ETH zürich

Stochastic Planning in Games: an AlphaGo Case-study

Tommaso Macrì

Distributed Computing

AlphaGo beats 18-time world champion Lee Sedol 4 games to 1

The New York Times

It isn't looking good for humanity.

The Guardian

In a major breakthrough for artificial intelligence, AlphaGo Zero took just three days to master the ancient Chinese board game of Go ... with no human help

The Game of Go

The Game of Go, Rules

Aim of the game is to surround more territory than the opponent

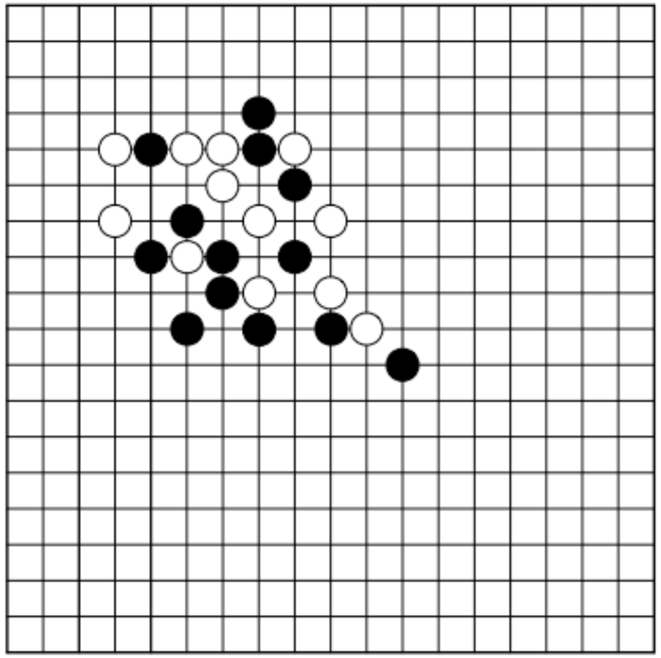
Comparison with Chess

- ~250 legal moves per position
- ~150 moves per game

- ~35 legal moves per position
- ~80 moves per game

Artificial Intelligence perspective

Go game has challenged artificial intelligence researchers for many decades



A Go board configuration

Mastering the Game of Go: a Major Breakthrough for AI

ARTICLE

nature

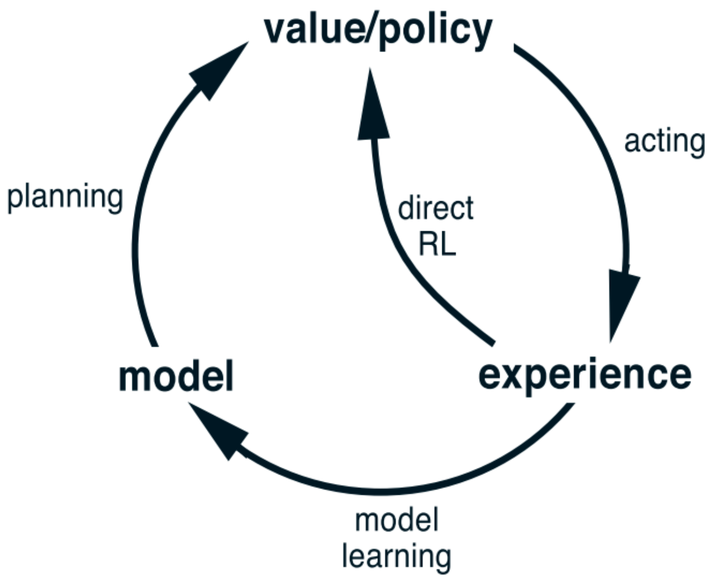
2016

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

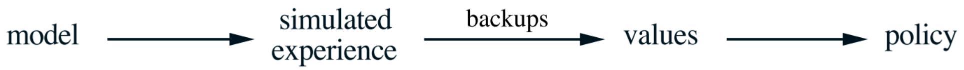
Planning and RL

Planning vs Learning



In Planning, we use the simulated experience to update the value function and policy

In Learning we use Experience Generated by the Environment (not simulated)



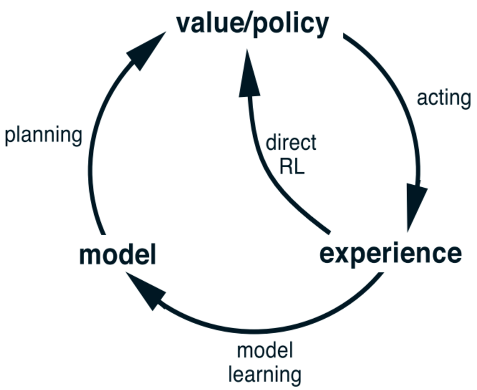
Planning vs Learning: the Dyna-Q example

Tabular Dyna-Q

Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Loop forever:
- (a) $S \leftarrow$ current (nonterminal) state
- (b) $A \leftarrow \varepsilon$-greedy(S,Q)

Direct RL:
- (c) Take action A; observe resultant reward, R, and state, S'
- (d) $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
- (e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment)
- (f) Loop repeat n times:

