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

Daisy Lin*, Sam Zheng*, Jake Topping*, Wei Ji Ma

Presenter: Daisy Lin

Ma Lab

CNS Lab Talk , 03/04/2022

OUTLINE

- **Intro to AlphaZero and cognitive model**
- **AlphaZero learning vs. Human learning**
- **Summary and future directions**

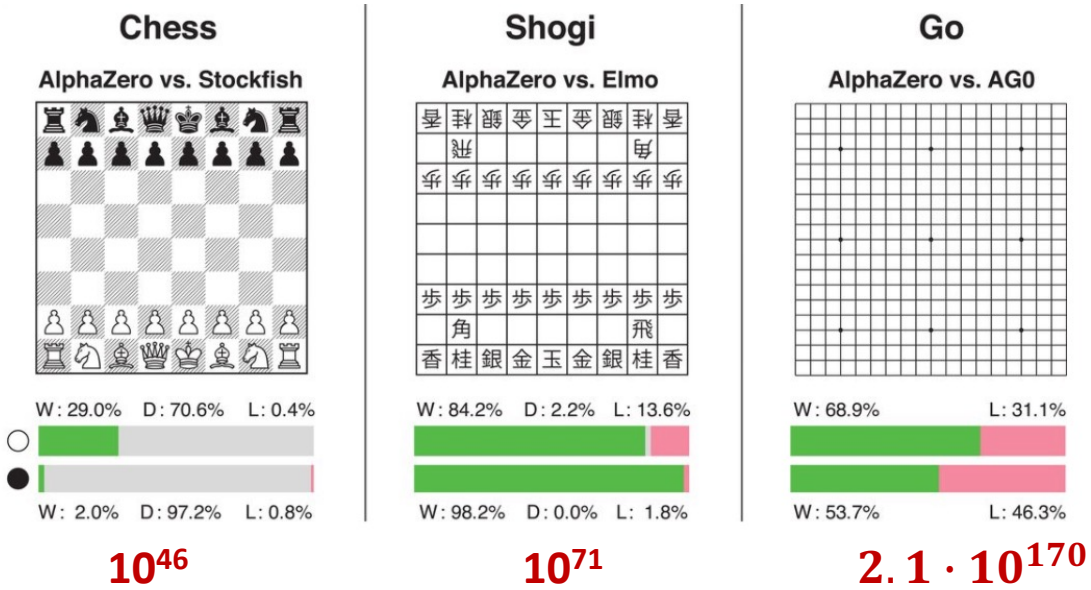
Planning in AI

Planning in Cognitive Science

Deepmind successes

AlphaZero

- superhuman performance

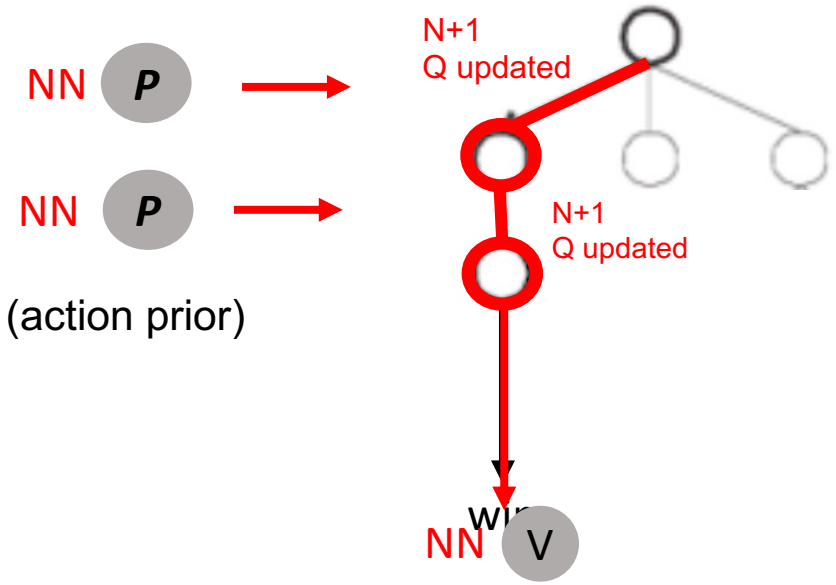


10^{46} 10^{71} $2.1 \cdot 10^{170}$

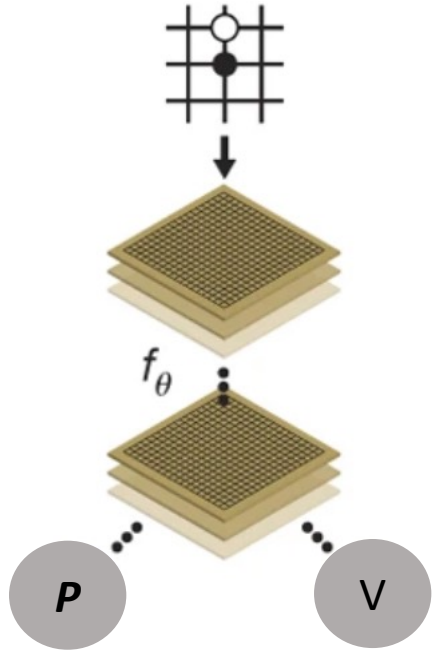
Silver et al., *Science*, 2017

AlphaZero: Neural network driven MCTS

Monte Carlo Tree Search:



