Monte Carlo Search

Tristan Cazenave

LAMSADE

Université Paris-Dauphine PSL CNRS

Tristan.Cazenave@dauphine.fr

16 Novembre 2023

Outline

- Monte Carlo Tree Search
- Nested Monte Carlo Search
- Nested Rollout Policy Adaptation
- AlphaGo and AlphaGo Zero
- Athénan
- AlphaMu

Monte Carlo Tree Search

Monte Carlo Tree Search

- Best search algorithm for many games since 2007
- All the winners of the General Game Playing competition since 2007 use MCTS
- AlphaGo, AlphaGo Zero and Alpha Zero use a MCTS variant named PUCT
- It can be combined with deep learning in deep reinforcement learning systems so as to learn to play games at a superhuman level from scratch

UCT

- UCT : Exploration/Exploitation dilemma for trees [Kocsis and Szepesvari 2006].
- Play random random games (playouts).
- Exploitation : choose the move that maximizes the mean of the playouts starting with the move.
- Exploration : add a regret term (UCB).

UCT

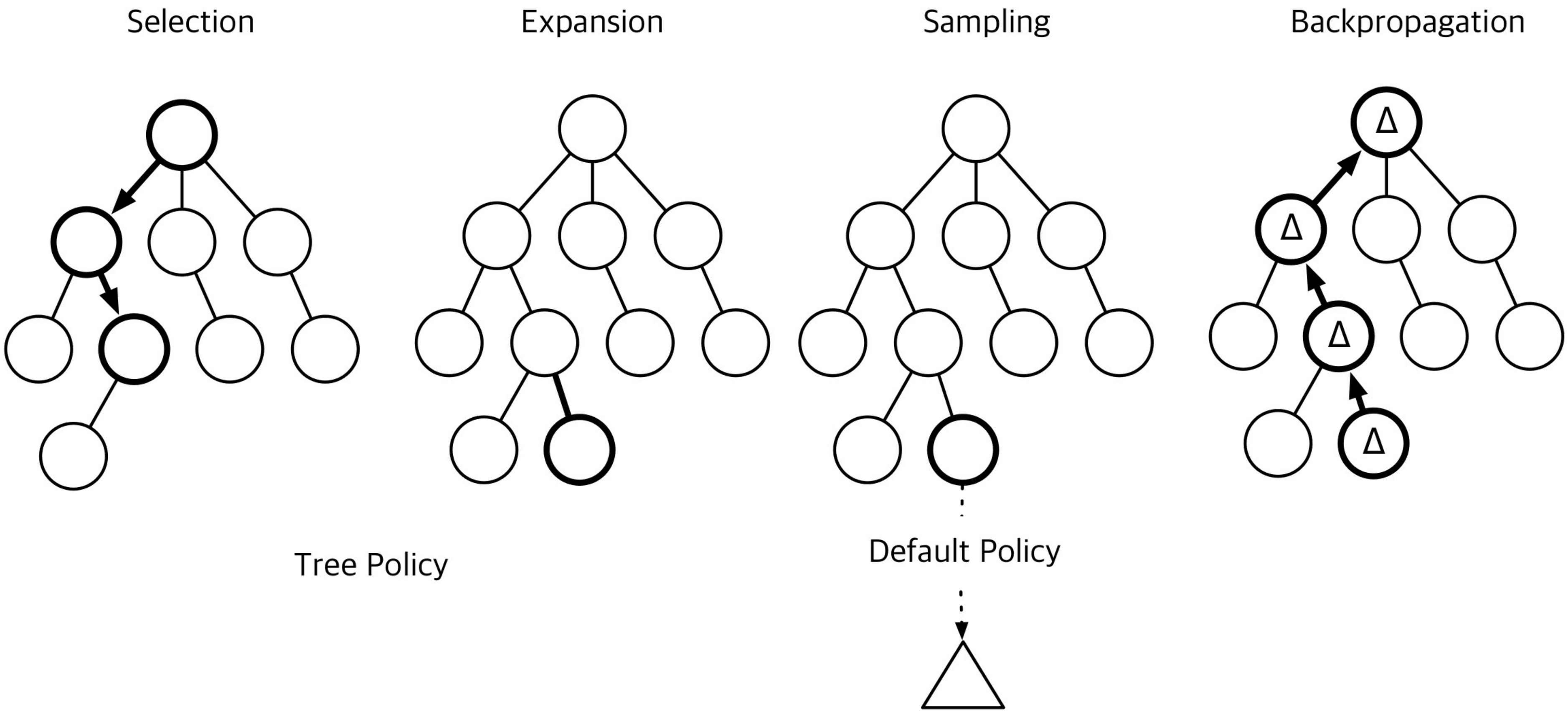
- UCT : exploration/exploitation dilemma.
- Play the move that maximizes

$$\frac{w_i}{n_i} + c\sqrt{\frac{\ln t}{n_i}}$$

In which

- w_i = number of wins after the i-th move
- n_i = number of simulations after the i-th move
- c = exploration parameter (theoretically equal to √2)
- t = total number of simulations for the parent node

UCT


Selection
Expansion
Sampling
Backpropagation
Δ
Δ
Δ
Δ
Tree Policy
Default Policy

RAVE

- A large improvement for Go, Hex and other games is Rapid Action Value Estimation (RAVE) [Gelly and Silver 2007, 2011].

- RAVE combines the mean of the playouts that start with the move and the mean of the playouts that contain the move (AMAF).

GRAVE

- Generalized Rapid Action Value Estimation (GRAVE) [Cazenave 2015] is a simple modification of RAVE.
- It consists in using the first ancestor node with more than n playouts to compute the RAVE values.
- It is a large improvement over RAVE for Go, Atarigo, Knightthrough and Domineering.
- State of the art in General Game Playing.

PUCT

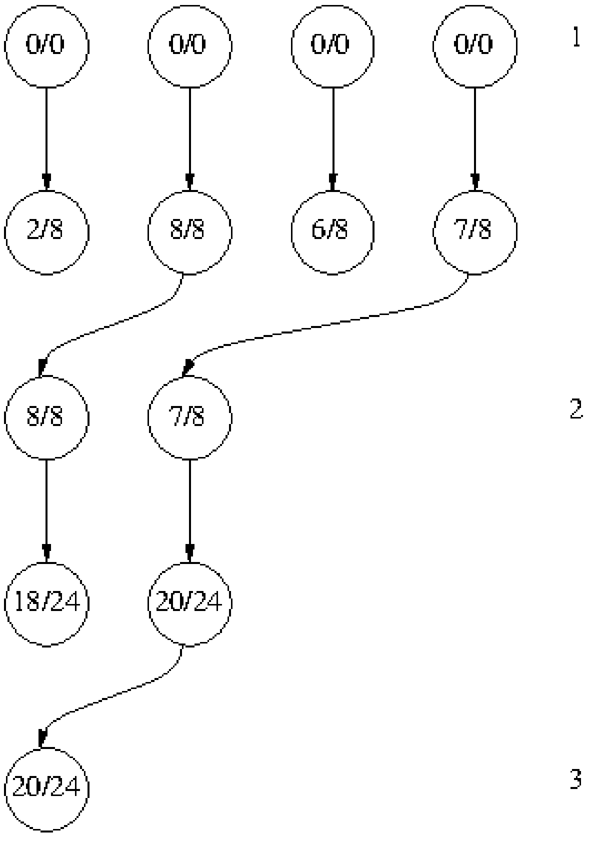
- MCTS used in AlphaGo and AlphaZero.
- A neural network gives a policy and a value.
- No playouts, evaluation with the value at the leaves.
- P(s,a) = probability for move a of being the best.
- Bandit for the tree descent:

$$U(s,a) = c_{\text{puct}} P(s,a) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)}$$

Sequential Halving

- Sequential Halving [Karnin & al. 2013] is a bandit algorithm that minimizes the simple regret.
- It has a fixed budget of arm pulls.
- It gives the same number of playouts to all the arms.
- It selects the best half.
- Repeat until only one move is left

Sequential Halving



Sequential Halving

- Combining Sequential Halving and UCT :
  - Sequential Halving at the root
  - UCT deeper in the tree

- The combination gives good results for Atarigo, Breakthrough, Amazons and partially observable games.