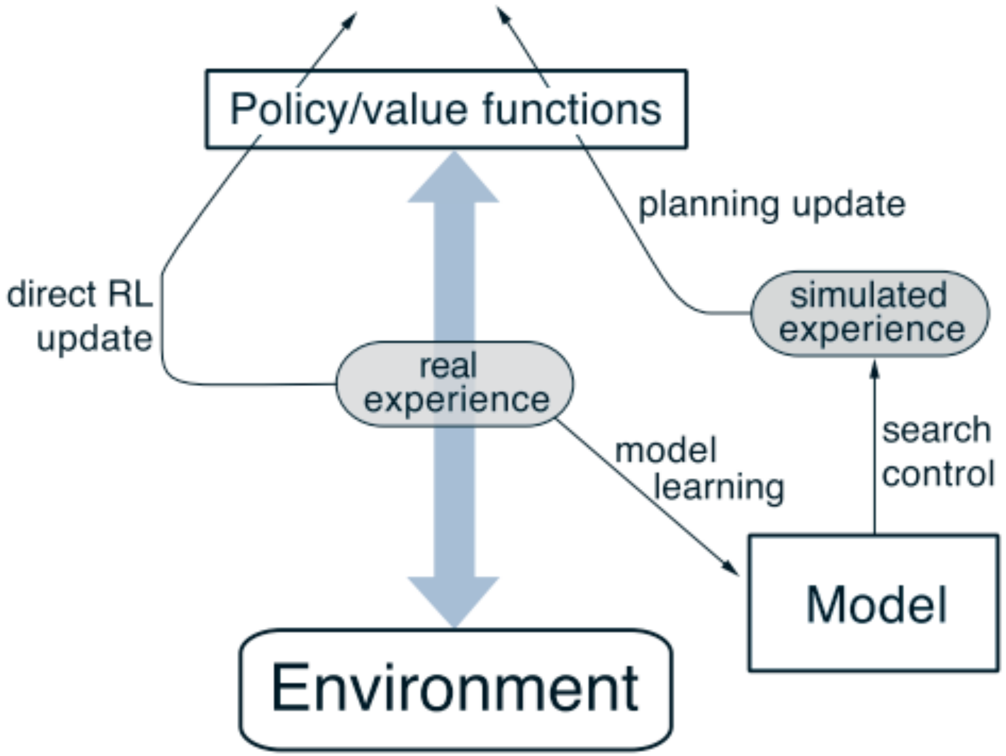
Planning:
  - $S \leftarrow$ random previously observed state
  - $A \leftarrow$ random action previously taken in S
  - $R, S' \leftarrow Model(S,A)$
  - $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$



Reinforcement Learning, Sutton et al.

Planning vs Learning: pros and cons

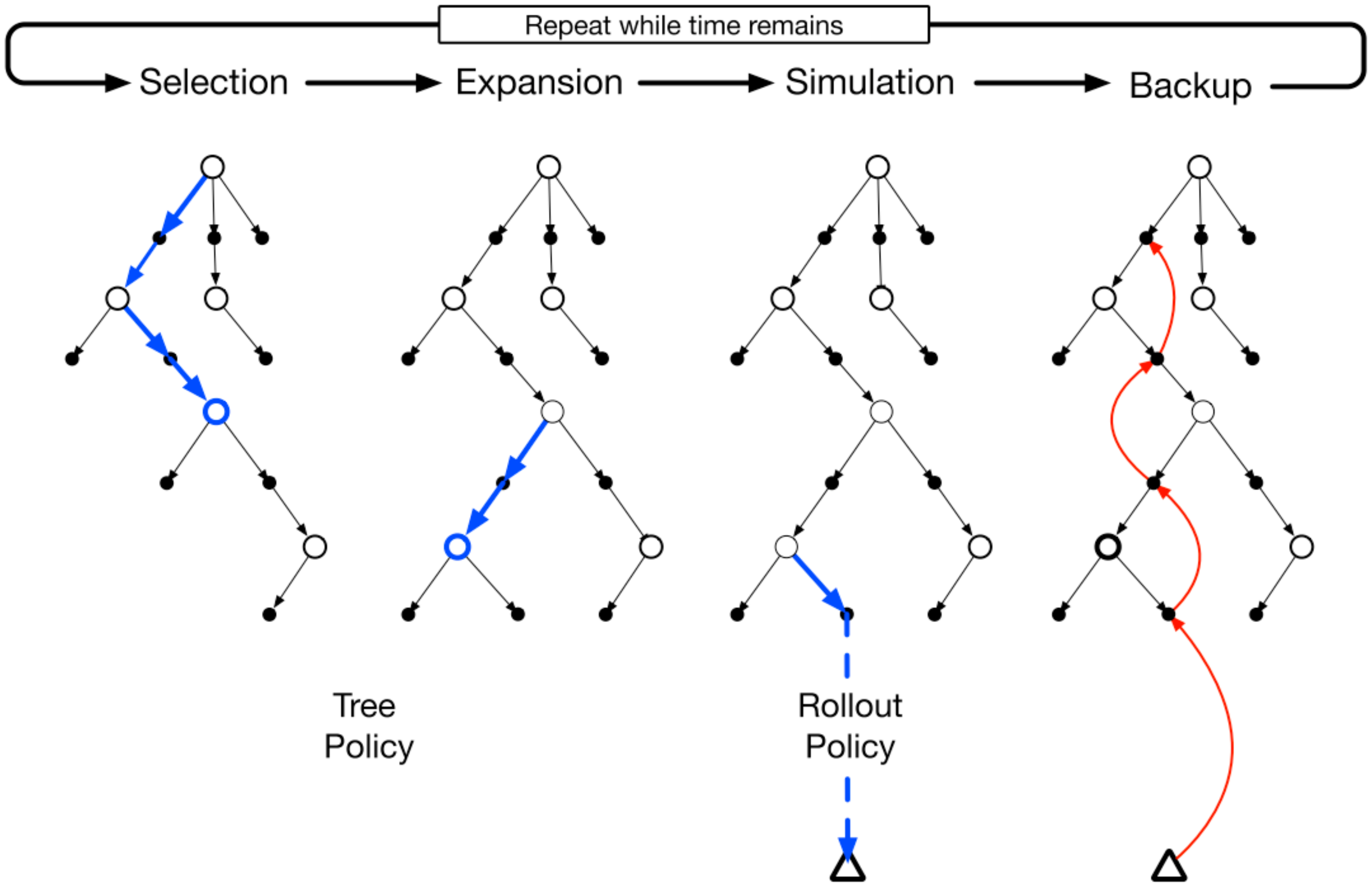
Direct vs Undirect Learning



Reinforcement Learning, Sutton et al.

