feature dropping rate is not included in our comparisons.

For each AlphaZero agent, we compute the value function quality by calculating the Pearson correlation between the DNN-returned immediate values of the probe boards and their game-theoretic values obtained in previous work (see Methods). We measure planning depth by having each agent make a move at probe boards, and average across the probes the length of the deepest branch of each resulting MCTS tree. Both value function quality and planning depth increase as training progresses (Figure 1B and C; left). We plot previous human results here to aid comparison (Figure 1B and C; right) (van Opheusden et al., 2021). Compared to human learning, the increase in planning metrics are more drastic in AlphaZero, which is expected given that human is not a blank slate to begin with.

The increase of planning depth in human learning has been attributed only to an increase in the number of search iterations (van Opheusden et al., 2021). By contrast, we discover that planning depth of AlphaZero increases over training despite the number of MCTS searches, N_{MCTS}, being fixed. This suggests either a potential difference between humans and machines in how their decision trees change during learning, or a potential alternative hypothesis for the mechanism of the depth increase in humans. Here we only provide a mechanism of depth increase in AlphaZero.

Entropy of action prior mediates the increase in planning depth

When the total search budget is fixed, one possible mechanism for the increase in planning depth could be a more targeted and less scattered search process. In AlphaZero the targetedness of the search is largely modulated by the policy. The policy starts out uniform and evolves to match the post-MCTS action probabilities. Since the search process makes the action probabilities be less uniform, the policy should become less uniform and thus have a lower entropy over training, defined as $H(s) = -\sum_a p(a|s) \log p(a|s)$. A decrease in entropy over training is confirmed in Figure 2B. (Similar to planning depth, the entropy here is also averaged across probe boards.)

Policy quality improves

A more concentrated prior does not necessarily imply "smarter" searches. A bad prior can lead a deep but misguided search. We therefore develop a metric, policy quality, to quantify how good AlphaZero's policies are. Policy quality reflects the correlation between AlphaZero's policies and the optimal policies, derived from the game-theoretic values (see Methods). The policy quality improves over training for all Networks (Figure 2A. So not only are the priors more concentrated, but they also align better with optimal policies, and thus lead the search in more promising directions.

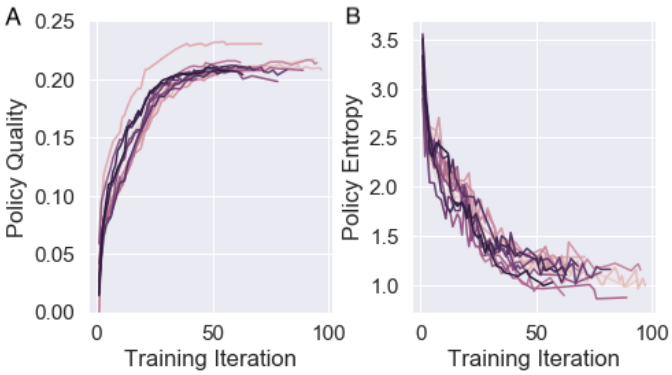

Figure 2: **Policy quality and policy entropy.** A) Policy quality of AlphaZero agents increases with training. B) Policy entropy of AlphaZero agents decreases with training

The increase in playing strength is mediated by planning metrics

Playing strength of AlphaZero agents increases with training (Fig 1A; left), and we hypothesize that the effect of training on playing strength is mediated by planning metrics. Mediation analysis shows that policy quality, value function quality and planning depth all have a significant mediated effect on Elo ratings (Figure 3). (We test the significance using bootstrapping procedures.)

Policy quality matters the most

Mediation analysis on each planning metric shows that learning-induced Elo changes are mediated by planning metrics. However we cannot reliably conclude the relative contribution of each metric, since the metrics are correlated with each other (value-policy:0.94, value-depth:0.80, depth-policy:0.85). We first demonstrate the dominant contribution of policy quality to performance through observational data and then in the later sections use causal manipulations to dissect the role of value function quality and planning depth.