SHUSS

- Sequential Halving combined with other statistics such as AMAF statistics.
- Instead of selecting the best half with the mean (mu_i), use:

$$mu_i + c * AMAF_i / p_i$$

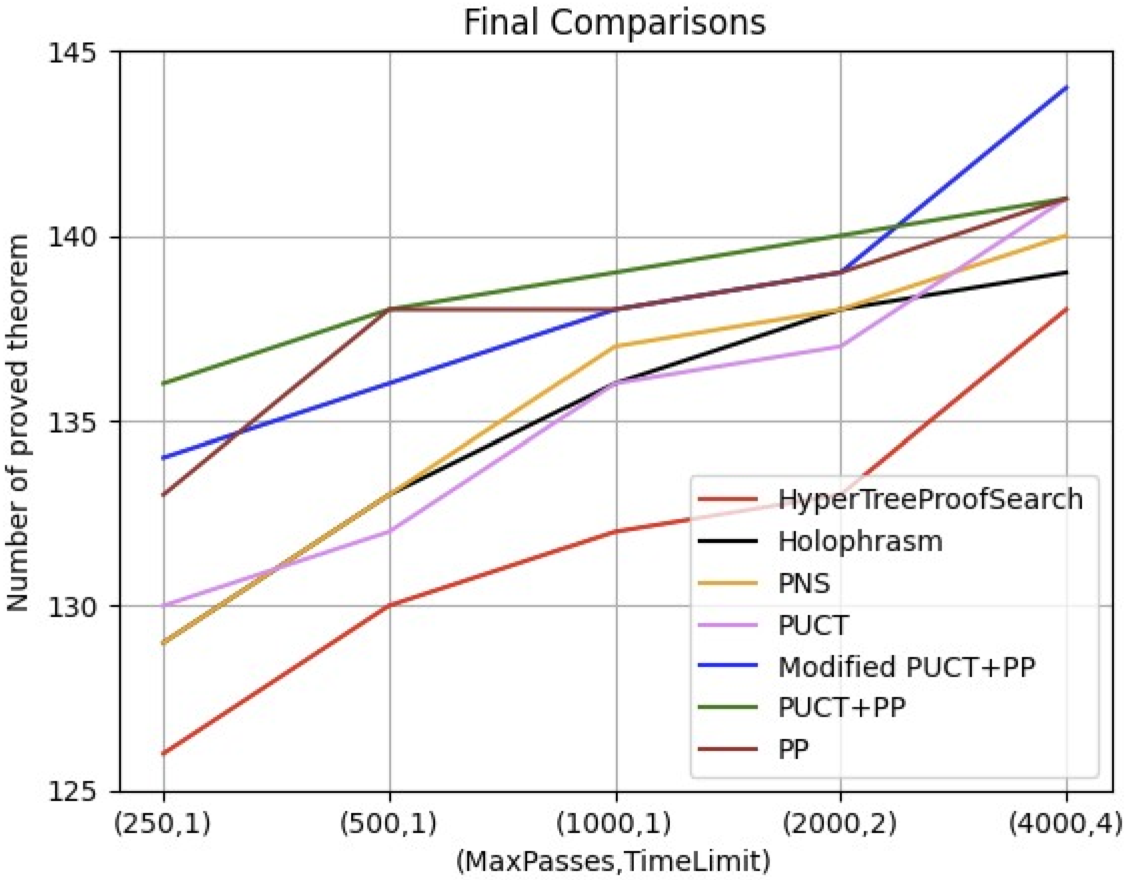
with p_i the number of playouts of move i and $c \geq 128$.

- Combining SH with AMAF = SHUSS (Sequential Halving Using Scores) [Fabiano et al. 2021]

Automated Theorem Proving

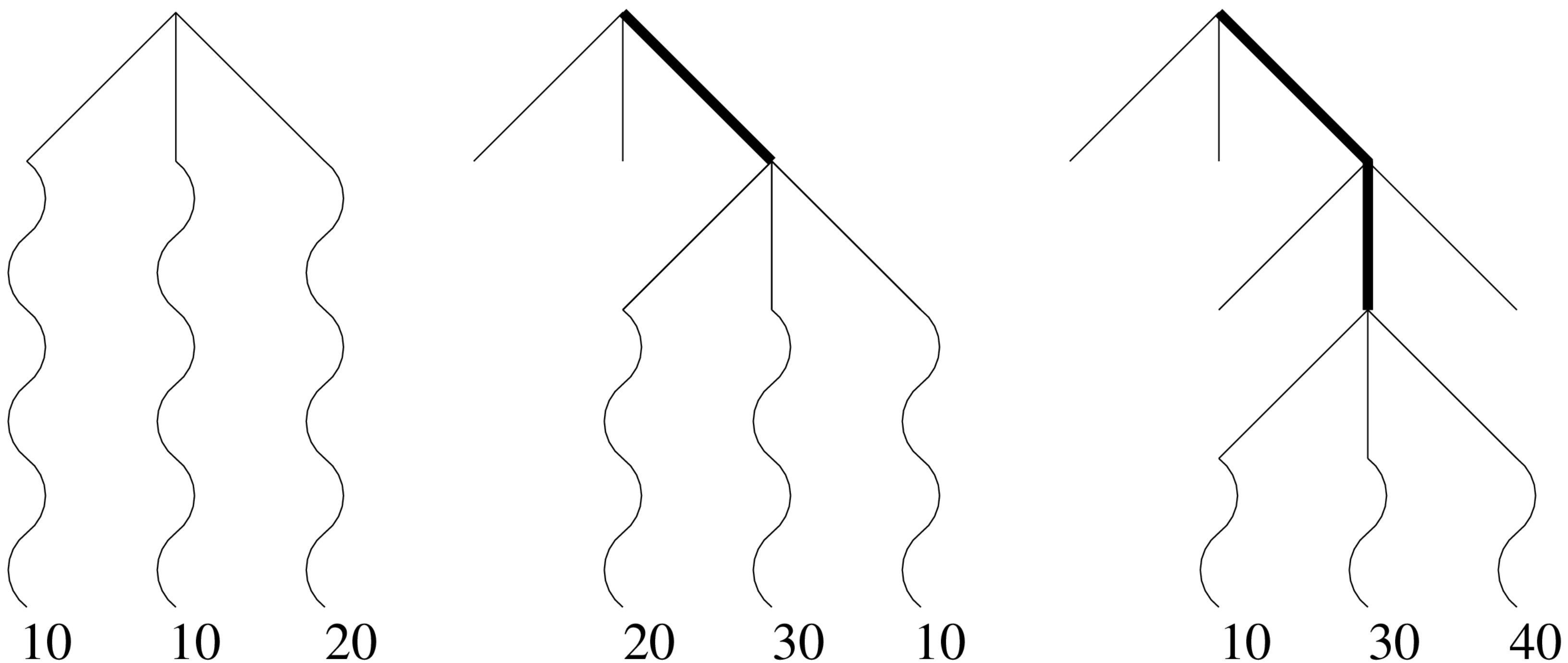
- The state space is an AND/OR tree as in games.
- Algorithms for solving games can be used to prove theorems.
- MCTS has been used in some theorem provers.
- Holophrasm [Daniel Whalen 2016].
- Tactictoe [Gauthier et al. 2021].

Automated Theorem Proving

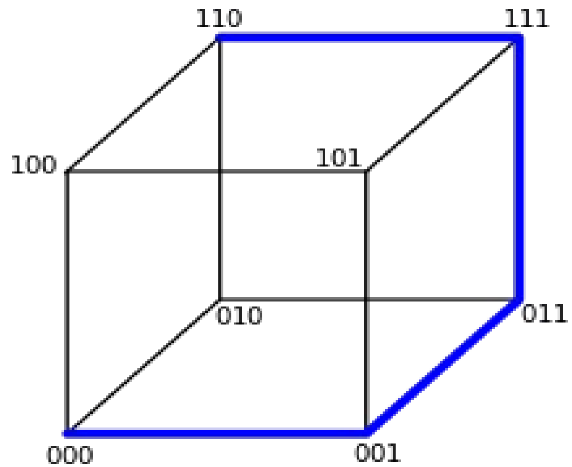


Nested Monte Carlo Search

Nested Monte-Carlo Search

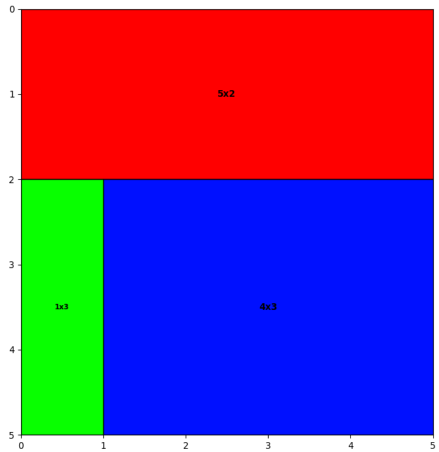


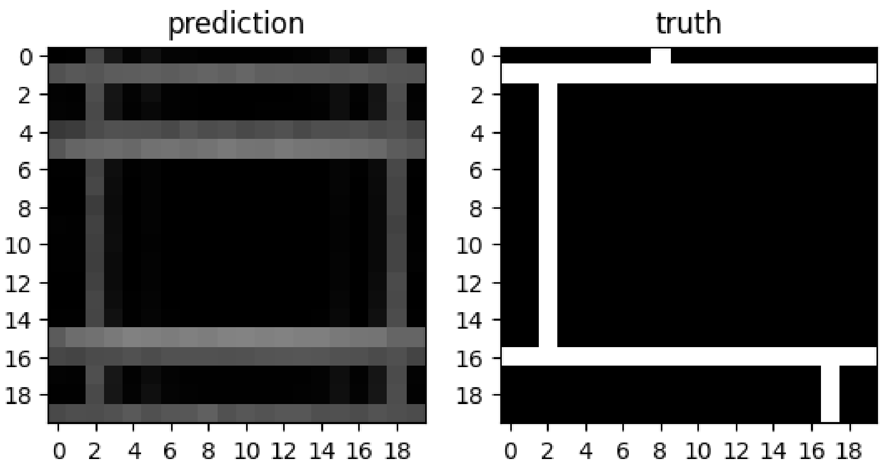
Snake in the box



- A path such that for every node only two neighbors are in the path.
- Applications: Electrical engineering, coding theory, computer network topologies.
- World records with NMCS [Kinny 2012].