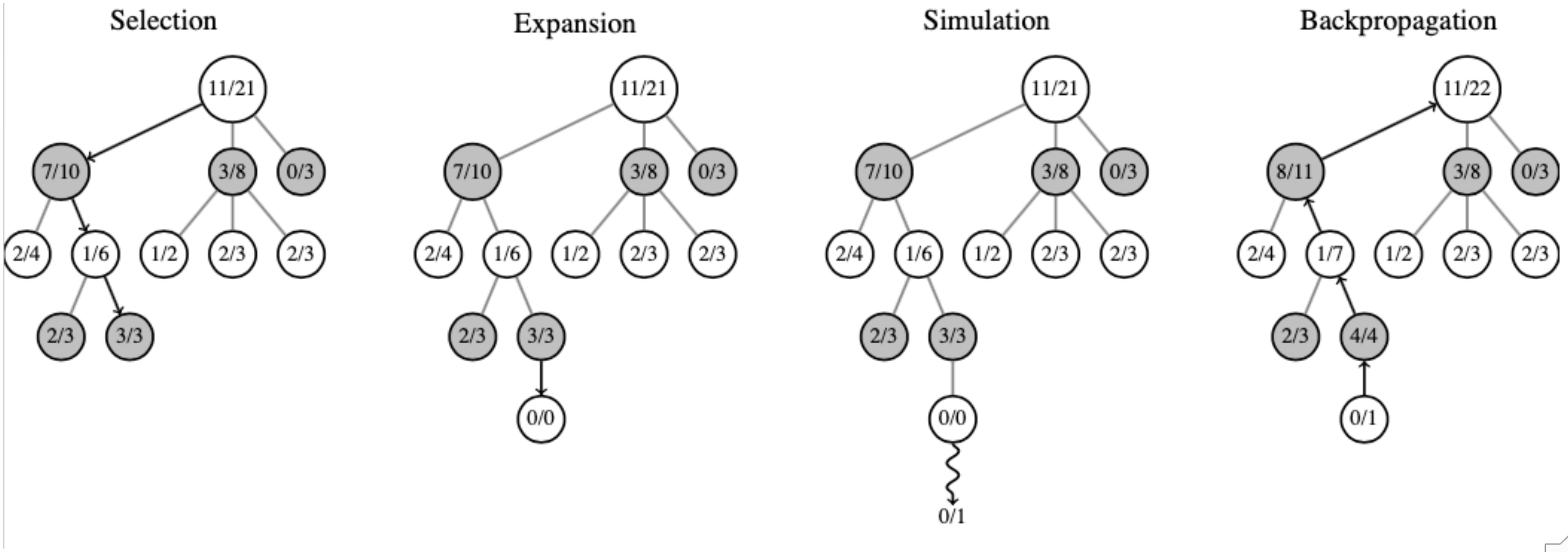
MCTS
(Monte Carlo Tree Search)

Monte Carlo Tree Search



Reinforcement Learning, Sutton et al.

Monte Carlo Tree Search



https://en.wikipedia.org/wiki/Monte_Carlo_tree_search

The AlphaGo Breakthrough

AlphaGo

ARTICLE

nature

2016

Mastering the game of Go with deep neural networks and tree search

David Silver[1*], Aja Huang[1*], Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

AlphaGo

- Go is a perfect information game.
- Compute optimal value function?