Neural Network



Probability distribution over moves

Probability of winning in current position

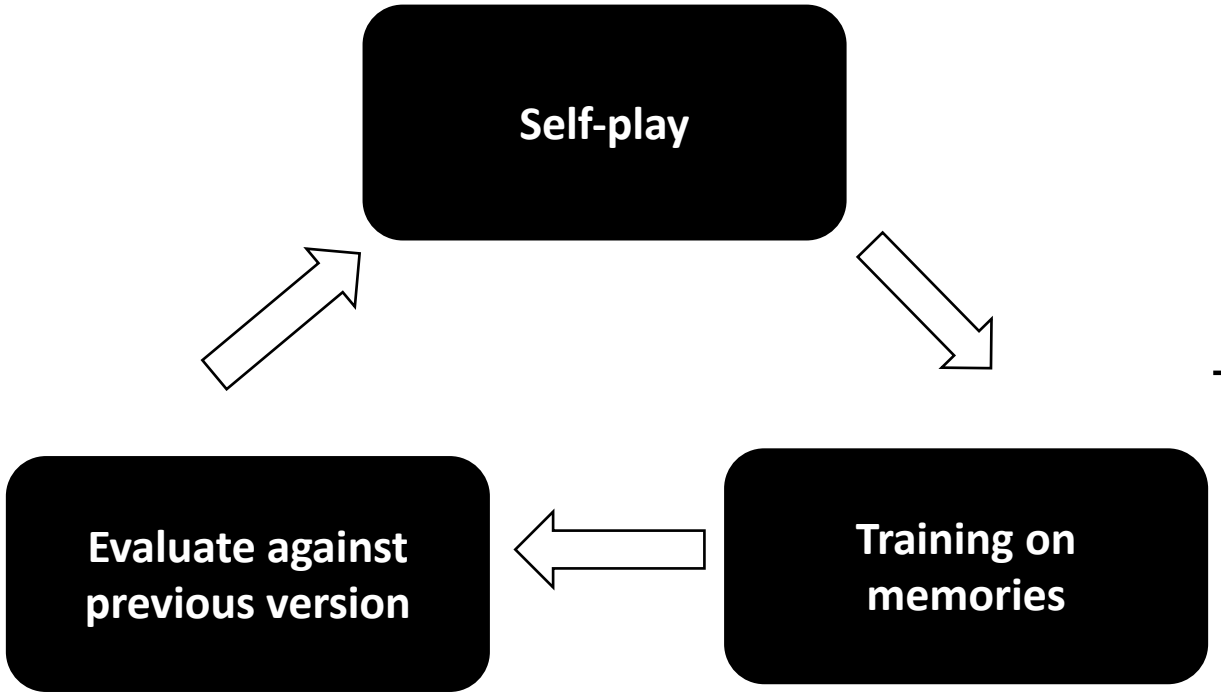
- Selection during search favors moves with a high prior, low visit count (to encourage exploration), and high mean action value (Q)

(Silver et al., 2017)

Learning cycle

At each move, store:

- board position
- MCTS outputs
- Game outcome



Train NN such that:

- Policy head predicts MCTS output
- Value head predicts game result

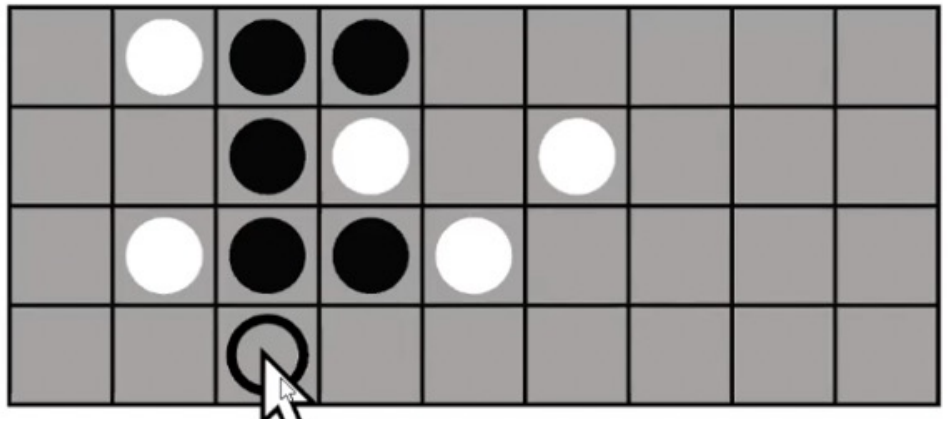
Newly trained agent becomes “new best” if it wins 50% playing 30 games with the current best version