Perfect Rectangle Packing



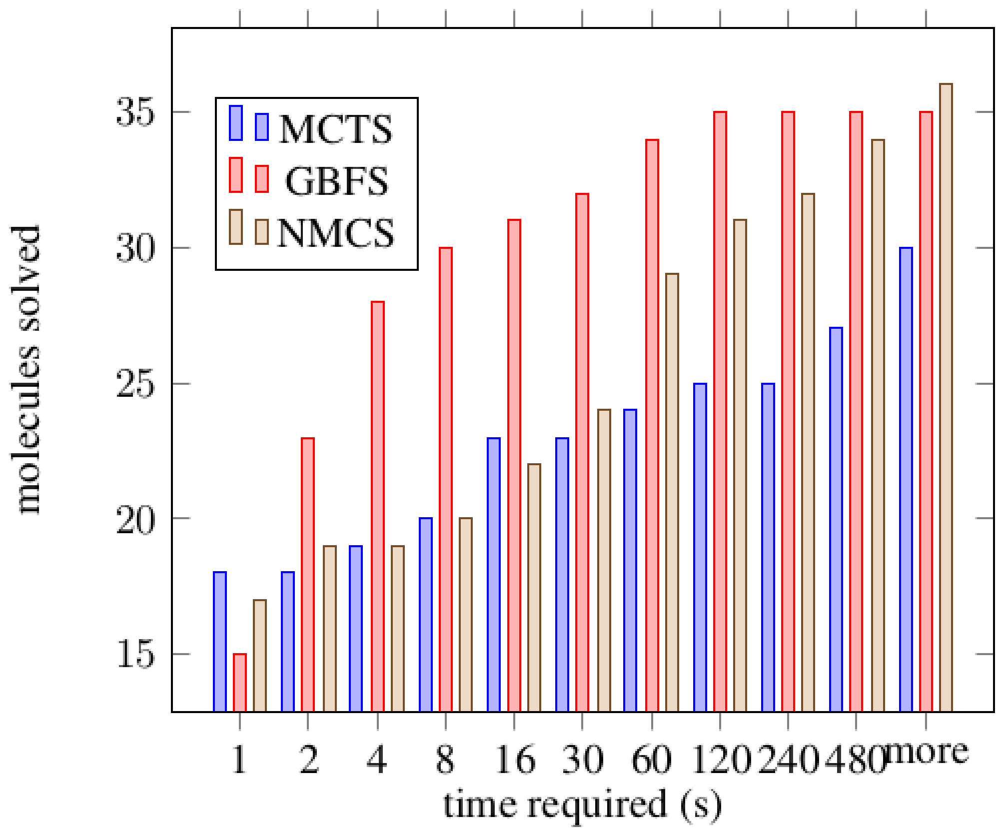


- Learning a policy with a neural network trained on solved instances.
- NMCS with playouts following the learned policy improves much on the uniform policy [Doux et al. 2022].

Retrosynthesis

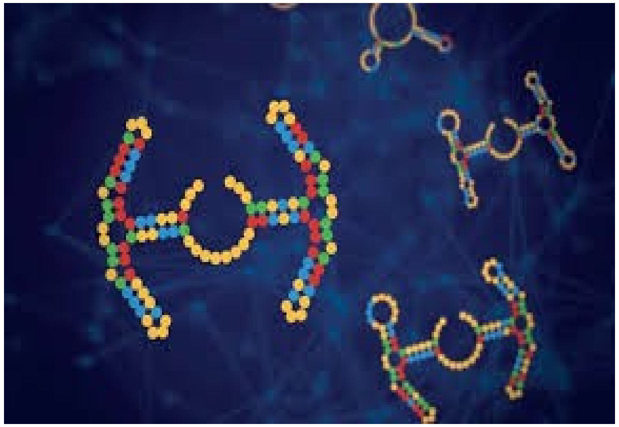
- Find a set of chemical reactions that enable to synthetize a given molecule.
- The state space is an AND/OR tree as in games.
- DF-PN and MCTS have been used to find retrosynthesis pathways.
- Alphachem [Segler et al. 2017].
- AiZynthFinder [Genheden et al. 2020].

Retrosynthesis



RNA Inverse Folding

- Find a sequence that has a given folding



RNA Inverse Folding

- Molecule Design as a Search Problem
- Find the sequence of nucleotides that gives a predefined structure.
- A biochimist applied Nested Monte Carlo Search to this problem [Portela 2018].
- Better than the state of the art.
- GNRPA generalizes the approach.

Applications

Nested Monte Carlo Search :

- Morpion Solitaire [Cazenave 2009]
- SameGame [Cazenave 2009]
- Sudoku [Cazenave 2009]
- Expression Discovery [Cazenave 2010]
- The Snake in the Box [Kinny 2012]
- Cooperative Pathfinding [Bouzy 2013]
- Software Testing [Poulding et al. 2014]
- Heuristic Model-Checking [Poulding et al. 2015]
- Pancake problem [Bouzy 2015]
- Games [Cazenave et al. 2016]
- Cryptography [Dwivedi et al. 2018]
- RNA inverse folding problem [Portela 2019]
- Perfect Rectangle Packing [Doux et al. 2022]
- Retrosynthesis [Roucairol et al. 2023]
- …

Nested Rollout Policy Adaptation

Nested Rollout Policy Adaptation

- NRPA is NMCS with policy learning.
- It adapts the weights of the moves according to the best sequence of moves found so far.
- There are multiple levels of best sequences.
- During adaptation each weight of a move of the best sequence is incremented and all the moves in the same state are decreased proportionally to their probability of being played.

Nested Rollout Policy Adaptation

- Each move is associated to a weight w_i
- During a playout each move is played with a probability:

$$\exp(w_i) / \Sigma_k \exp(w_k)$$

Nested Rollout Policy Adaptation

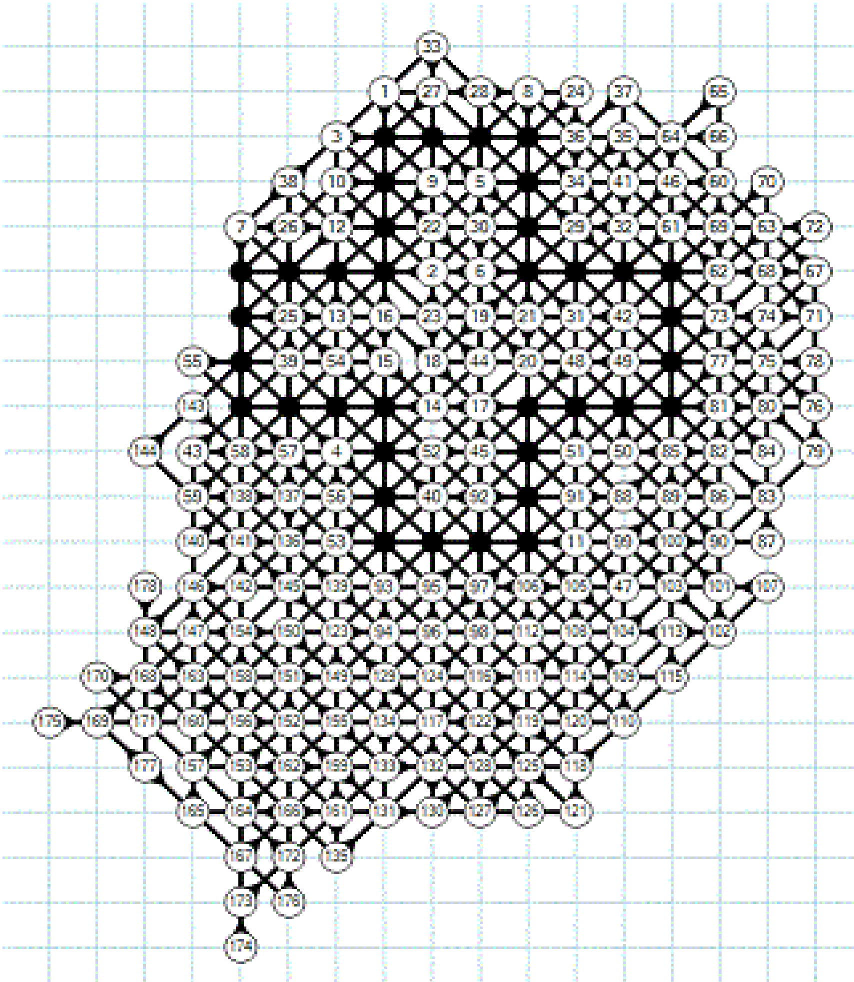
- For each move of the best sequence:

  $w_i = w_i + 1$

- For each possible move of each state of the best sequence:

  $w_j = w_j - \exp(w_j) / \Sigma_k \exp(w_k)$

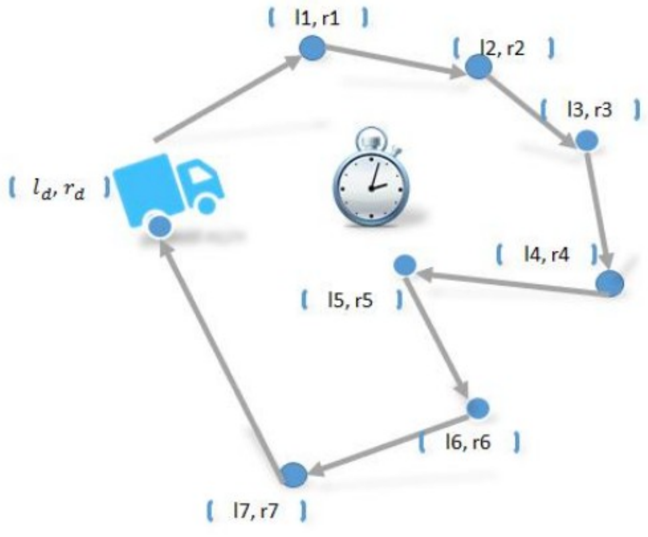
Morpion Solitaire



World record [Rosin 2011]

Applications of NRPA

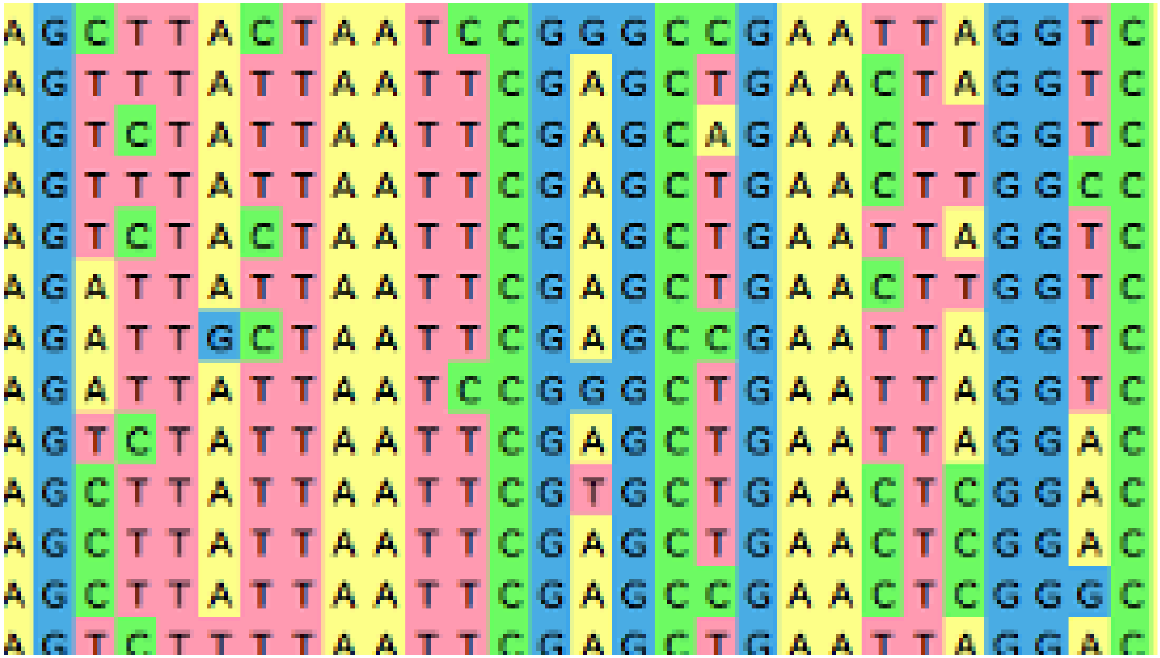
- Traveling Salesman Problem with Time Windows [Cazenave 2012].



- Physical traveling salesman problem.

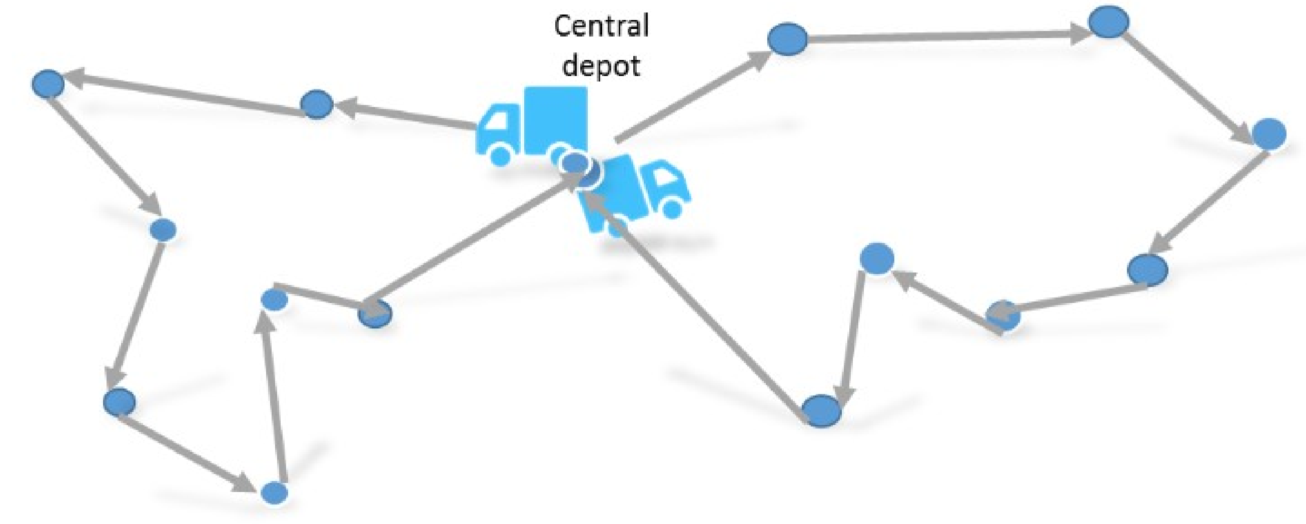
Applications of NRPA

- Improvement of some alignments for Multiple Sequence Alignment [Edelkamp & al. 2015].



Applications of NRPA

- State of the art results for Logistics [Edelkamp & al. 2016].

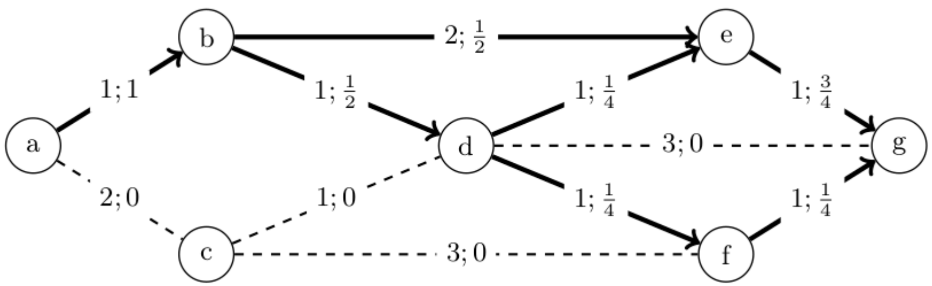


EDF Agents

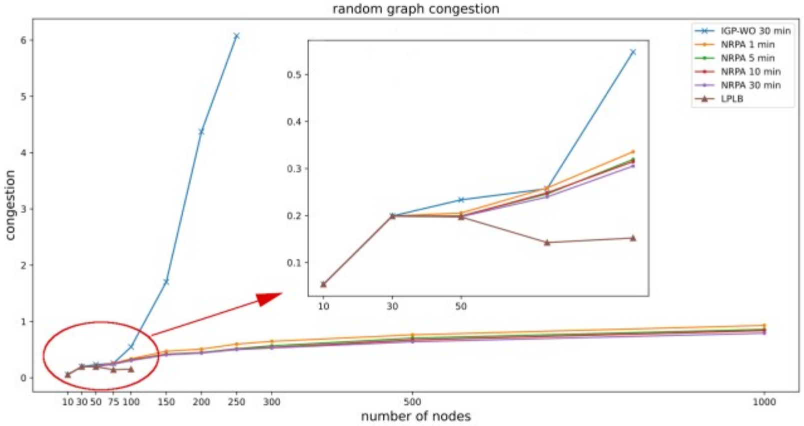
- EDF fleet of vehicles is one of the largest.
- They plan lot of visits every day.
- Monte Carlo Search is 5% better than the specialized algorithms they use [Cazenave & al. 2021].
- Millions of kilometers saved each year.
- Hundreds of tons of CO_2 saved each year.

Network Traffic Engineering

- Provide routing configurations in networks that:
  - Miminize ressources
  - Preserve QoS.

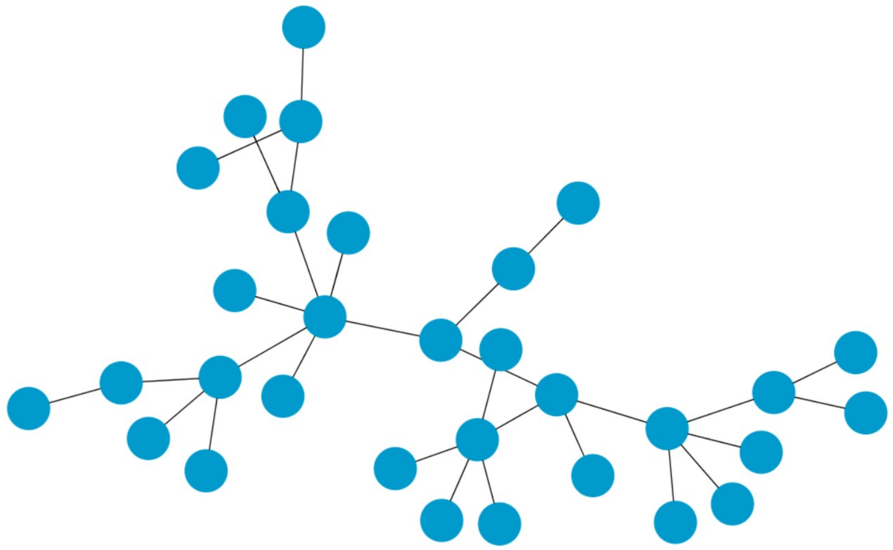


- Better than local search [Dang et al. 2021]:



Refutation of Spectral Graph Theory Conjectures

- Conjecture 3. Collins. Given a tree T, CPA(T) form an unimodal sequence and its peak pA(T) is at the same place as pD(T).