Policy quality, planning depth and value function quality together explain 0.95 of the variance in Elo in a linear regression, with weights: $\beta_{policy} = 0.92$ ($p < 10^{-20}$), whereas $\beta_{depth} = 0.09$ ($p < 10^{-5}$) and $\beta_{value} = -0.03$ ($p = 0.334$). We also test the dominance of policy quality by first regressing

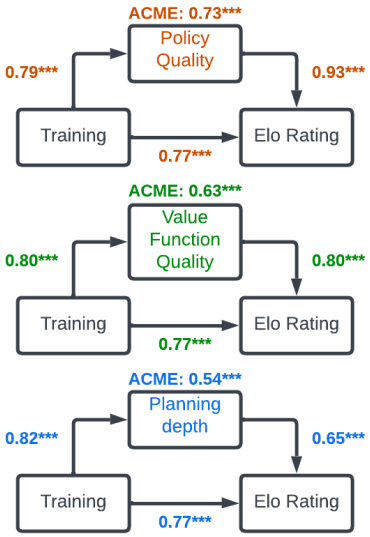

Figure 3: **Mediation analysis: illustration and results.** The effect of training on Elo ratings is mediated via three planning metrics: policy quality, value quality and planning depth. ACME is the average causal mediation effect. Numbers next to arrows represent the regression coefficients between variables. Asterisks denote statistical significance.

out all the other confounders (training iterations, value quality and depth) from Elo. Policy quality explains the residuals significantly well ($F = 29.11$, $p < 10^{-6}$). By contrast, neither a similar residual regression for value function quality ($F = 0.18$, $p = 0.668$) nor one for depth ($F = 2.69$,$p = 0.101$) is significant.

Causal manipulations reveal contributions from all planning metrics

The planning depth and value function quality do not show significant contribution to Elo once all the confounding factors are regressed out. But this fact by itself does not rule out the possibility of their contribution. To arbitrate the role of planning depth in playing strength, we causally manipulate planning depth for each iteration in the best Network, while holding everything else about an agent constant. To do this, we replicate an agent and then set its number of MCTS searches (N_{MCTS}) to four different levels. The resulting agents are then included in the tournament. For the same iteration (dots connected by the same line in Figure 4A.), a higher N_{MCTS} induces a high planning depth, which correlates positively with Elo in all iterations. As training progresses, the positive effect of planning depth on Elo diminishes, as seen from the decreasing of the slopes of the lines in Figure 4A. We perform a linear regression (Elo $\sim$ depth) for each iteration within the Network to obtain its "depth efficiency" (Figure 4B). The depth efficiency decreases as training progresses. One possible explanation for why the benefit of depth diminishes is that the good action priors of well-trained models are sufficient to guide actions. Adding more depth in the search might not advise major changes to the preferences provided by the prior.

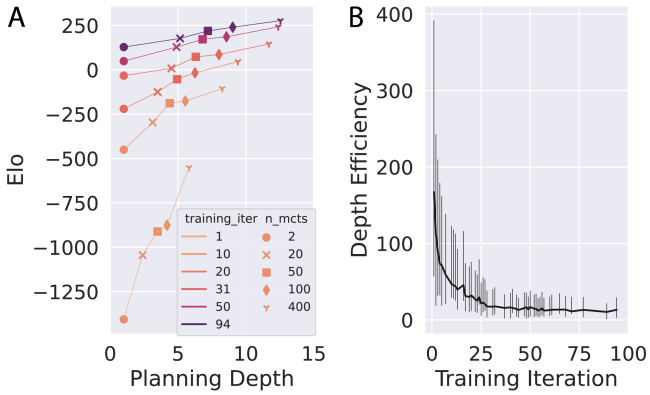

Figure 4: **The effect of N_{MCTS} manipulation on depth and Elo.** A) Elo vs planning depth for selected iterations and all N_{MCTS} manipulation of the chosen Network. Color indicates training iteration, and marker style indicates the number of MCTS searches. Agents with the same training iteration are connected by a line. B) Depth efficiency vs Training iteration for all agents. Depth efficiency is defined as the slope of each line in A (as well as the lines for other iterations not shown in A), which represents the efficiency of depth increase in increasing Elo. Error bars reflects the 95% confidence intervals.