Planning in AI

Complex tasks: chess, go
AlphaZero trained on the task

Planning in Cognitive Science

4-in-a-row

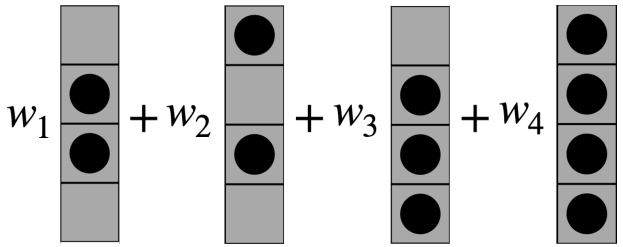


- 2 player game, on a 4-by-9 board
- The goal is to connect four pieces
- State space complexity: $1.2 \cdot 10^{16}$

4-in-a-row

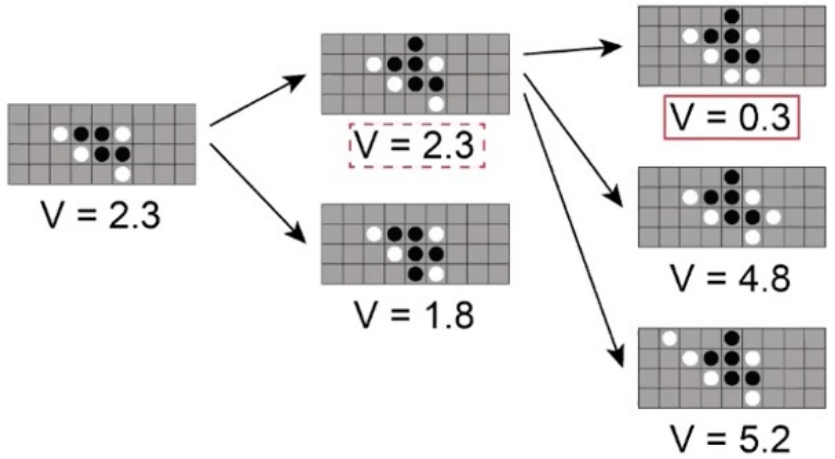
- Cognitive Model that can reliably predict human moves

value function



Replaced by in NN in AlphaZero

Best First Search



(Van Opheusden et al. ,2021)