- Better than Deep RL [Roucairol et Cazenave 2022]

Generalized Nested Rollout Policy Adaptation

We propose to generalize the NRPA algorithm by generalizing the way the probability is calculated using a temperature τ and a bias β_{ij}:

$$p_{ik} = \frac{e^{\frac{w_{ik}}{\tau} + \beta_{ik}}}{\Sigma_j e^{\frac{w_{ij}}{\tau} + \beta_{ij}}}$$

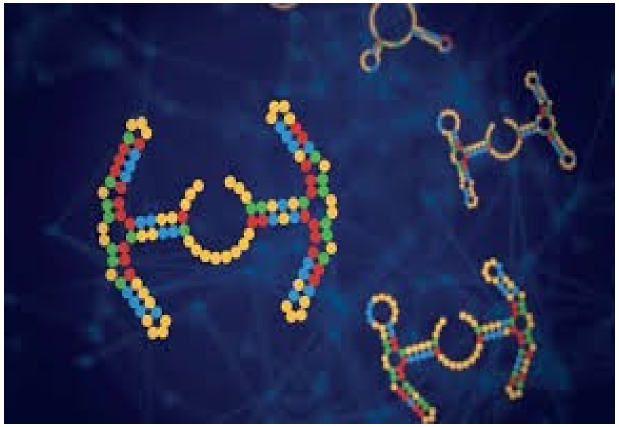
TSPTW

Table 2: Results for the TSPTW rc204.1 problem

Time	NRPA	GNRPA.beta	GNRPA.beta.t.1.4	GNRPA.beta.t.1.4.opt
40.96	-3745986.46 (245766.53)	-897.60 (1.32)	-892.89 (0.96)	-892.17 (1.04)
81.92	-1750959.11 (243210.68)	-891.04 (1.05)	-886.97 (0.87)	-886.52 (0.83)
163.84	-1030946.86 (212092.35)	-888.44 (0.98)	-883.87 (0.71)	-884.07 (0.70)
327.68	-285933.63 (108975.99)	-883.61 (0.63)	-880.76 (0.40)	-880.83 (0.32)
655.36	-45918.97 (38203.97)	-880.42 (0.30)	-879.35 (0.16)	-879.45 (0.17)

Eterna 100

- Find a RNA sequence that has a given folding



- 95/100 sequences found [Cazenave et al. 2020]

Nested Rollout Policy Adaptation

- Morpion Solitaire [Rosin 2011]
- CrossWords [Rosin 2011]
- Traveling Salesman Problem with Time Windows [Cazenave et al. 2012]
- 3D Packing with Object Orientation [Edelkamp et al. 2014]
- Multiple Sequence Alignment [Edelkamp et al. 2015]
- SameGame [Cazenave et al. 2016]
- Vehicle Routing Problems [Edelkamp et al. 2016, Cazenave et al. 2020]
- Graph Coloring [Cazenave et al. 2020]
- RNA Inverse Folding [Cazenave & Fournier 2020]
- Network Traffic Engineering [Dang & al. 2021]
- Slicing 5G [Elkael et al. 2021]
- Refutation of Spectral Graph Theory Conjectures [Roucairol & Cazenave 2022]
- …

AlphaGo

AlphaGo

Lee Sedol is among the strongest and the most famous 9p Go player :

AlphaGo Lee won 4-1 against Lee Sedol in march 2016.

AlphaGo

Ke Jie was the world champion of Go according to Elo ratings :

AlphaGo Master won 3-0 against Ke Jie in may 2017.

AlphaGo Zero

AlphaGo Zero

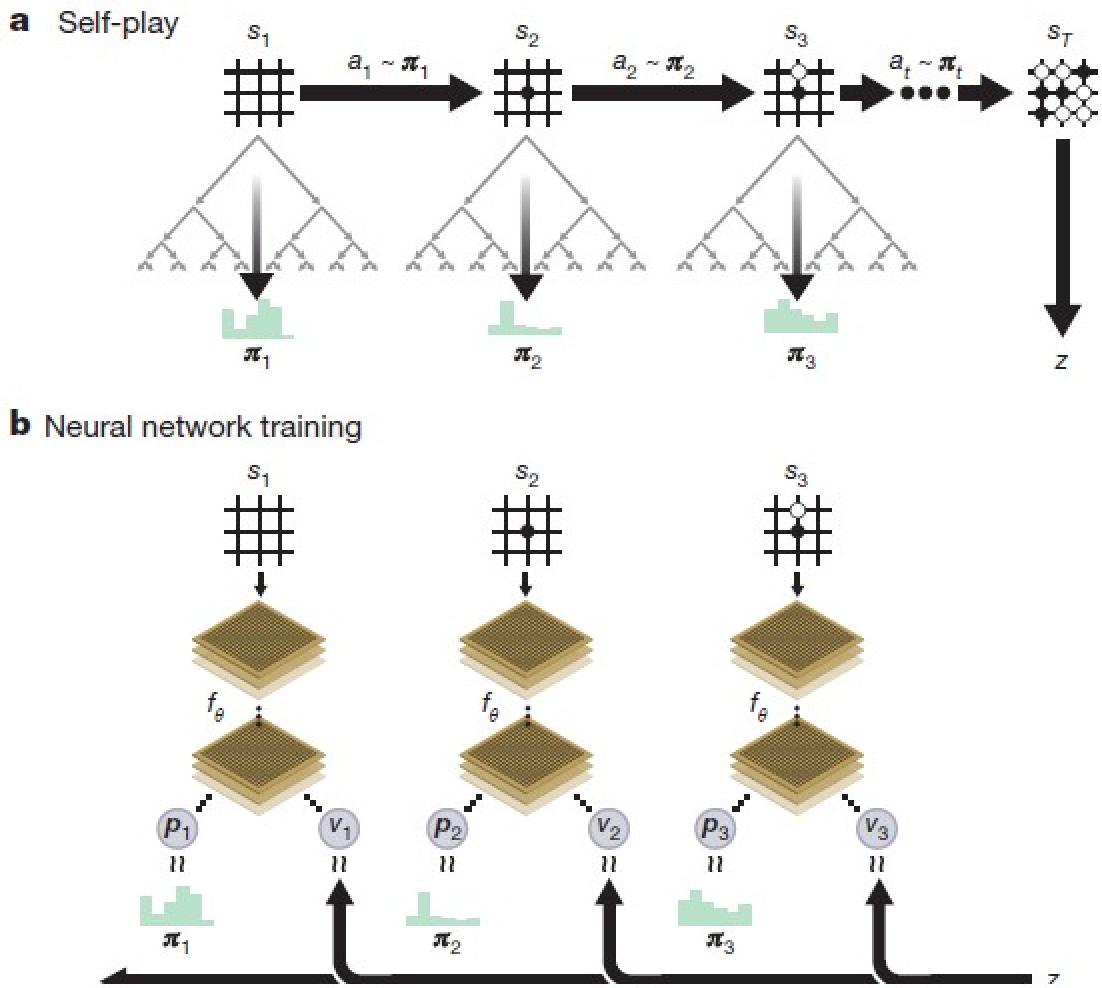
- It plays against itself using PUCT and 1,600 tree descents