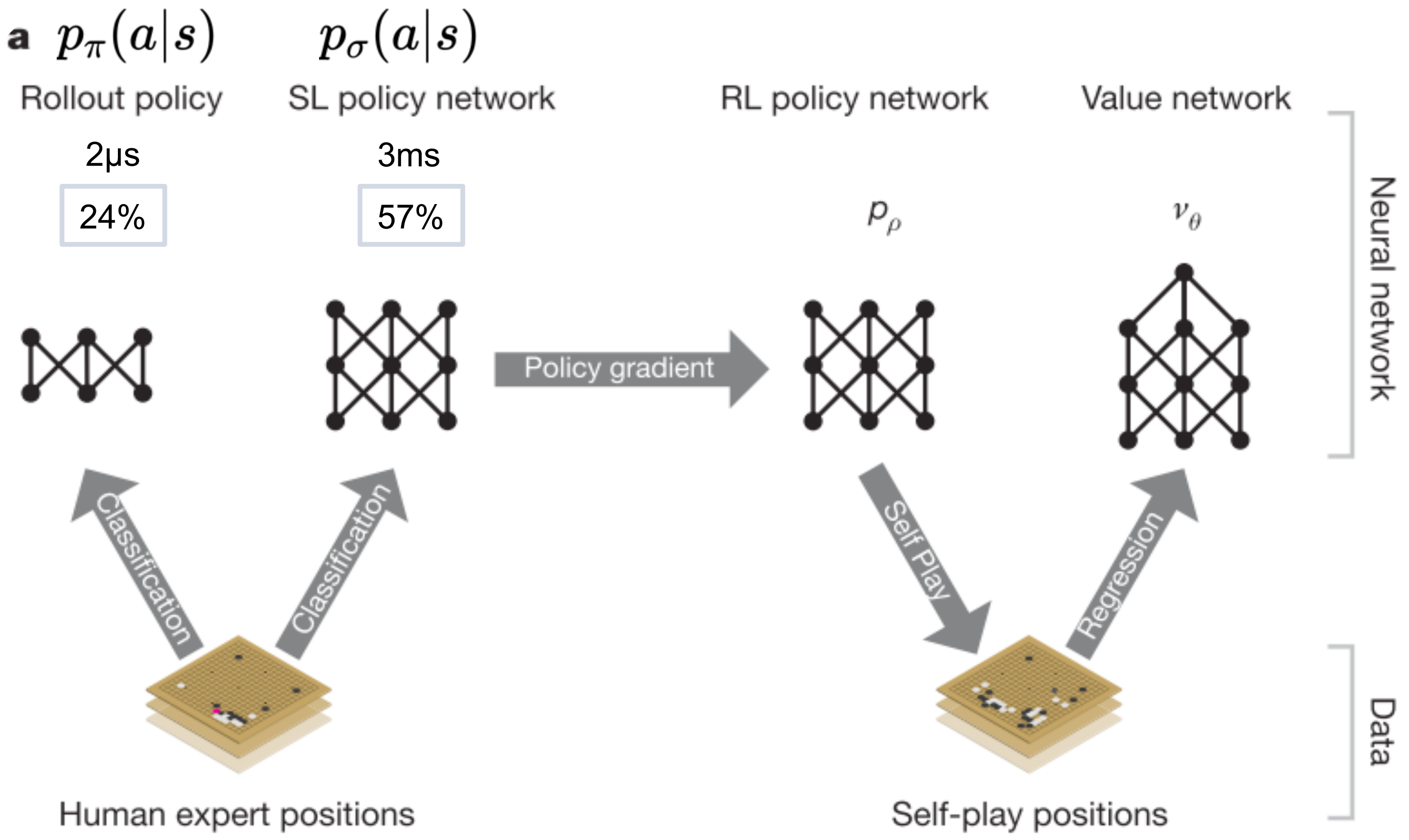
- Tree search = b^d
- b = 250; d = 150

- Exhaustive search is infeasible

AlphaGo: the 4 Neural Networks

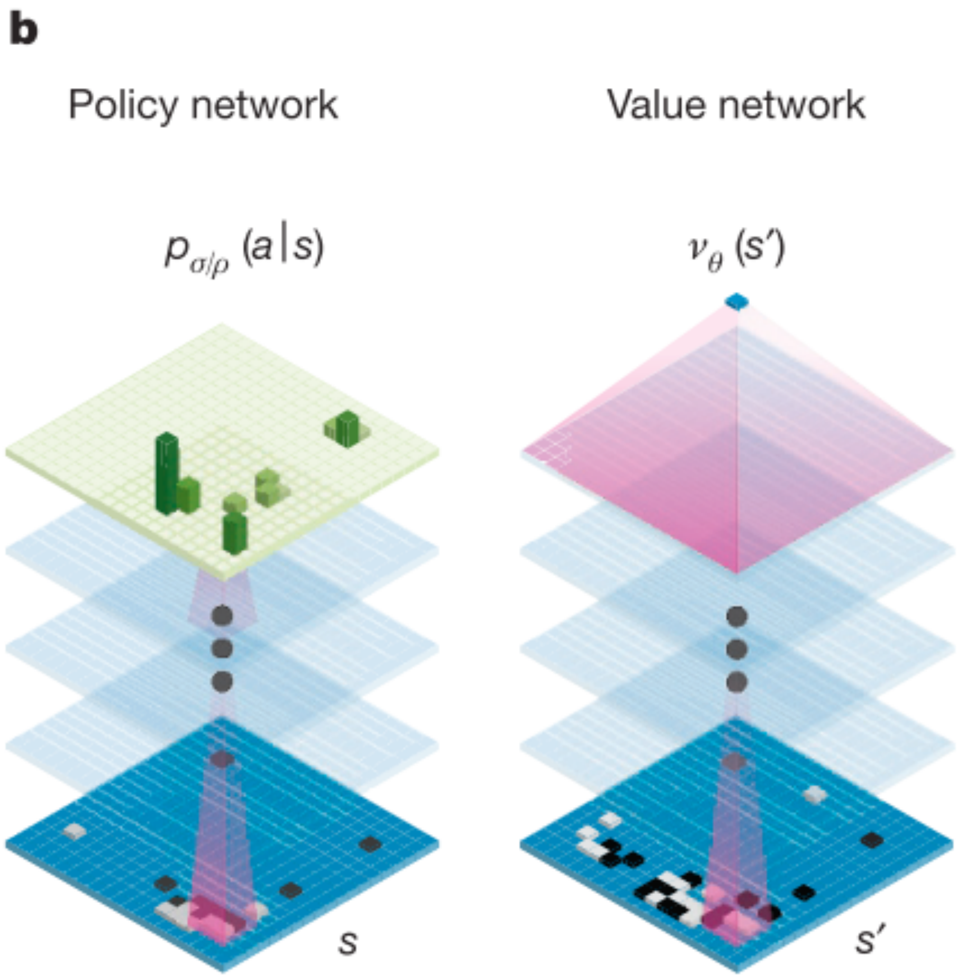
Forward Propagation: 2µs 3ms

Accuracy: 24% 57%



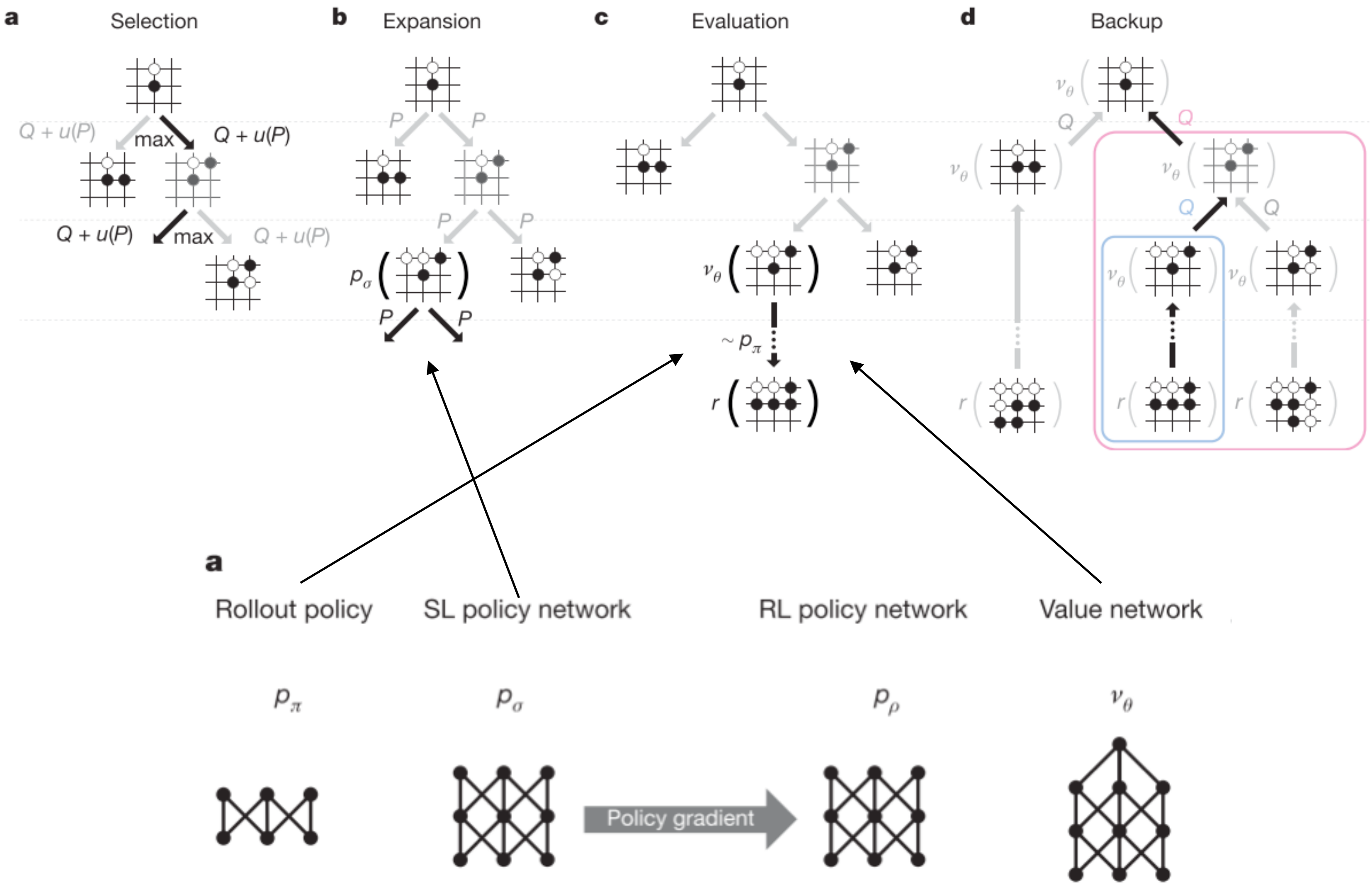
Mastering the game of go with deep neural networks and tree search,
David Silver et al.

AlphaGo: multiple outputs for the policy and single for the value



Mastering the game of go with deep neural networks and tree search, David Silver et al.

AlphaGo: combining Policy Network and Value Network with MCTS



Mastering the game of go with deep neural networks and tree search, David Silver et al.