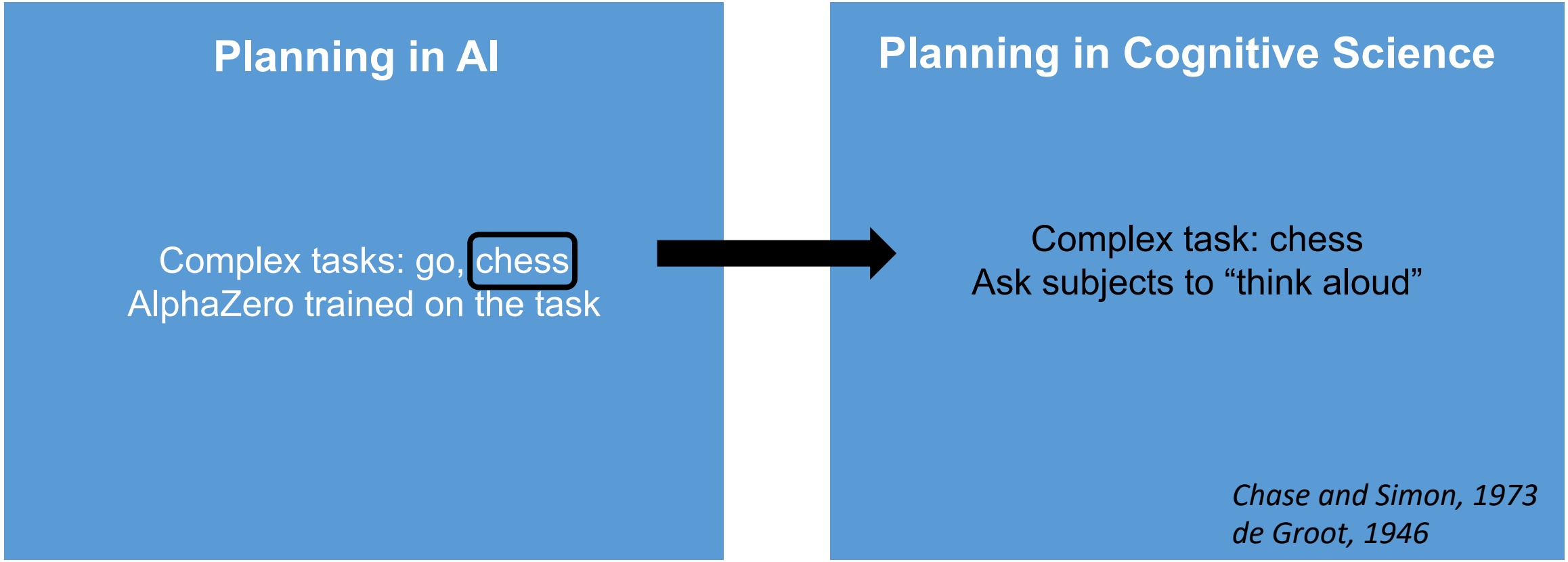
Planning in AI

Complex tasks: chess, go
AlphaZero trained on the task

Planning in Cognitive Science

Intermediate complexity task: 4-in-a-row
Cognitive model fitted to human play

Attempts to align machine and human planning

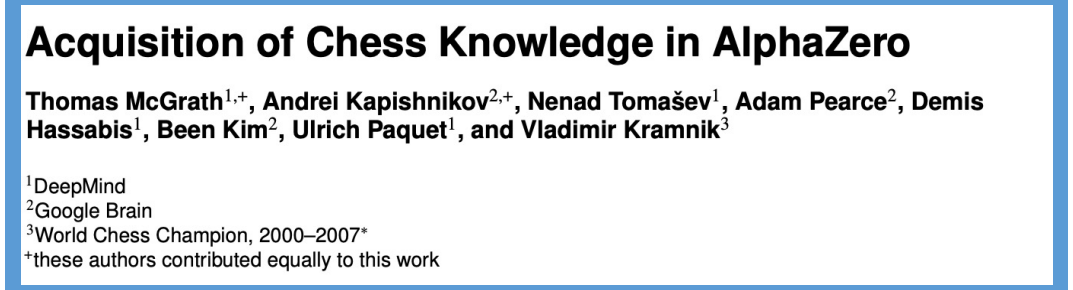


It’s difficult to build precise models of human behavior in complex games

Attempts to align machine and human planning

Planning in AI

Complex tasks: go, chess
AlphaZero trained on the task



Acquisition of Chess Knowledge in AlphaZero

Thomas McGrath[1,+], Andrei Kapishnikov[2,+], Nenad Tomašev[1], Adam Pearce[2], Demis Hassabis[1], Been Kim[2], Ulrich Paquet[1], and Vladimir Kramnik[3]

[1]DeepMind
[2]Google Brain
[3]World Chess Champion, 2000–2007*
[+]these authors contributed equally to this work

Planning in Cognitive Science

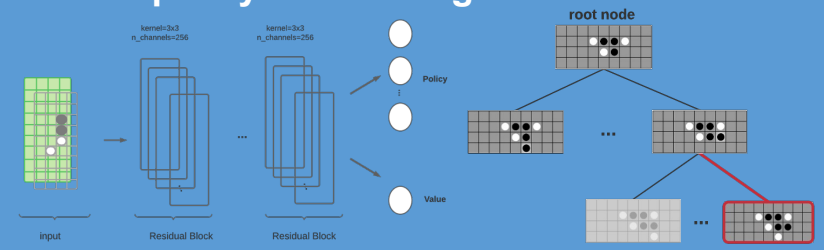
Intermediate complexity task: 4-in-a-row
Cognitive model fitted to human

- AlphaZero can acquire human concepts in chess

Planning in AI

Intermediate complexity task: 4-in-a-row
AlphaZero trained on the task

NN policy and value guided MCTS

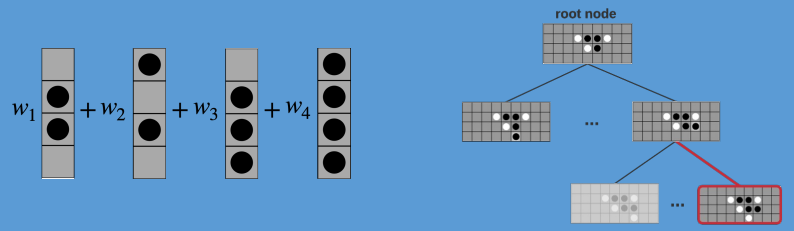


Metrics: value function quality, planning depth

Planning in Cognitive Science

Intermediate complexity task: 4-in-a-row
Cognitive model fitted to human

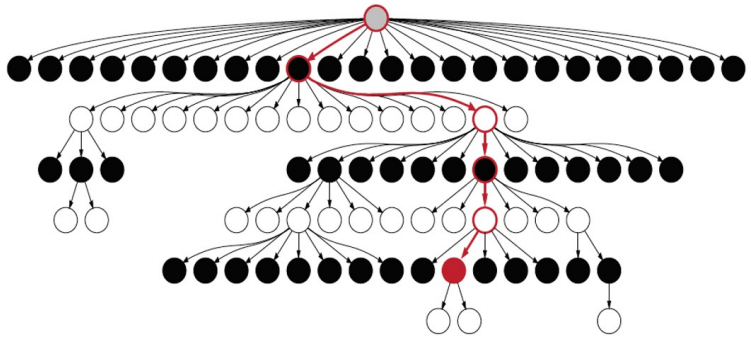
Feature-based value function guided BFS



Metrics: value function quality, planning depth

Planning Metrics

- **Planning depth**: how far one looks into the future



- **Value function quality**: how good the value estimate of a board is
    - Pearson correlation between estimated value and objective value of a board

Does AlphaZero learn to play 4-in-a-row in a similar way compared to humans?

