LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Praktikum Autonome Systeme

Applications II

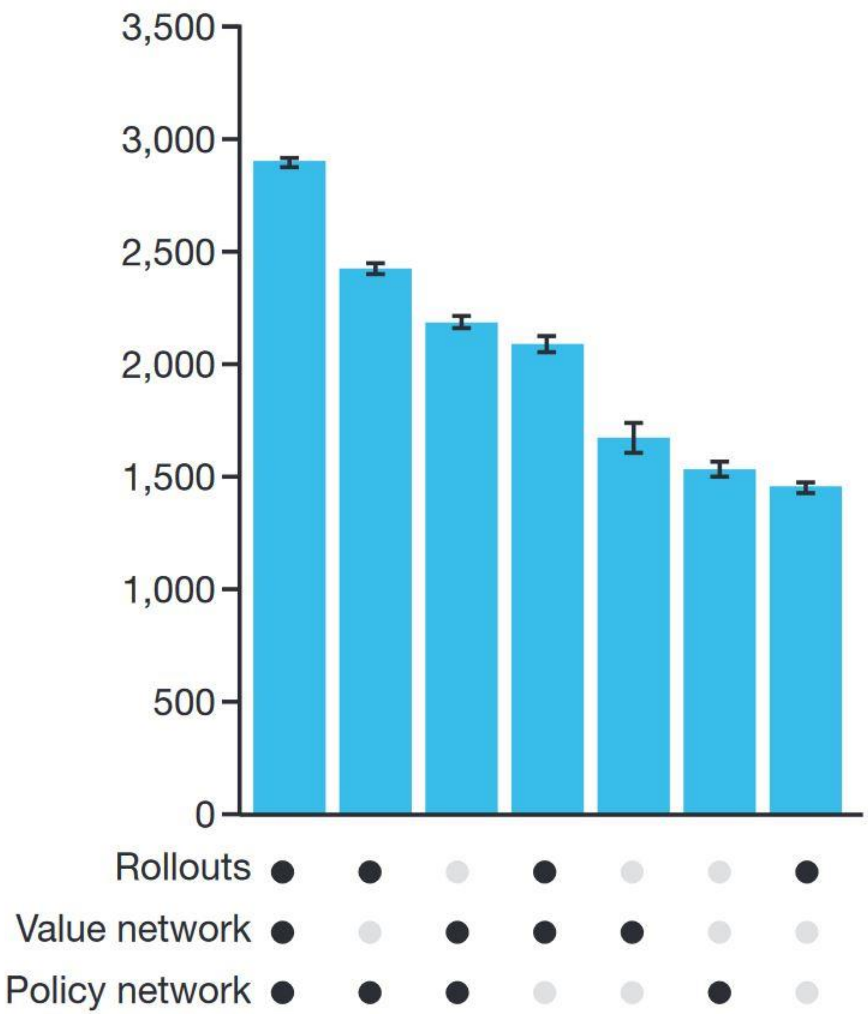
Prof. Dr. Claudia Linnhoff-Popien
Thomy Phan, Andreas Sedlmeier, Fabian Ritz
http://www.mobile.ifi.lmu.de

WiSe 2020/21

→ **Recap**

Recap: AlphaGo (2016)