For the value and policy manipulation, we select eleven models from a Network that spans early, middle and late epochs of training (this Network has the highest policy quality and is different from the one in the N_{MCTS} manipulation). We swap either the value or the policy function of a model with the value/policy function of low, middle or the best quality. These swap targets come from the initial, a middle (iter 22) or the final iteration of the model within the Network, respectively. Models from one training epoch do not have swaps with models from the same training epoch (e.g. the policy/value functions of models from early iterations (iter 1-20) will only be replaced by those from the middle and final models).

The result shows that both value and policy contribute to performance, as equipping early models with a well-trained policy or value function improves their Elo (Figure 5). The gain is larger with the policy swap initially, but the value swap catches up during the middle epoch (20-35 iters), suggesting value and policy quality can complement each other in this intermediate range, i.e. a good performance does not require both value and policy to be really high, but only one to be high and the other intermediate. Swapping the policy function of a well-trained model to a naive policy is unambiguously more disruptive than swapping the value function, again echoing the previous section in terms of the overall dominance of policy. Swapping the components of the late models to those of the middle ones produces qualitatively similar but

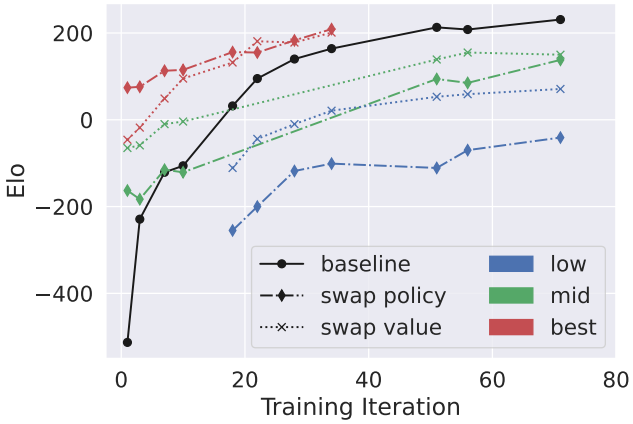

Figure 5: **Elo ratings as a result of the value or policy function manipulation.** Black markers (solid line) are the original agents across training. Colored diamond markers (half solid line) are agents with policy functions swapped. Colored cross markers (dotted lines) are agents with value functions swapped. Color reflects the quality of the target of the swap: low-blue, middle-green, best-red.

quantitatively less drastic reduction, compared to swapping to early ones. Surprisingly, replacing the value functions of the early models with those of the middle ones provides a larger improvement than swapping the policy functions. This phenomenon awaits further investigations.

Discussion

We analyzed what capacity changes underlie AlphaZero's learning in a game of intermediate complexity. We found that although value function quality, planning depth and policy quality all improve during training, the improvement in performance is driven the most by policy quality. The intermediate complexity of 4-in-a-row allowed us to compute the game-theoretical values of board positions and therefore the value function quality and policy quality metrics, which are difficult to obtain in complex games such as Go.

The distinction between value and policy is worth emphasizing. Value provides a context independent common scale for measuring all possible states (immediate value), and can be updated during the planning/reinforcement learning process (action value). Policy provides a context dependent relative scale for comparing available actions at one state. In MCTS it biases the search and does not change during the search. The machine learning community has noted the distinct and crucial role of policy. Hessel et al. (2021) argued that action values alone are not enough for representing the best stochastic policy. Hamrick et al. (2020) studied the contribution of planning in MuZero and showed that planning is most useful in the learning process, while post-learning, shallow trees are often as performant. The AlphaGo Zero study also showed strong play from a myopic policy network without MCTS (Silver et al., 2017). These results are consistent with our result of manipulating the N_{MCTS}, all of which point to the effect of a good policy in maintaining performance in the absence of many searches. The appropriate complexity of the task allows us to further quantify policy and value quality and directly demonstrate the dominant role of policy among the different components of AlphaZero.