$$U(s,a) = c_{\text{puct}} P(s,a) \frac{\sqrt{\sum_b N(s,b)}}{1 + N(s,a)}$$

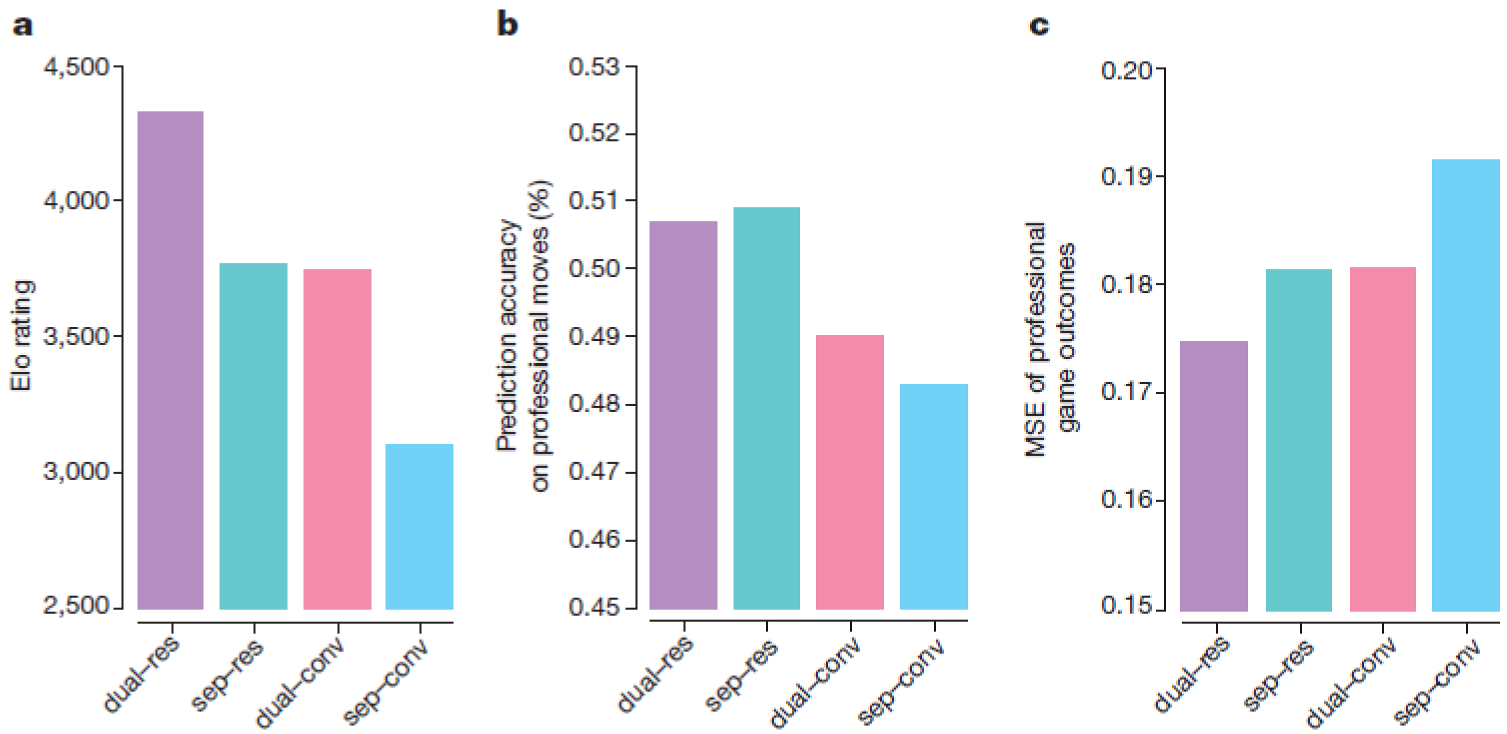
- It uses a residual neural network with two heads.

- One head is the policy, the other head is the value.

AlphaGo Zero



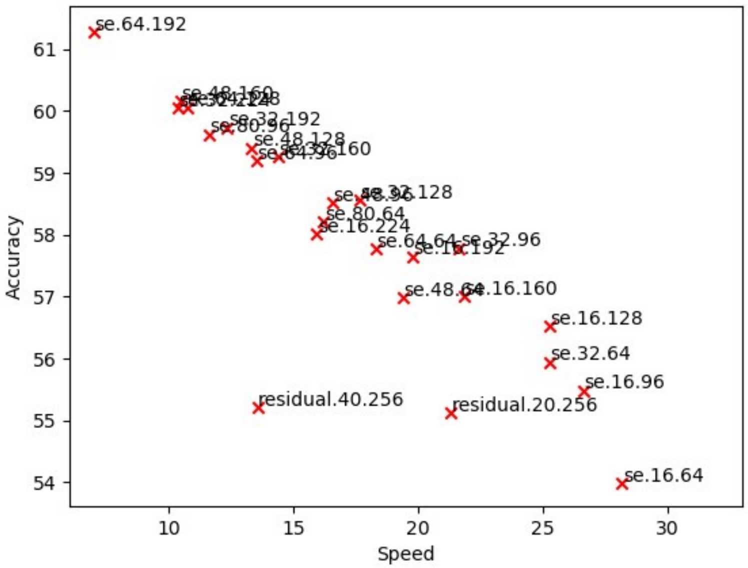
AlphaGo Zero

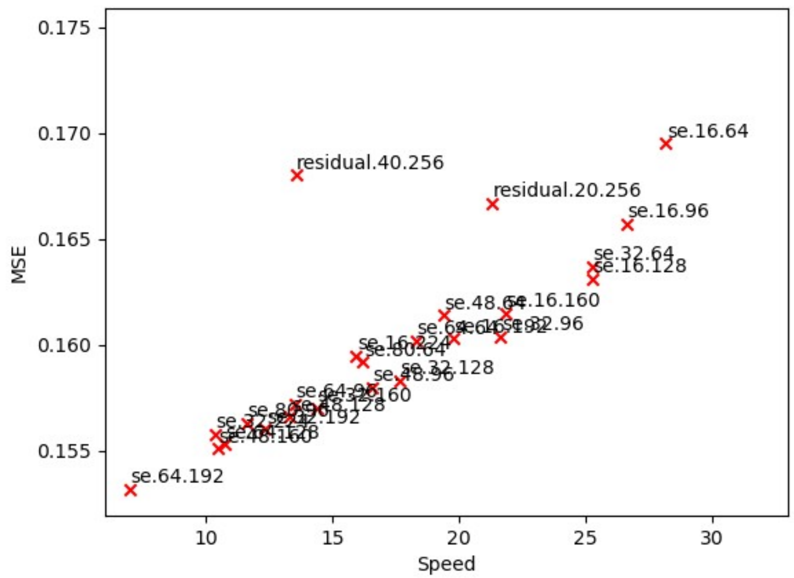


Polygames

- Polygames [Cazenave & al. 2020] is the open source implementation by Facebook FAIR of Alpha Zero.
- It has been applied to many games.
- It uses a fully convolutional policy head.
- It uses average global pooling in the value head.
- It makes it invariant to board size.
- It has beaten the best Havannah player.

Mobile Networks for Computer Go





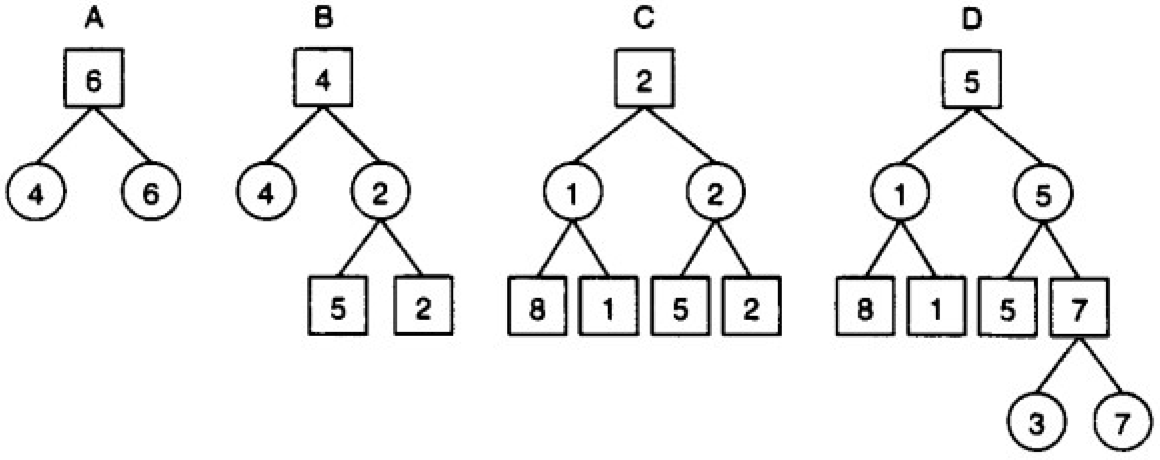
MCTS and Deep RL

Monte Carlo Tree Search and Deep Reinforcement Learning to discover new fast matrix multiplication algorithms:

Athénan

Unbounded Minimax

- Principle = Extend the most promising leaf.
- Asymmetric growing of the search tree.

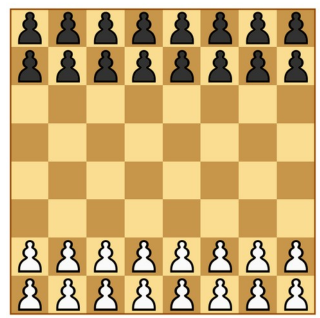


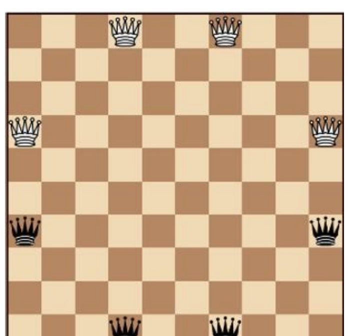
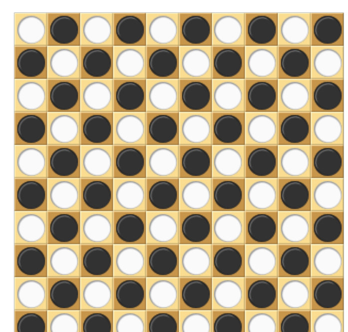
Descent

- Only uses a value network.
- Self play without prior knowledge.
- Learns the scores inside the trees developed by the Unbounded MiniMax.
- Minimax Strikes Back [Cohen-Solal & Cazenave 2023].