OUTLINE

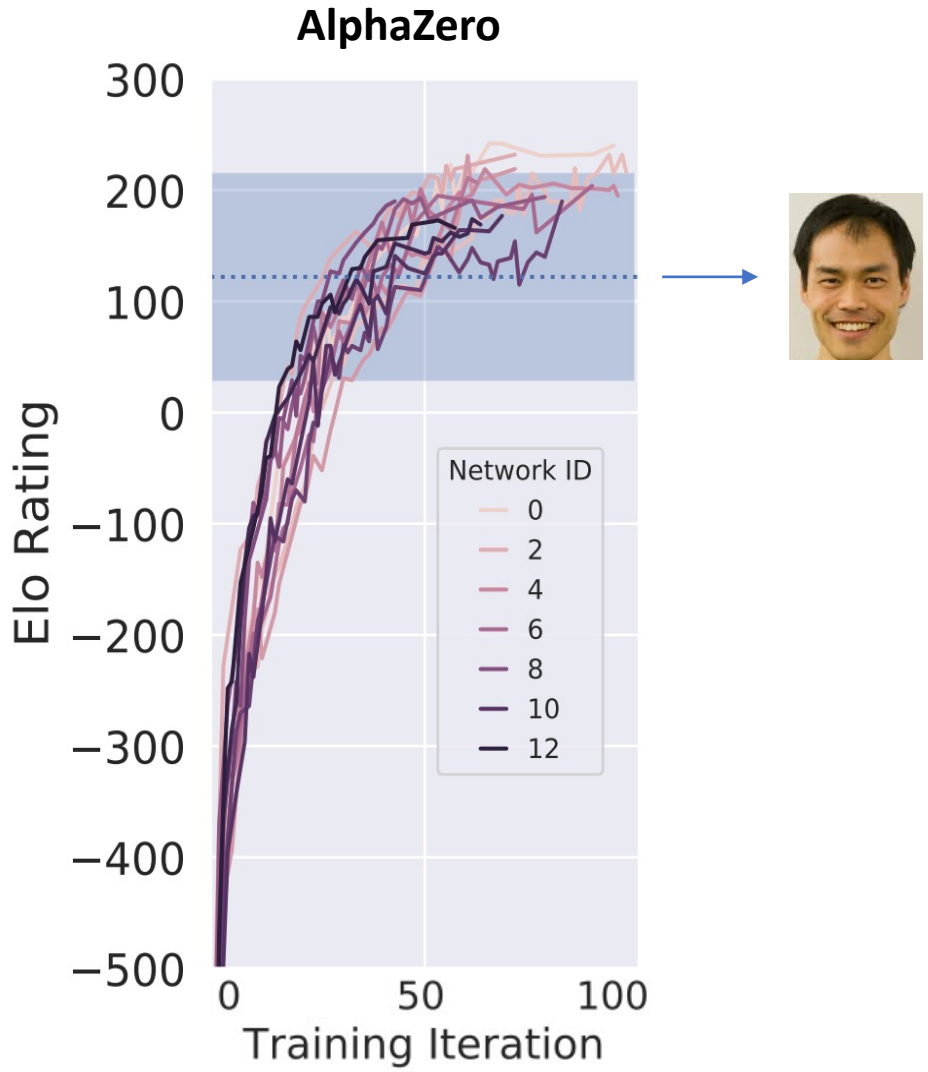
- Intro to AlphaZero and cognitive model
- **AlphaZero learning vs. Human learning**
- Summary and future directions

Methods

- We train 13 AlphaZero agents to play 4-in-a-row with different hyperparameter configurations

- We compare our AlphaZero agents' learning with human learning in previous 4-in-a-row study (*Van Opheusden et al. ,2021)*

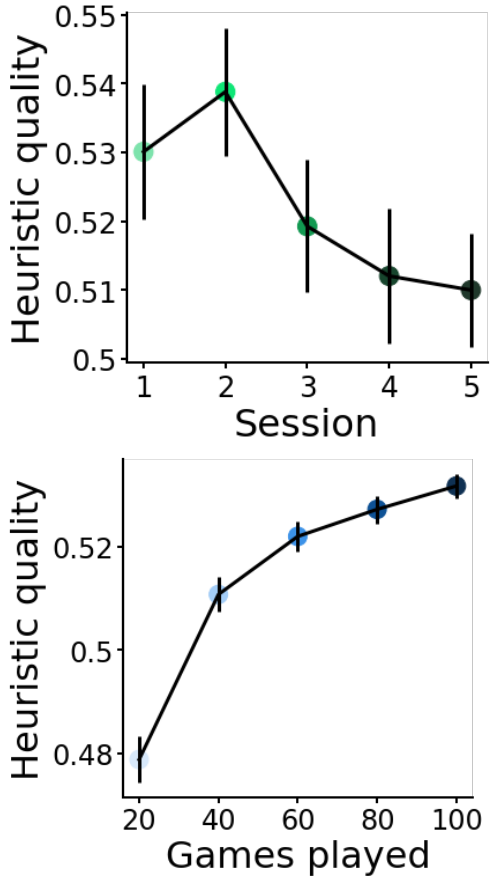
Playing strength vs. human



- AlphaZero agents surpass human expert performance!
- But in what ways do the agents improve?