The previous cognitive model on 4-in-a-row assumed an objective immediate value function based on counts of desirable (e.g. self's 3-in-a-row) and undesirable features (e.g. opponent's 3-in-a-row), and updated action values using best-first search (see Method). It did not implement a policy. On the other hand, in several human decision making studies, direct policy learning methods better explained subjects' choices in complete feedback tasks, falsifying the assumption of learning option-values on an objective scale (Klein, Ullsperger, & Jocham, 2017; Li & Daw, 2011). Others, however, showed direct policy learning by itself was not sufficient and favored a hybrid scenario where values were computed, but in a context-dependent way (Palminteri, Khamassi, Joffily, & Coricelli, 2015; Bavard, Lebreton, Khamassi, Coricelli, & Palminteri, 2018). Since these observations have been made using simple reinforcement learning tasks where the goal was to maximize stochastic reward, it would be interesting to experimentally test whether policy and value can be disentangled in more complex scenarios like 4-in-a-row, especially given that the game is strategic and deterministic.

Our agents improve planning depth through a more concentrated policy induced by training. By contrast, in humans, a depth increase in 4-in-a-row seemed to be due to an increase in the number of searches (Van Opheusden, Galbiati, Bnaya, Li, & Ma, 2017). What is the origin of this difference? It would be possible that the depth increase in humans could at least be partly explained by "smarter" searches, but such an explanation was not within the degree of freedoms of the cognitive model. In the cognitive model for the human (briefly described in Method), after the current player selects the best leaf node and simulates an opponent move, that move would often block the current player's feature, turning the best node into the worst. Only after a thorough search through the alternative actions, can the initial best node regain its status. Only then can the search advance one level deeper. Additionally, although the human study did rule out MCTS as a good model for explaining human behavior, the MCTS they examined differed in crucial ways. The immediate value of a board was obtained from a rollout instead of a value function as in AlphaZero. Nor did they incorporate a policy in selecting which node to explore. In this light, perhaps aspects of MCTS, such as using a policy to guide tree search, could possibly help explain human behavior, but were overshadowed by aspects that were not realistic in previous model comparisons. In any case, it would be an interesting future direction in human planning research to arbitrate between "more searches" and "smarter searches" as mechanisms for depth increase.

The diminishing contributions of the planning depth on Elo as training progresses has not been reported in human 4-in-a-row studies. One possibility is that the human subjects recruited were not on the extreme end of expertise such that deeper searches stop providing marginal utilities. Intriguingly, the early work on chess by de Groot (1965) was unable to find gross differences in the number of moves considered, search heuristics, and depth of search, between masters and weaker players. But masters are good at coming up with the "right" moves to search further. This finding allows us to draw similarities between the masters and AlphaZero agents at a later learning stage. It also provides some confidence that AlphaZero could be valuable in showing a wider range of possible behaviors than those shown in the recruited human subjects. More importantly, the chess result indicates that something similar to the policy in AlphaZero might indeed play a crucial role in human decision making.

Method

AlphaZero

Our neural network architecture and MCTS are largely the same as described in the AlphaGo Zero work (Silver et al., 2017). We used 3 or 9 residual blocks in the DNN.

Training For each training iteration, we use the current best agent to play 100 games against itself to generate the training examples. During the first 15 steps of each self-play game, the temperature is set to 1 to induce variability in the data. In the AlphaZero paper, the temperature is later set to 0. Here we include one Network whose temperature does not switch. For some Networks, during self-play, Dirichlet noise with a hyperparameter Dir_α is added to the current root of the MCTS tree to encourage exploration. We also include Networks without Dirichlet noise, which deviates from AlphaZero. The motivation for different hyperparameters is exploratory and not systematic.

Each training example contains a tuple of (board positions s, MCTS output $\pi(a|s)$, game outcome r). The DNN value output v is trained to match the game result r under a mean-squared error loss and the policy output $\mathbf{p}$ is trained to match the action probabilities π under a cross-entropy loss, with $L2$ weight regularization. The DNN parameters are optimized by the Adam Optimizer, using the training examples from the last 20 training 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. If the updated network can win more games than it loses, it will be accepted and become the current best network for data generation and network comparison. We use this looser selection criterion compared to AlphaGo Zero to encourage easier network update. Because if the updated network is not accepted, we revert it back to the current best network, forgoing the parameter update (different from AlphaGo Zero). But if the "continuous training" hyperparameter is true, the updated network continues training in the next iteration, similar to AlphaGo Zero.