AlphaGo Zero

ARTICLE

nature

2017

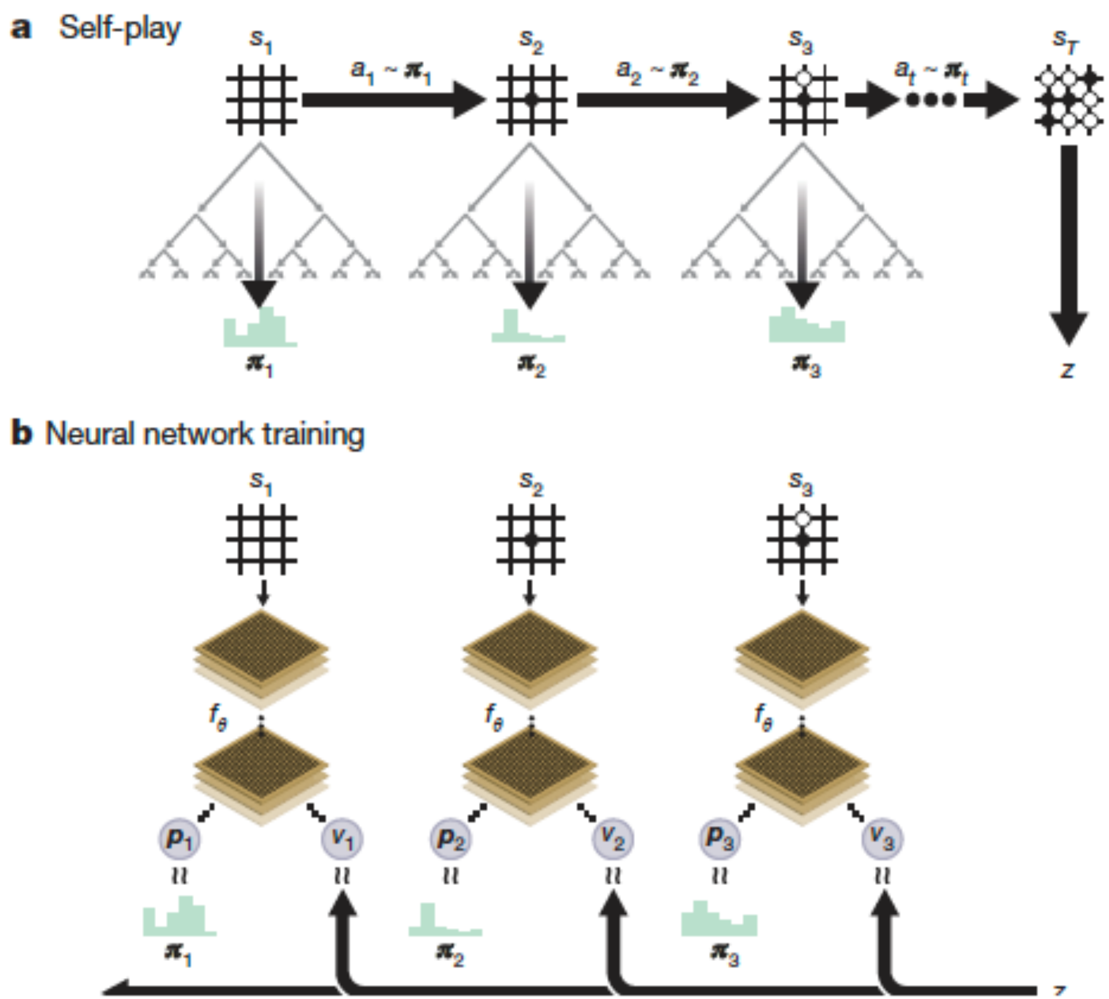
Mastering the game of Go without human knowledge

David Silver[1*], Julian Schrittwieser[1*], Karen Simonyan[1*], Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

AlphaGo Zero: Only one Neural Network for Policy and Value

$$(\boldsymbol{p}, v) = f_\theta(\boldsymbol{s})$$

AlphaGo Zero: Using MCTS to select moves throughout self-play



Mastering the game of go without human knowledge, David Silver et al.

Alpha Zero

RESEARCH

nature 2018

COMPUTER SCIENCE

A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play

David Silver[1,2]*†, Thomas Hubert[1]*, Julian Schrittwieser[1]*, Ioannis Antonoglou[1], Matthew Lai[1], Arthur Guez[1], Marc Lanctot[1], Laurent Sifre[1], Dharshan Kumaran[1], Thore Graepel[1], Timothy Lillicrap[1], Karen Simonyan[1], Demis Hassabis[1]†

Alpha Zero: Learning and MCTS do not assume symmetry

Alpha Go and Alpha Go Zero assumed symmetry for:
- Training data augmentation
- Bias removal in Monte Carlo evaluations

Alpha Zero considers also the “Drawn” outcome

Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.

Alpha Zero: The Algorithm Architecture

Alpha Go Zero, Alpha Zero:

- one Neural Network (Policy + Value)
- Monte Carlo Tree Search

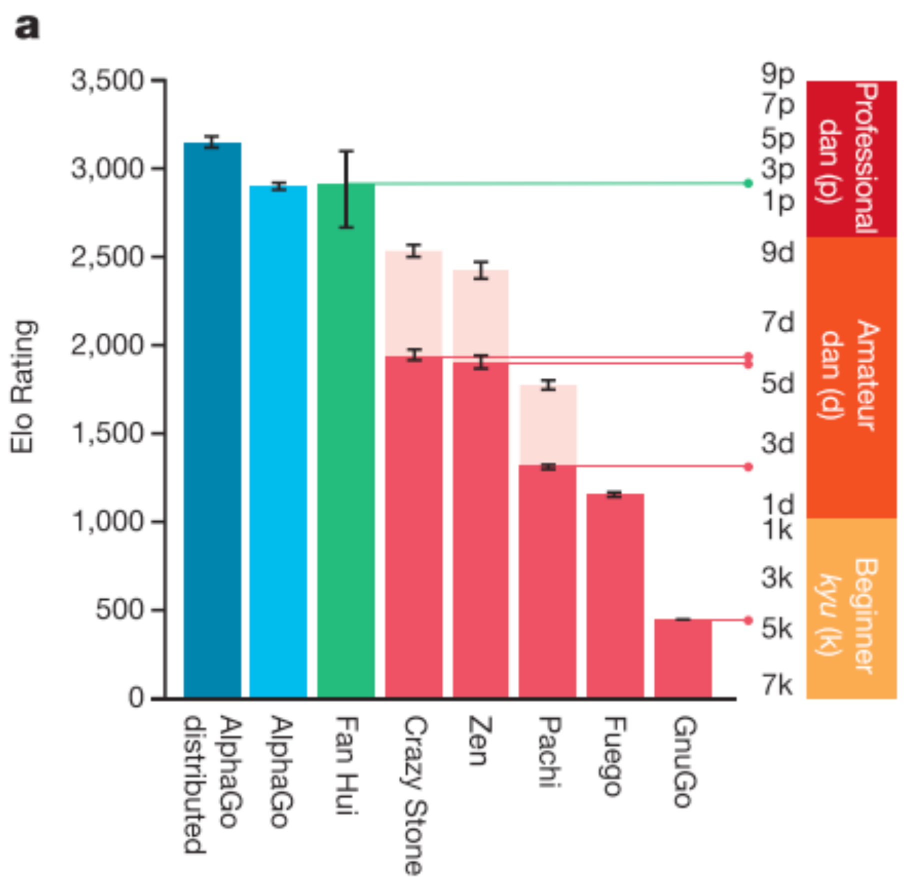
Alpha Go Zero

- Wait for an iteration to conclude to update NN
- Compare the new policy to the best

Alpha Zero

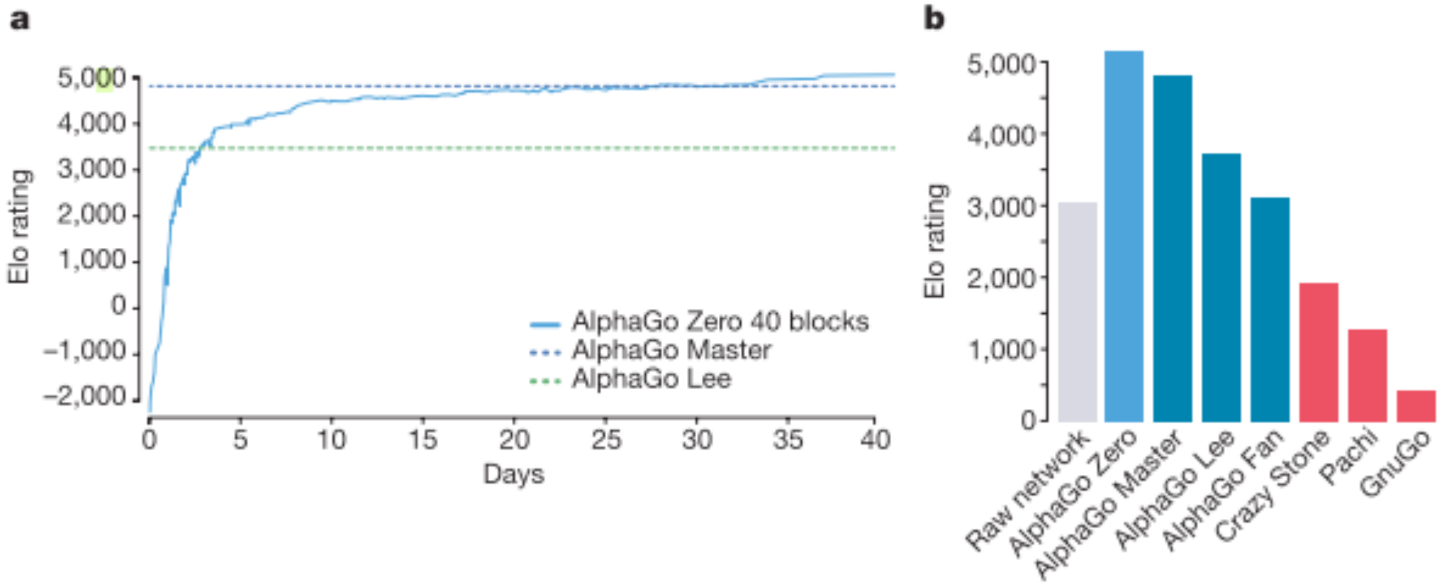
- Update the NN continuously
- Always generate Self-Play with the latest NN

Results: AlphaGo beats the European Go Champion



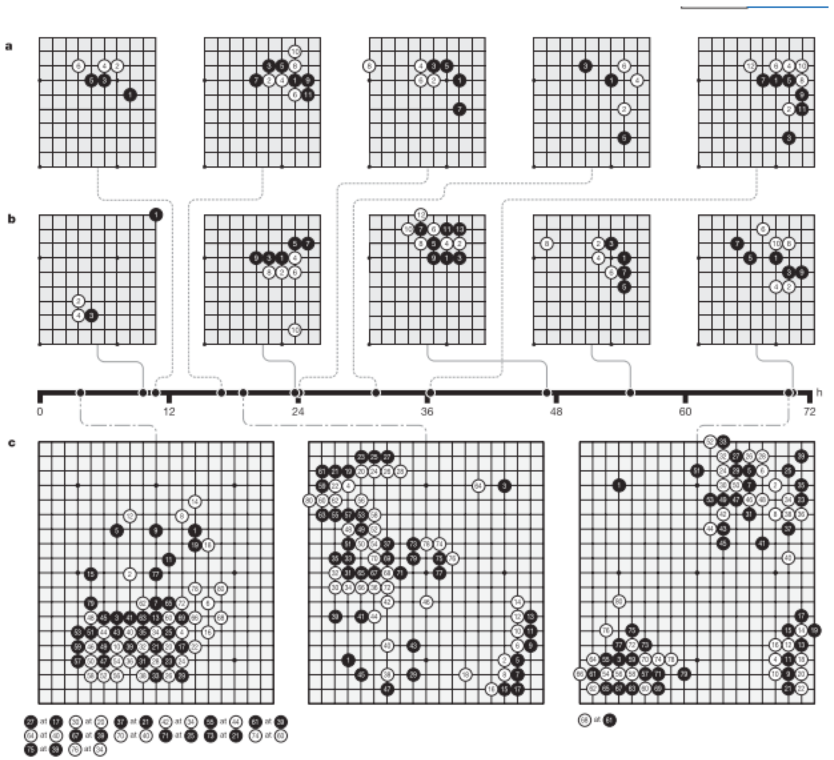
Mastering the game of go with deep neural networks and tree search,
David Silver et al.