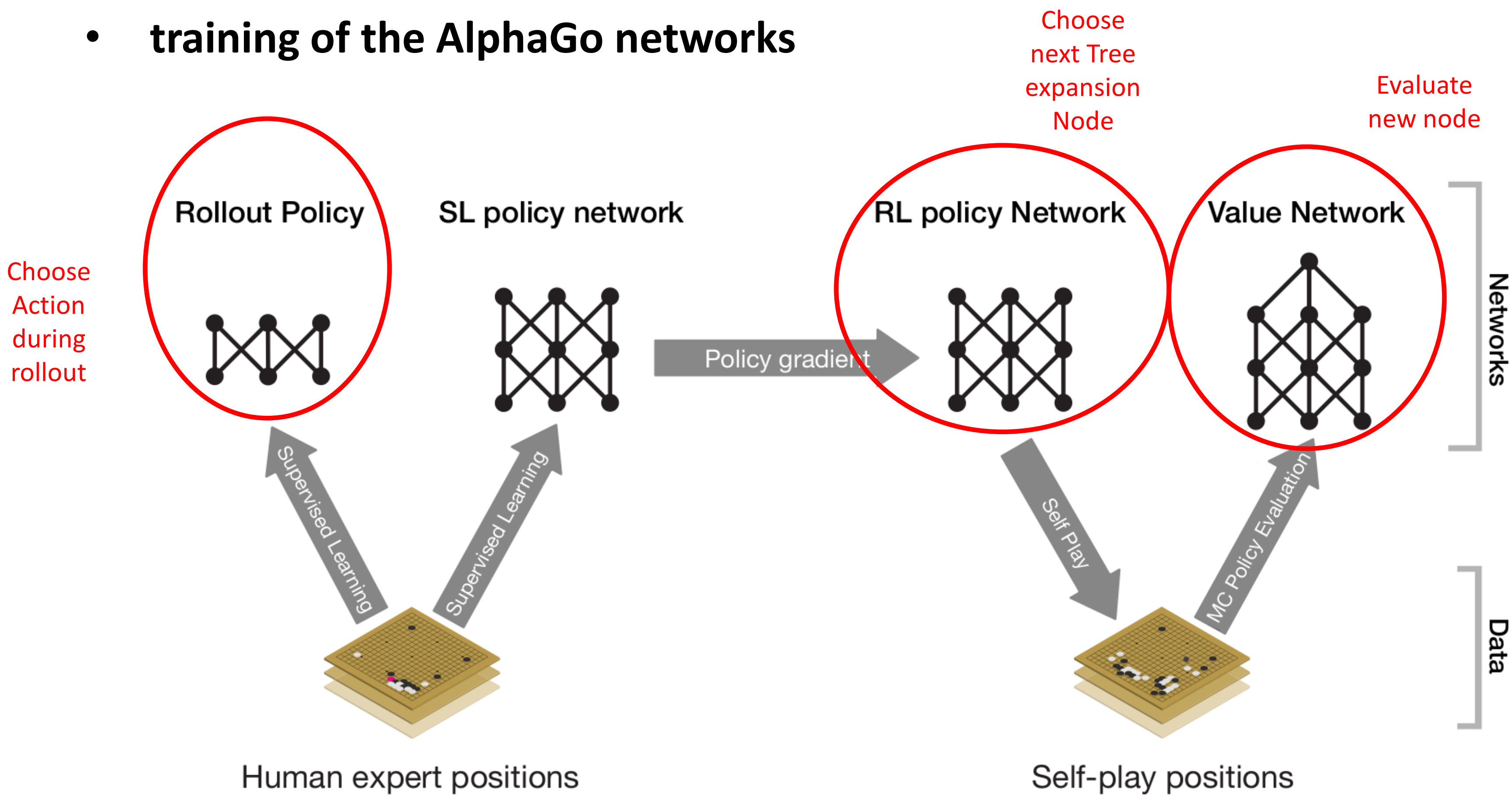
- **combines**
  - Fast Rollout Policy (rollouts)
  - Deep Reinforcement Learning (Value network)
  - Deep Supervised Learning (Policy network)
  - MCTS (combines above)