Measuring playing strength

We hold a tournament in which each agent plays against every other agent once as both colors. There are 789 agents in total, including all accepted iterations from the thirteen Networks, as well as the agents whose N_{MCTS} have been modified, and those whose value or quality functions have been swapped. The temperature is fixed at 0.1. Playing strength is quantified by Elo ratings, computed by the BayesElo program (Coulom, 2008). The Elo ratings are computed such that the difference between the Elo ratings of two players maps monotonically to the probability that one player will defeat the other.

Probe boards and game-theoretic values

The probe boards are all positions (5482 positions) which occurred in human-vs-human experiments conducted by van Opheusden et al. (2021). The game-theoretic values of these boards are defined as game outcomes in which both sides play perfectly. van Opheusden et al. (2021) approximated the game-theoretic values by searching each board for 200,000 iterations using the cognitive model. The result for most boards converges to a game-theoretic value, while the undetermined ones are assigned a 0 value, indicating a draw. We used these pre-computed game-theoretic values for our value quality calculation.

The cognitive model The cognitive model is a best-first search algorithm combined with a feature-based value function ((van Opheusden et al., 2021)). It contains other components like feature dropping rate, which is not used for the game-theoretic value calculation, and thus will not be described in this section. The value function evaluates a board by a weighted sum of feature counts, including desirable features like 3-in-a-row one owns and undesirable features like 3-in-a-row the opponent owns. For each iteration of the best-first search, the algorithm performs mini-max when traversing down a search tree, expands the tree at the leaf node by evaluating all the children nodes of the leaf, and updates the traversed nodes with the best mini-max value among the newly expanded children.

During the process of writing the paper, we became aware of the work of Uiterwijk (2019), who developed a solver for the 4-by-9 4-in-a-row game. It would be more accurate to derive the exact game-theoretic values using the solver, but the current method suffices for a good approximation when a large number of boards are included.

Policy quality

After computing the game-theoretic values for each child board of all probe boards (used in value quality and depth calculation), we applied softmax to the values of those children boards to get an "optimal" policy for each probe board. We then concatenate these optimal policies of all probe boards, and correlate the long vector with the concatenated policy vector returned by a DNN.

Acknowledgement

This work was supported by grant number R01MH118925 from the National Institutes of Health.

References

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature communications*, *9*(1), 1–12.

Coulom, R. (2008). Whole-history rating: A bayesian rating system for players of time-varying strength. In *International conference on computers and games* (pp. 113–124).

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague, Mouton.

Dou, Z.-L., Ma, L., Nguyen, K., & Nguyen, K. X. (2020). Paradox of alphazero: Strategic vs. optimal plays. In *2020 ieee 39th international performance computing and communications conference (ipccc)* (pp. 1–9).

Hamrick, J. B., Friesen, A. L., Behbahani, F., Guez, A., Viola, F., Witherspoon, S., ... Weber, T. (2020). On the role of planning in model-based deep reinforcement learning. *arXiv preprint arXiv:2011.04021*.

Hessel, M., Danihelka, I., Viola, F., Guez, A., Schmitt, S., Sifre, L., ... Van Hasselt, H. (2021). Muesli: Combining improvements in policy optimization. In *International conference on machine learning* (pp. 4214–4226).

Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature communications*, *8*(1), 1–12.

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*.

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, *31*(14), 5504–5511.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Hassabis, D., Kim, B., ... Kramnik, V. (2021). Acquisition of chess knowledge in alphazero. *arXiv preprint arXiv:2111.09259*.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature communications*, *6*(1), 1–14.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning* (pp. 2940–2949).

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... others (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359.

Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., & Zitnick, L. (2019). Elf opengo: An analysis and open reimplementation of alphazero. In *International conference on machine learning* (pp. 6244–6253).

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

Uiterwijk, J. W. (2019). Solving strong and weak 4-in-a-row. In *2019 ieee conference on games (cog)* (p. 1-8). doi: 10.1109/CIG.2019.8848010

Van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2017). A computational model for decision tree search. In *Cogsci.*

van Opheusden, B., Galbiati, G., Kuperwajs, I., Bnaya, Z., Ma, W.-J., et al. (2021). Revealing the impact of expertise on human planning with a two-player board game.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860–868.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.