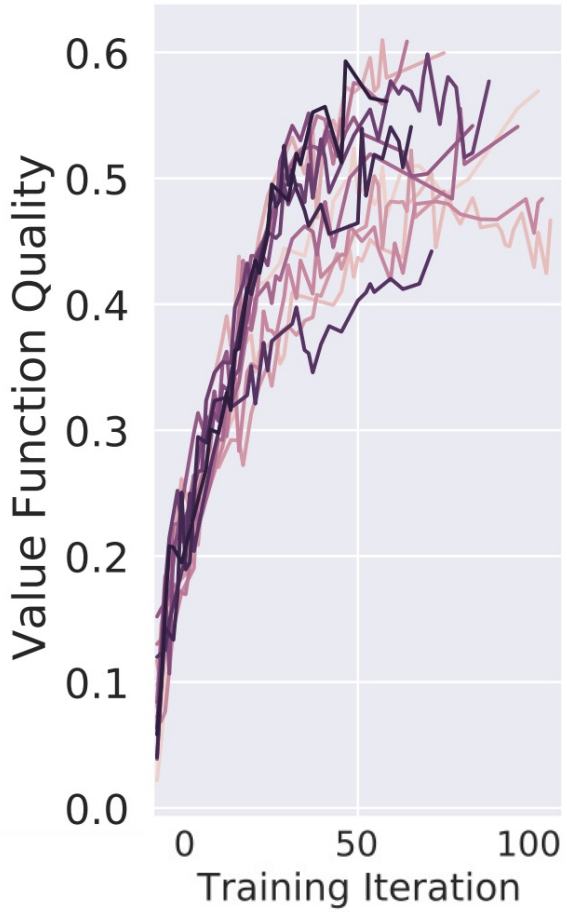
Value function quality

human



(Van Opheusden et al. ,2021)