https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf

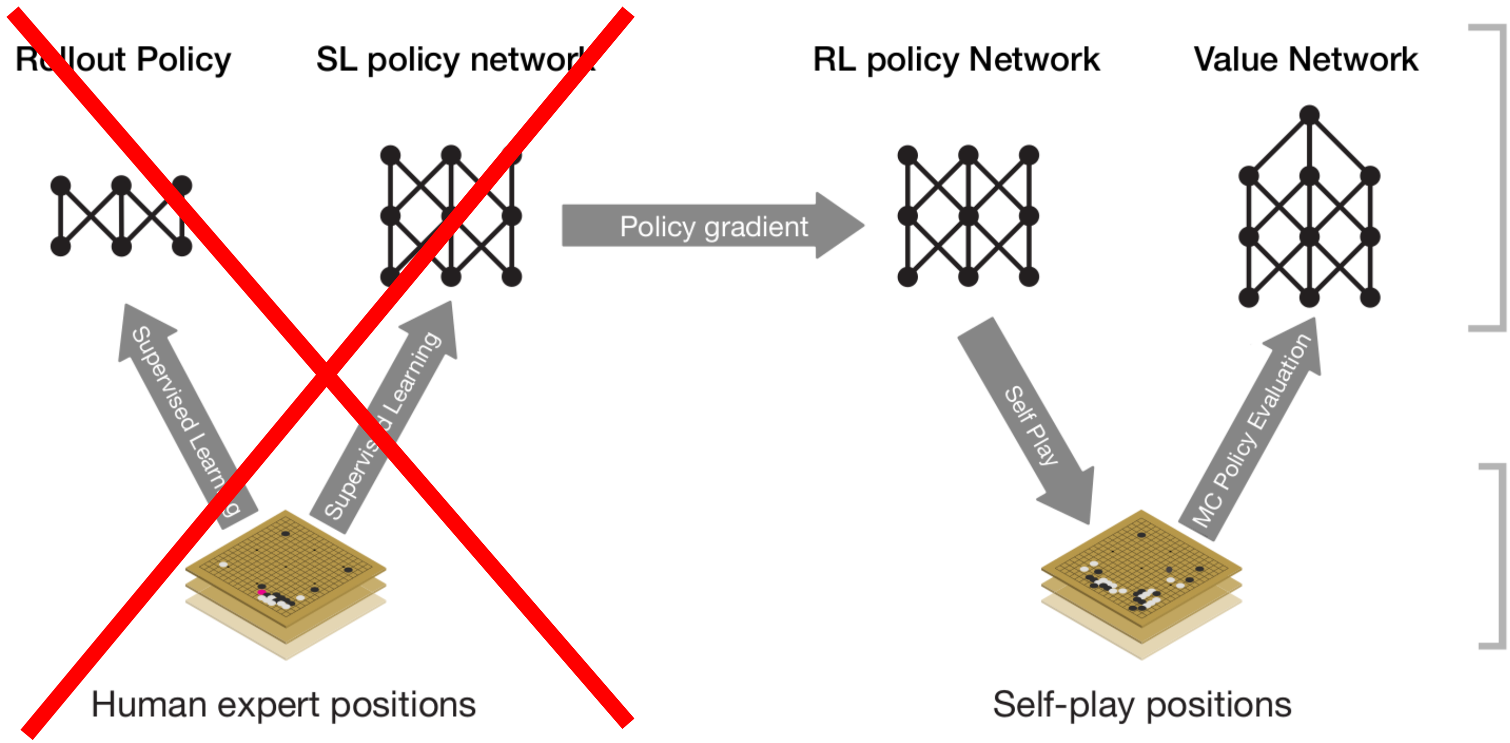
Recap: AlphaGo (2016)

- **training of the AlphaGo networks**



https://www.nature.com/articles/nature16961

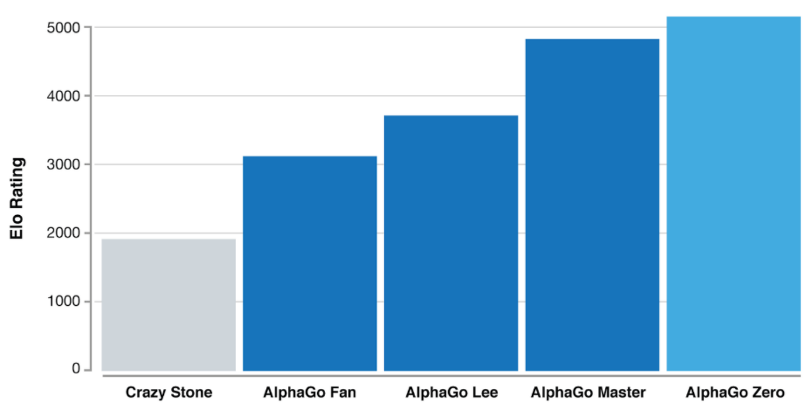
Recap: AlphaGo Zero (2017)

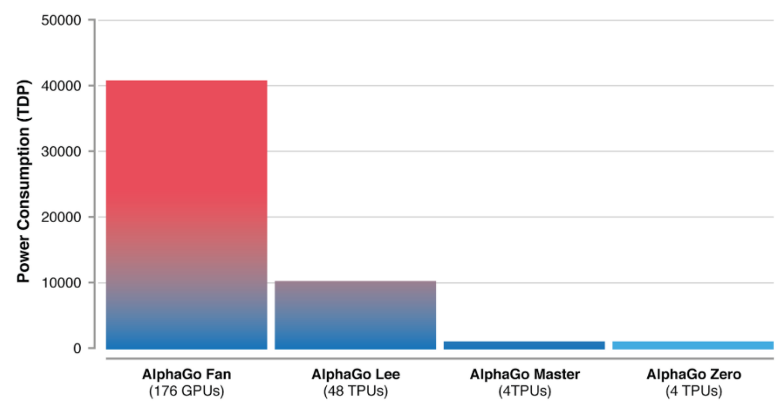


https://deepmind.com/documents/119/agz_unformatted_nature.pdf

Recap: AlphaGo Zero (2017)