Athénan

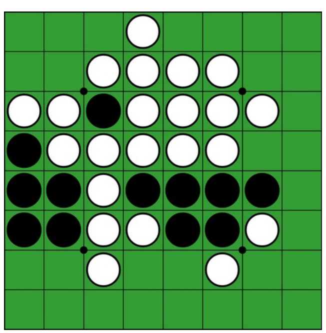
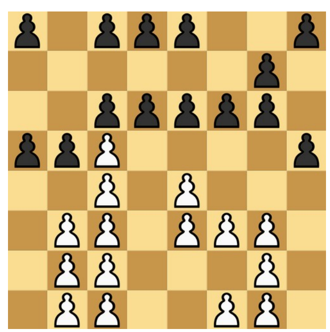
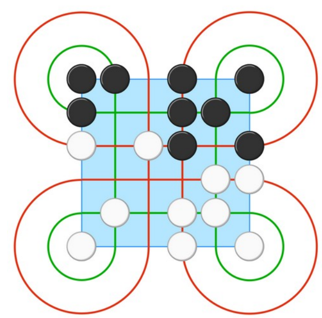
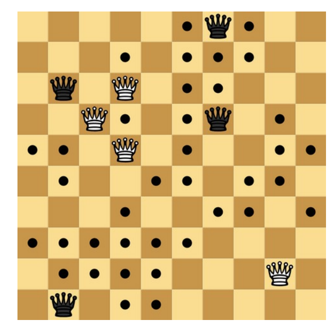
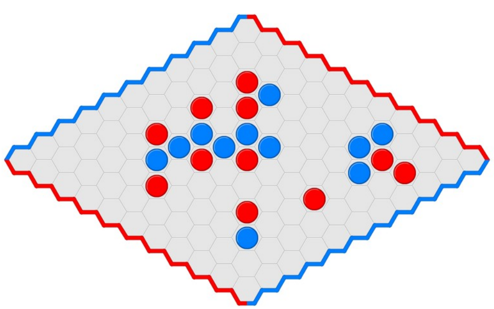
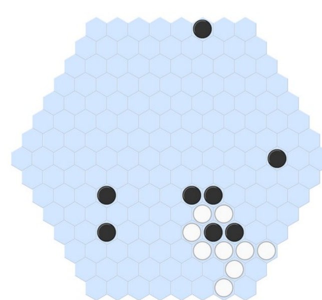
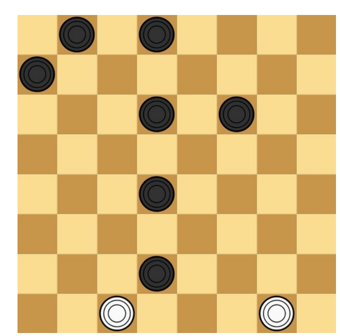
- 5 gold medals at the 2020 Computer Olympiad.
- Othello 10x10, Breakthrough, Surakarta, Amazons, and Clobber.



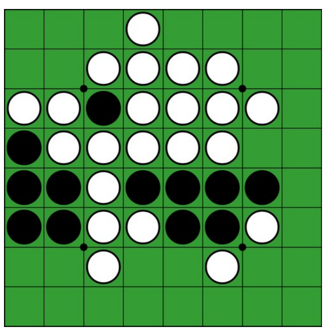
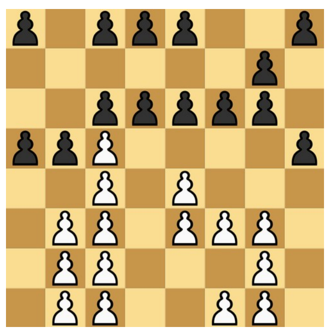
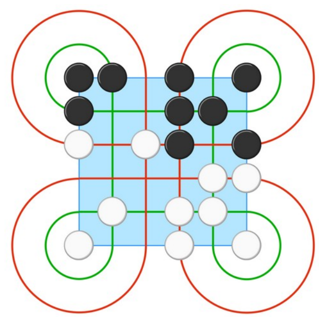
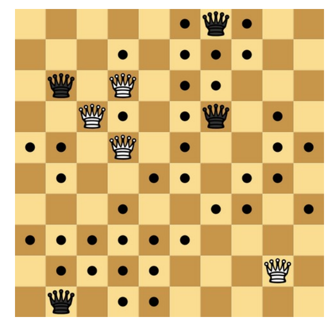
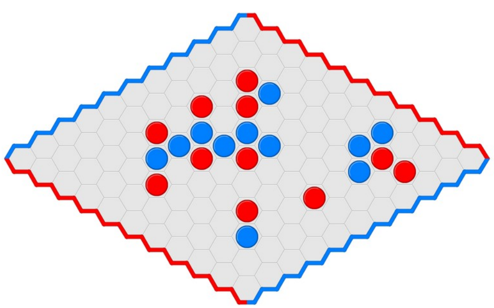
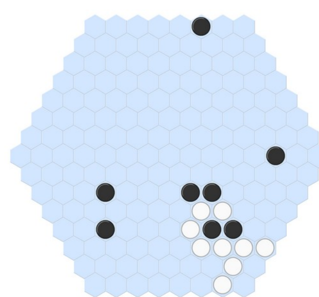
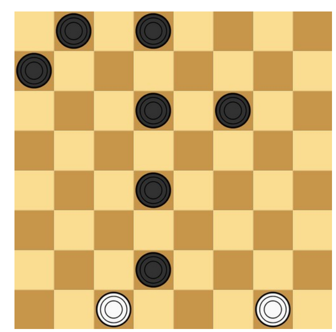
Athénan

- 11 gold medals at the 2021 Computer Olympiad!
- Othello 8x8, Breakthrough, Surakarta, Amazons, Hex 11x11, Hex 13x13, Hex 19x19, Havannah 8x8, Havannah 10x10, Canadian Draughts, Brazilian Draughts.

Athénan

- 16 gold medals at the 2023 Computer Olympiad!
- Amazons, Arimaa, Ataxx, Breakthrough, Canadian Draughts, Chinese Chess, Clobber, Havannah (8×8), Havannah (10×10), Hex (11×11), Hex (13×13), Hex (19×19), Lines of Action, Othello (10×10), Santorini, Surakarta.