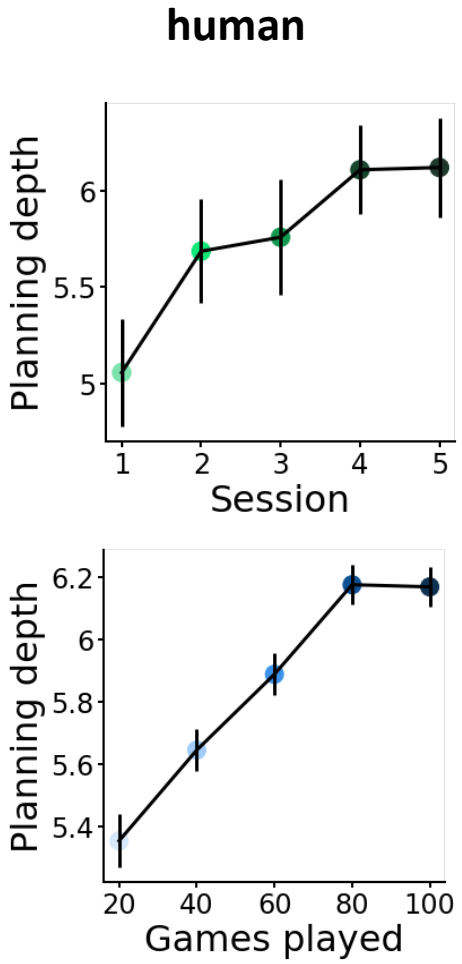
AlphaZero



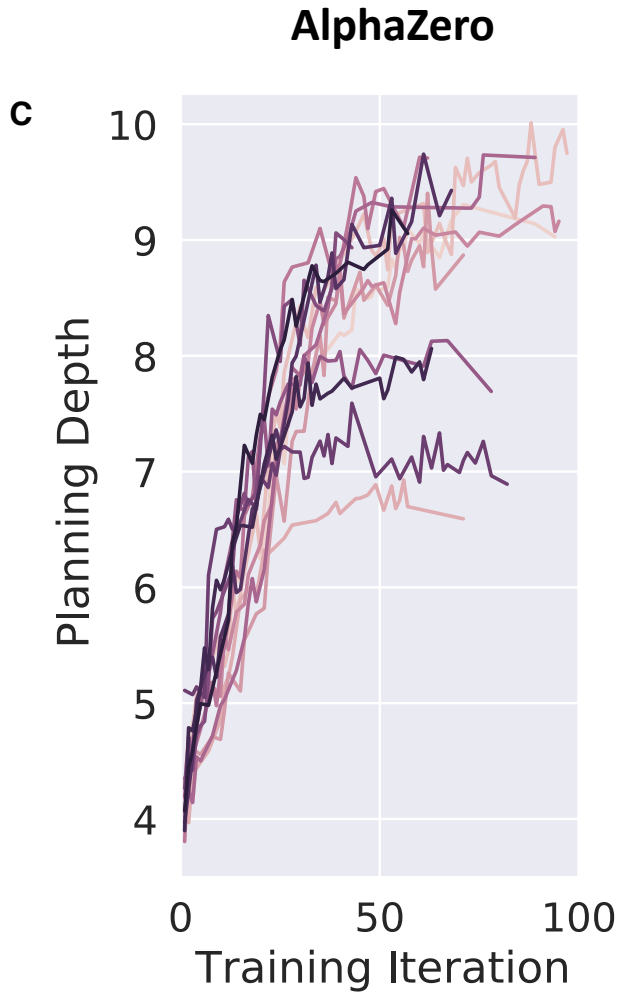
Human improvement

- Both AlphaZero and humans improve their value function quality, but the range of improvement is different
- Humans already start with approximately correct inductive biases, while AlphaZero starts with really bad value function quality

Planning depth



(Van Opheusden et al. ,2021)



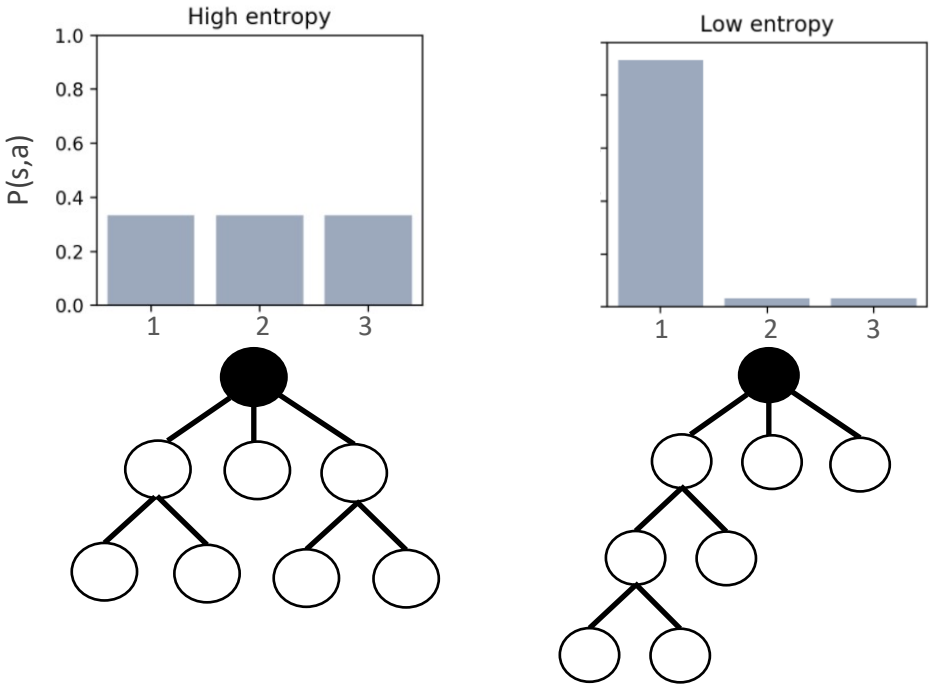
- Planning depth ≈ number of steps into the future one looks ahead
- Both AlphaZero and humans improve their planning depth
- Improvement of planning depth in humans is attributed to more searches
- How does AlphaZero improve planning depth?

Entropy of Action Prior Mediates AlphaZero Planning Depth Increase

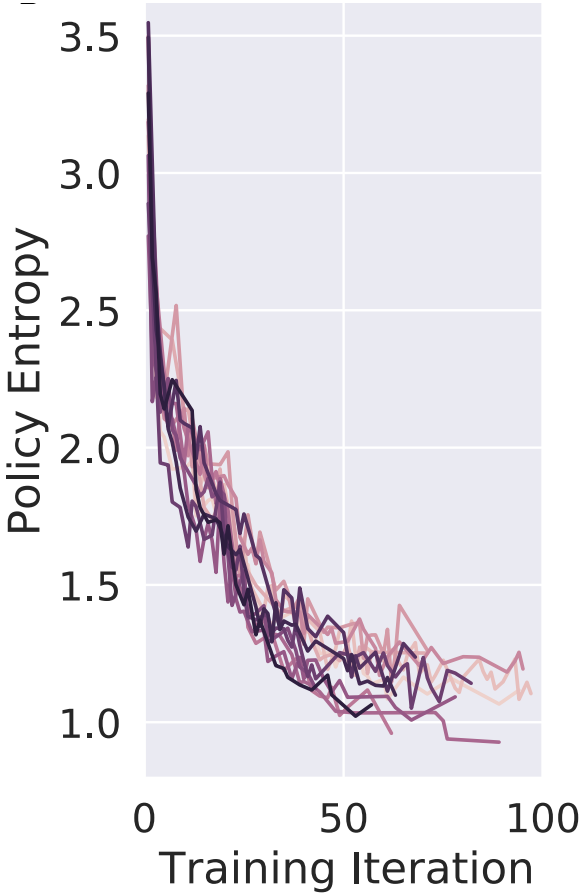
- Entropy: quantify how evenly distributed/concentrated a prior is

$$H(s) = -\sum_a P(s,a) \log P(s,a)$$

- More concentrated prior lead to deeper trees



"smarter" search!