- **differences with AlphaGo**
  - only stones from the Go board as input
  - only one neural network with two heads (no rollouts)
  - only Self-Play Reinforcement Learning
  - MCTS during RL Self-Play (training)





https://deepmind.com/blog/article/alphago-zero-starting-scratch

→ Alpha Zero

Shedding new light on the grand games of chess, shogi and Go

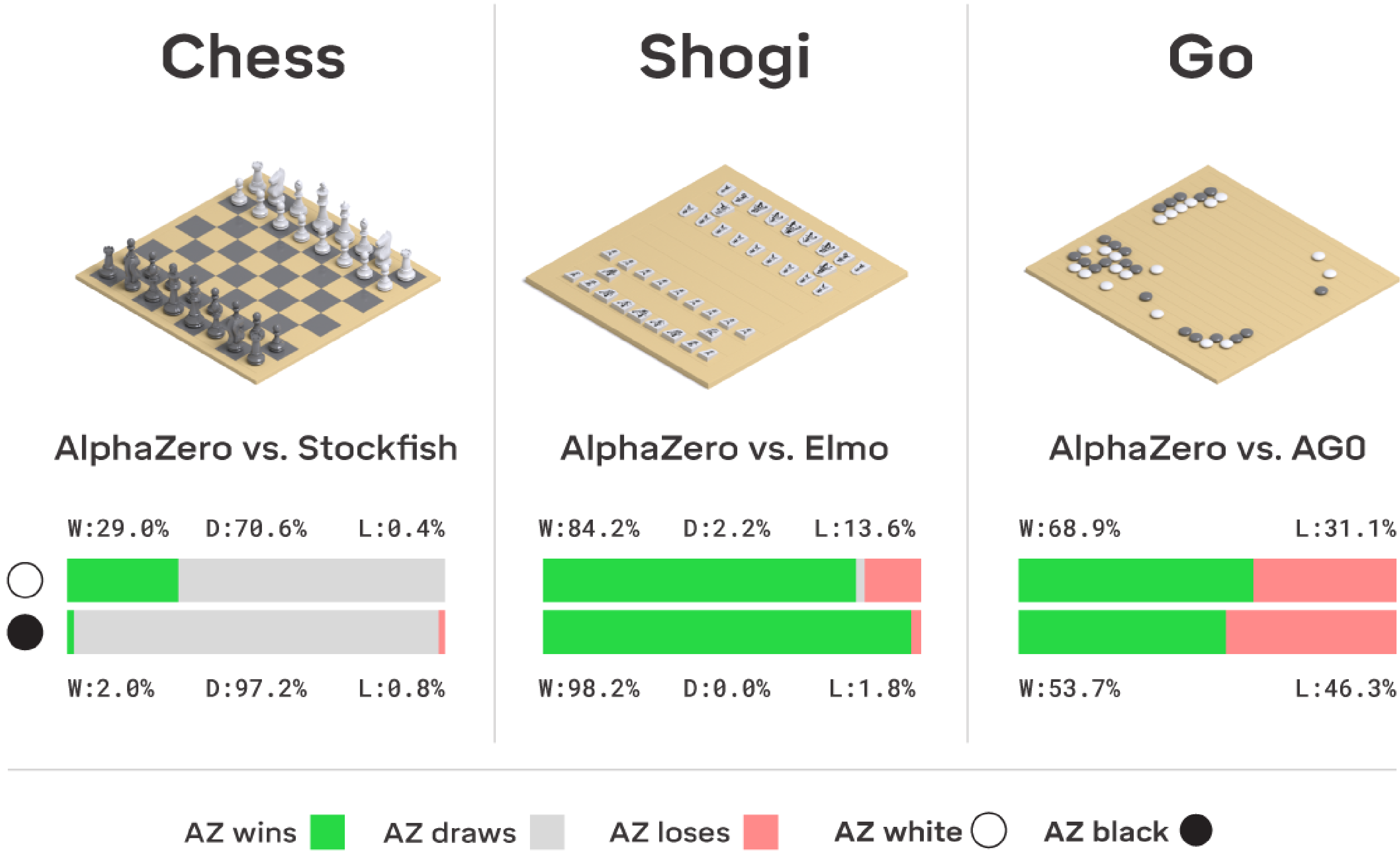
https://www.doi.org/10.1126/science.aar6404