AlphaMu

PIMC

For all possible moves

For all possible worlds

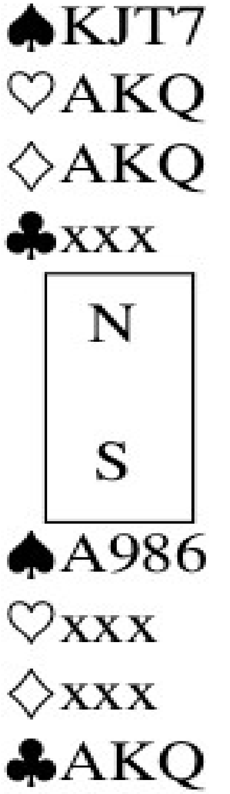
Exactly solve the world

Play the move winning in the most worlds

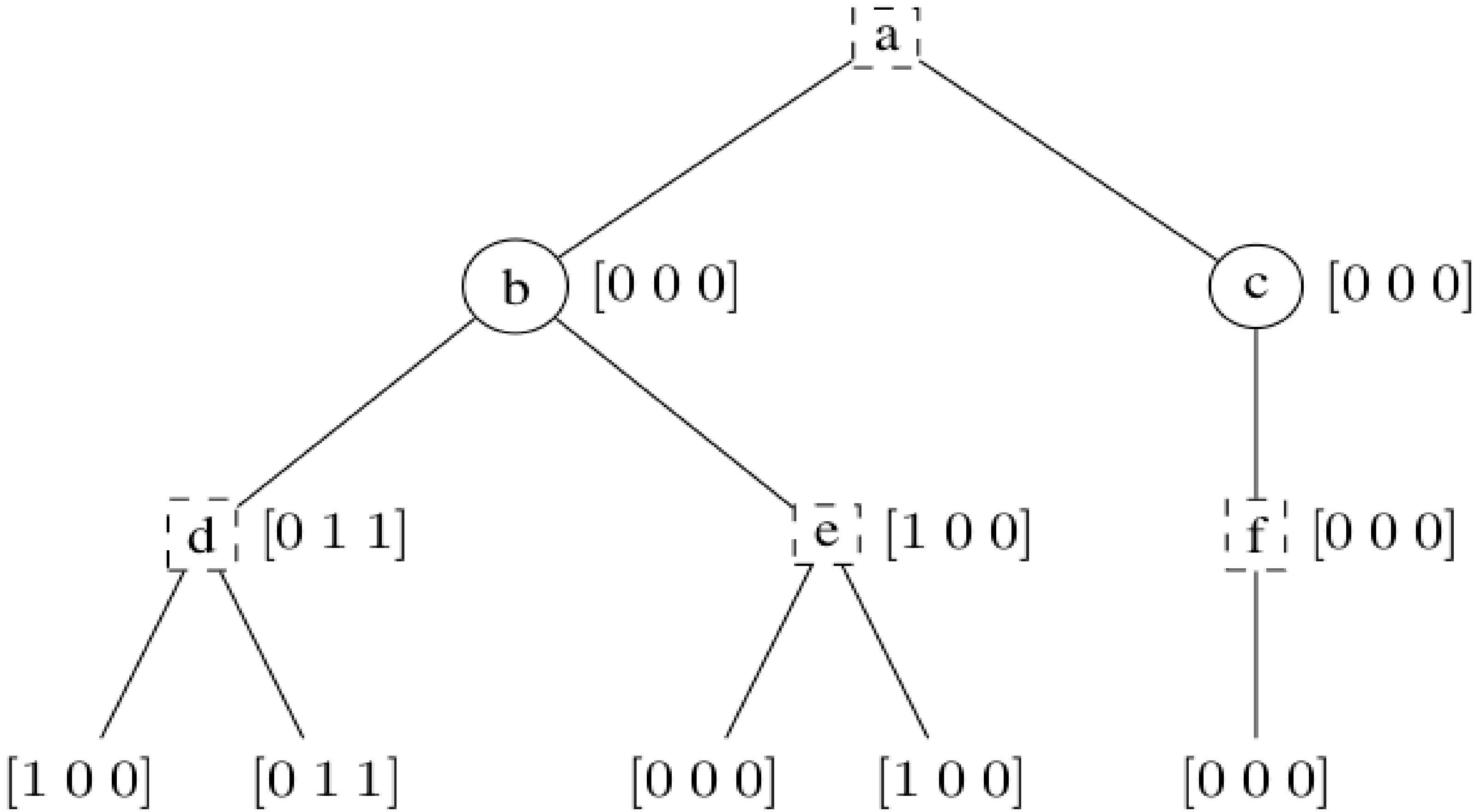
Strategy Fusion

- Problem = PIMC can play different moves in different worlds.

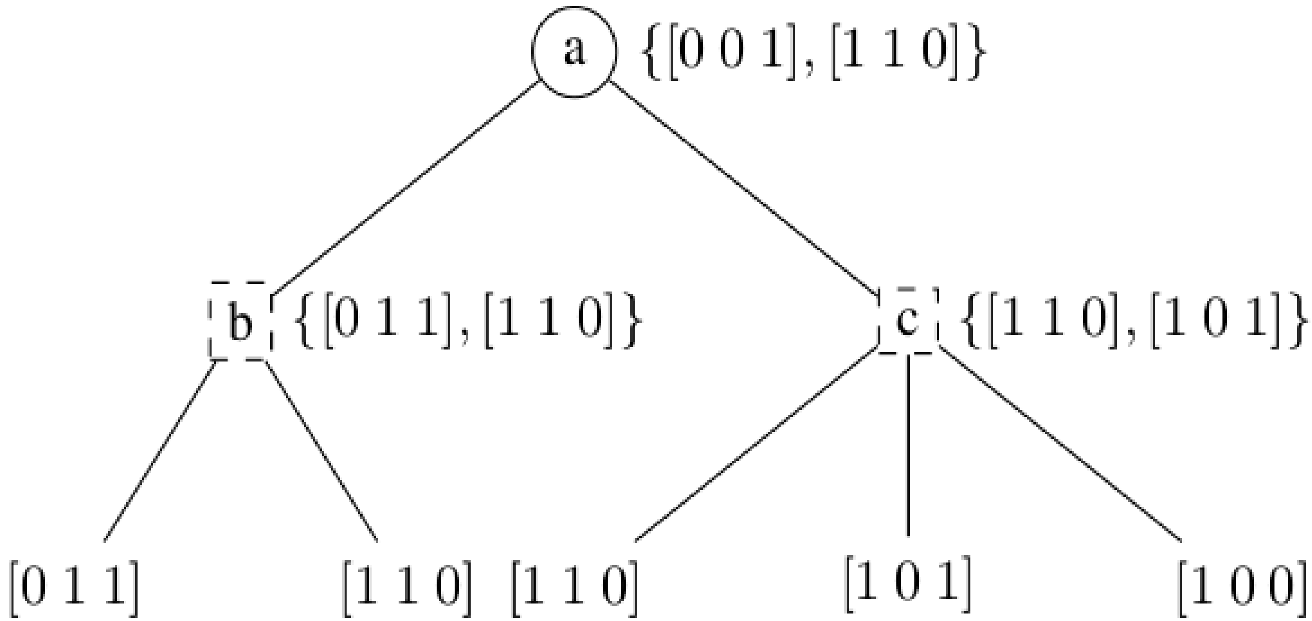
- Whereas the player cannot distinguish between the different worlds.



Non Locality



Product of Pareto Fronts at Min Nodes



- AlphaMu [Cazenave & Ventos 2020].

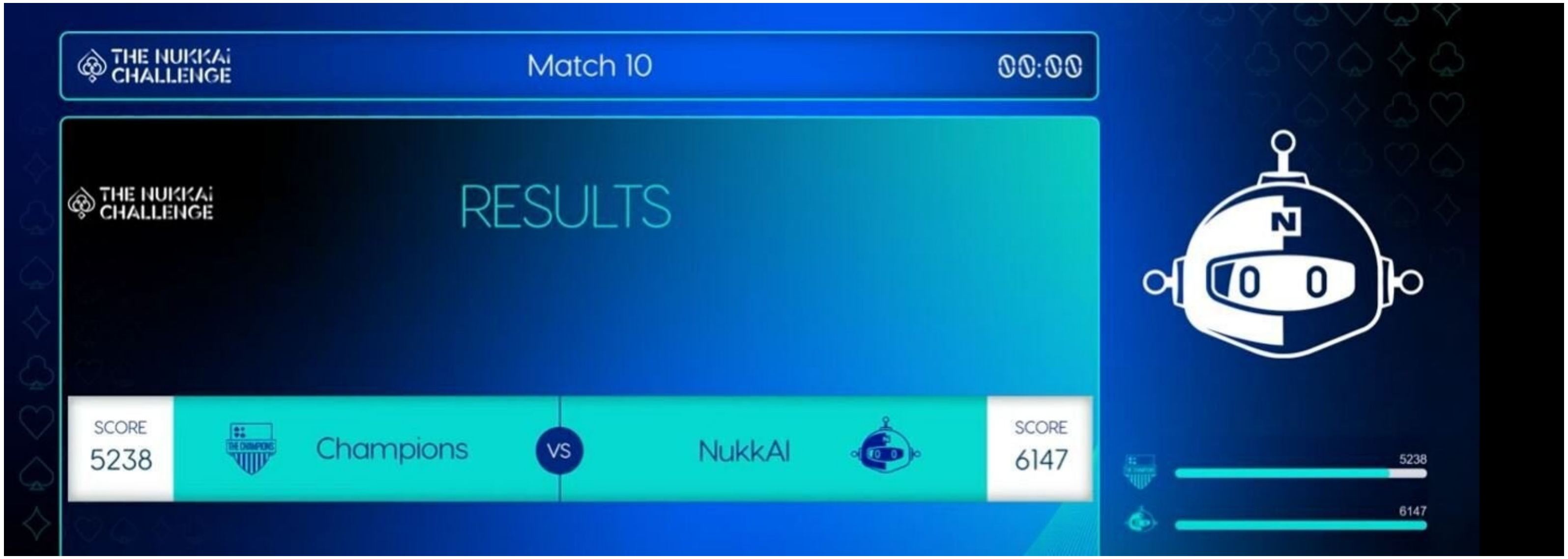
Nook

- Opponent Modeling
- Alpha-Beta on each possible world
- AlphaMu
- Rule based opening lead
- Contract : 1NT 2NT 3NT
- Declarer

Nook


NECTER
IL FAUT FAIRE AUJOURD'HUI
CE QUE TOUT LE MONDE FERA DEMAIN
JEAN COCTEAU
SSANCE
changement
SIMPLICITÉ
bpifrance
OSER
agilité
VOLONTÉ
Le HUB
OPTIMISME
PROXIMITÉ
AMBITION
PARTAGER
DESIGN
ÉCHANGER
DÉCLOISONNER
1820
1816
1879
1824
1884
1890

Nook



THE NUKKAi CHALLENGE
Match 10
00:00
THE NUKKAi CHALLENGE
RESULTS
SCORE
5238
Champions
VS
NukkAI
SCORE
6147
5238
6147

Nook

Print subscriptions | Sign in | Search jobs | Search | International edition

Support the Guardian

Available for everyone, funded by readers

Support us →

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

World UK Coronavirus Climate crisis Environment Science Global development Football **Tech** Business Obituaries

Artificial intelligence (AI)

This article is more than **7 months old**

Artificial intelligence beats eight world champions at bridge

Victory marks milestone for AI as bridge requires more human skills than other strategy games

Laura Spinney

Tue 29 Mar 2022 06.00 BST

The AI, NooK, was able to read its opponents and explain its decision-making. Photograph: switas/Getty Images/iStockphoto

An artificial intelligence has beaten eight world champions at bridge, a game in which human supremacy has resisted the march of the machines until now.

The victory represents a new milestone for AI because in bridge players work with incomplete information and must react to the behaviour of several other players - a scenario far closer to human decision-making.

In contrast, chess and Go - in both of which AIs have already beaten human champions - a player has a single opponent at a time and both are in possession of all the information.

Conclusion

- Monte Carlo Tree Search
- Nested Monte Carlo Search
- Nested Rollout Policy Adaptation
- Alpha Zero and Deep Reinforcement Learning
- Athénan = Unbounded Minimax and self play learning of the evaluation
- AlphaMu = Planning in Bridge