Planning in AI

Intermediate complexity task: 4-in-a-row
AlphaZero trained on the task

- **value function quality and planning depth improve with training**
- **Planning depth increases due to "smarter" searches**

Planning in Cognitive Science

Intermediate complexity task: 4-in-a-row
Cognitive model fitted to human

- **value function quality and planning depth improve with training**
- **Planning depth increases due to "more" searches**

Summary

- We study how AlphaZero learns 4-in-a-row and use metrics that are comparable between AlphaZero and human modelling results (thanks to the intermediate task complexity)

- Similar to human modelling studies, the value function quality and planning depth improve during training, but the range of improvement is different.

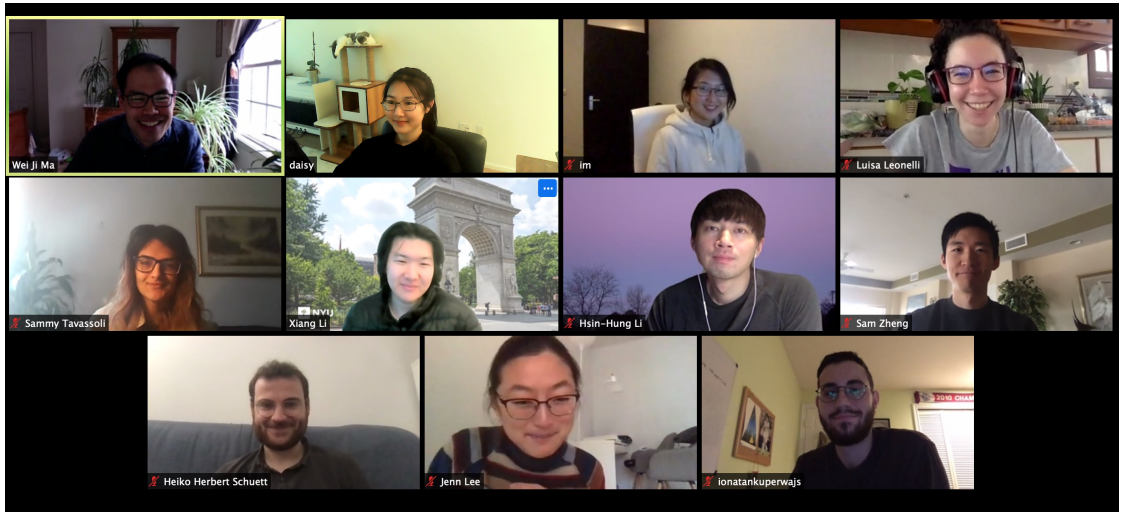
- Different from human modelling studies, AlphaZero improves planning depth through “smarter search” rather than “more search”, which provides new hypothesis for improving the existing cognitive model

Future work

- Develop an action prior metric to access the quality of action prior
- Understand playing strength increase at different training stages
- Analyze features learned in the network and compare it against features humans learn
- Compare AlphaZero choices with choice probability in previous human data on specific board positions to compare the choice bias

Acknowledgement

Wei Ji Ma
Heiko Schütt
Hsin-Hung Li
Xiang Li
Jennifer Laura Lee
Peiyuan Zhang
Ionatan Kuperwajs
Luisa Leonelli
Sam Zheng
Jake Topping



NYU/CNS
Center for Neural Science

MA LAB
HOMINES STERCVS NESCIVNT
MMXIII

NYU
NEUROSCIENCE

Thanks for your attention!

Hyperparameters							
Network ID	$Dir(\alpha)$	c_{PUCT}	color	n_{res}	switch temp.	cont. train-ing	final Elo
0	0.30	2.0	True	3	True	True	242
1	0.03	2.0	True	3	True	True	232
2	0.00	2.0	False	3	False	False	232
3	0.03	2.0	True	9	True	False	231
4	0.00	2.0	True	3	True	False	219
5	0.00	2.0	True	9	True	False	211
6	0.00	2.0	True	3	True	False	204
7	0.00	2.0	False	3	True	True	194
8	0.00	2.0	False	3	True	False	190
9	0.00	0.5	False	3	True	False	190
10	0.03	2.0	True	9	True	True	177
11	0.00	2.0	False	9	True	True	174
12	0.00	2.0	False	9	True	False	173

Table 2: Hyperparameters. $\mathrm{Dir}(\alpha)$: controls the Dirichlet noise added to the action prior. c_{PUCT}: controls the MCTS exploration-exploitation trade-off. *Color*: whether the input feature to DNN includes player color. n_{res}: the number of residual blocks in the DNN. *Switch temp.*: whether the temperature parameter is switched to 0 after 15 moves. *cont. training*: whether an updated network continues its training after losing to the current best network. *Final Elo*: the Elo of the latest iteration.