AlphaGo Zero: better performances than AlphaGo



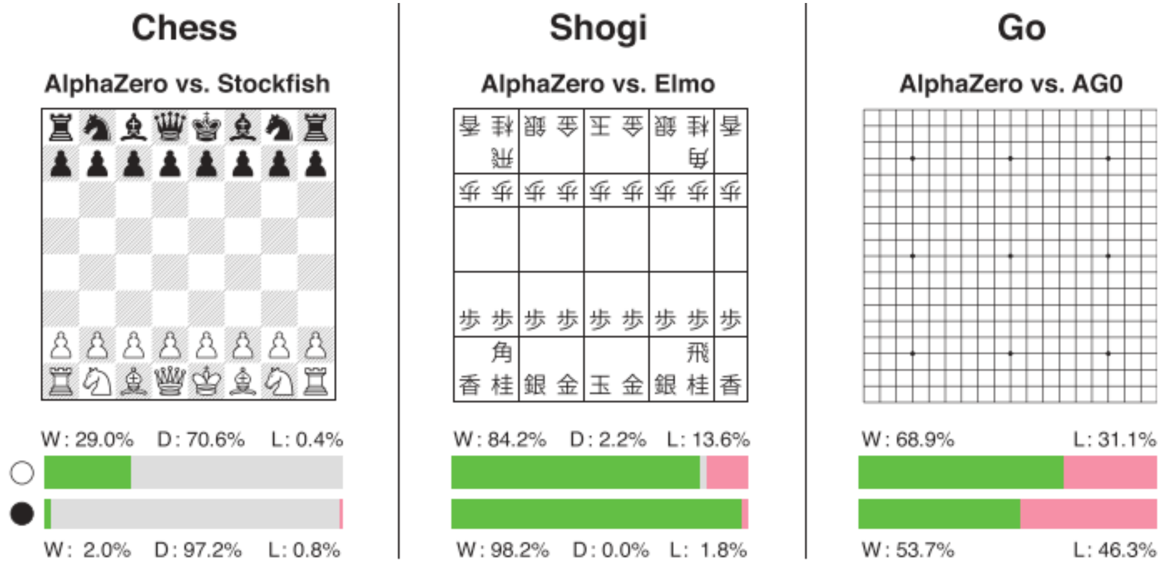
Mastering the game of go without human knowledge, David Silver et al.

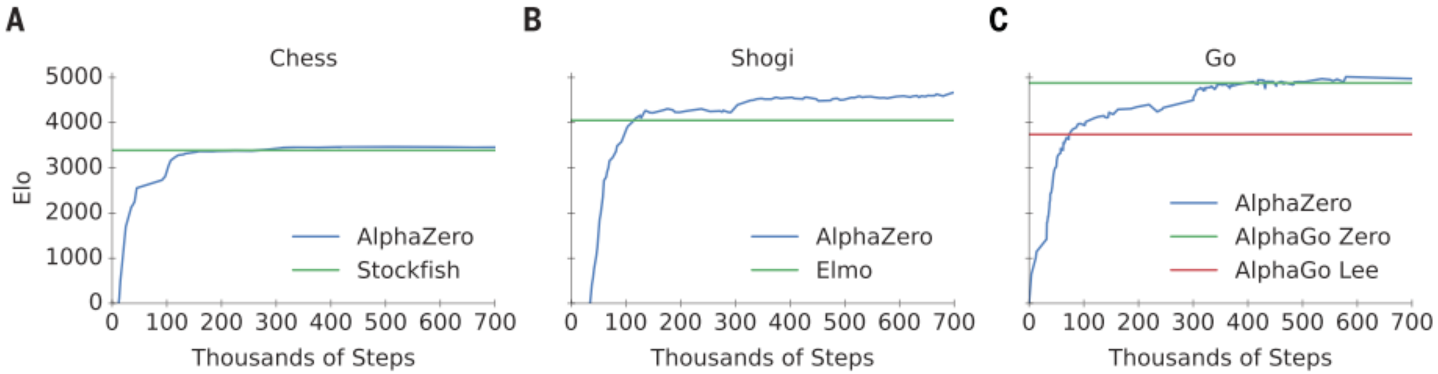
AlphaGo Zero: Learns human expert moves and beyond



Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.

Alpha Zero: program applied to Chess and Shogi





Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, David Silver et al.

Conclusions

"I can't disguise my satisfaction that it plays with a very dynamic style, much like my own!"

GARRY KASPAROV
FORMER WORLD CHESS CHAMPION

"The implications go far beyond my beloved chessboard... Not only do these self-taught expert machines perform incredibly well, but we can actually learn from the new knowledge they produce."

GARRY KASPAROV
FORMER WORLD CHESS CHAMPION

https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

ETH zürich

Tommaso Macrì

Stochastic Planning in Games: an AlphaGo Case-study