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

AlphaZero – Shedding new light… (2018)

https://www.youtube.com/watch?v=7L2sUGcOgh0

AlphaZero (2018)



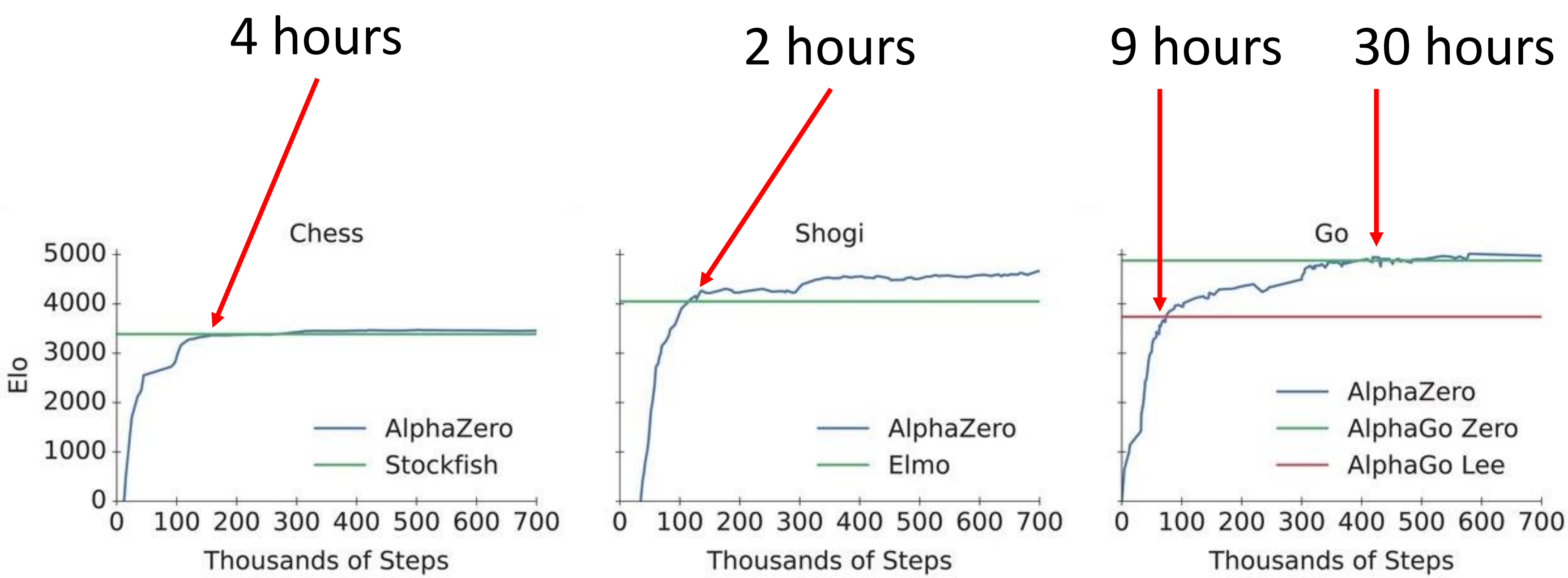
https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

AlphaZero (2018)

- **further differences to AlphaGo Zero:**
  - estimates and optimizes the expected outcome (instead of the probability of winning)
  - no augmentation of training data (instead of transforming the board position during MCTS)
  - neural network is updated continually (instead of waiting for an iteration to complete)
  - self-play games always generated by the latest parameters for the neural network

AlphaZero (2018)

- **training time to beat the State of the Art:**



take-away: MCTS and RL go well together

https://science.sciencemag.org/content/362/6419/1140

Game Theory: Chess, Shogi and Go are

- finite
- two-player zero-sum games with
- perfect information and
- no stochasticy

→ How about other ~~problems~~ games?

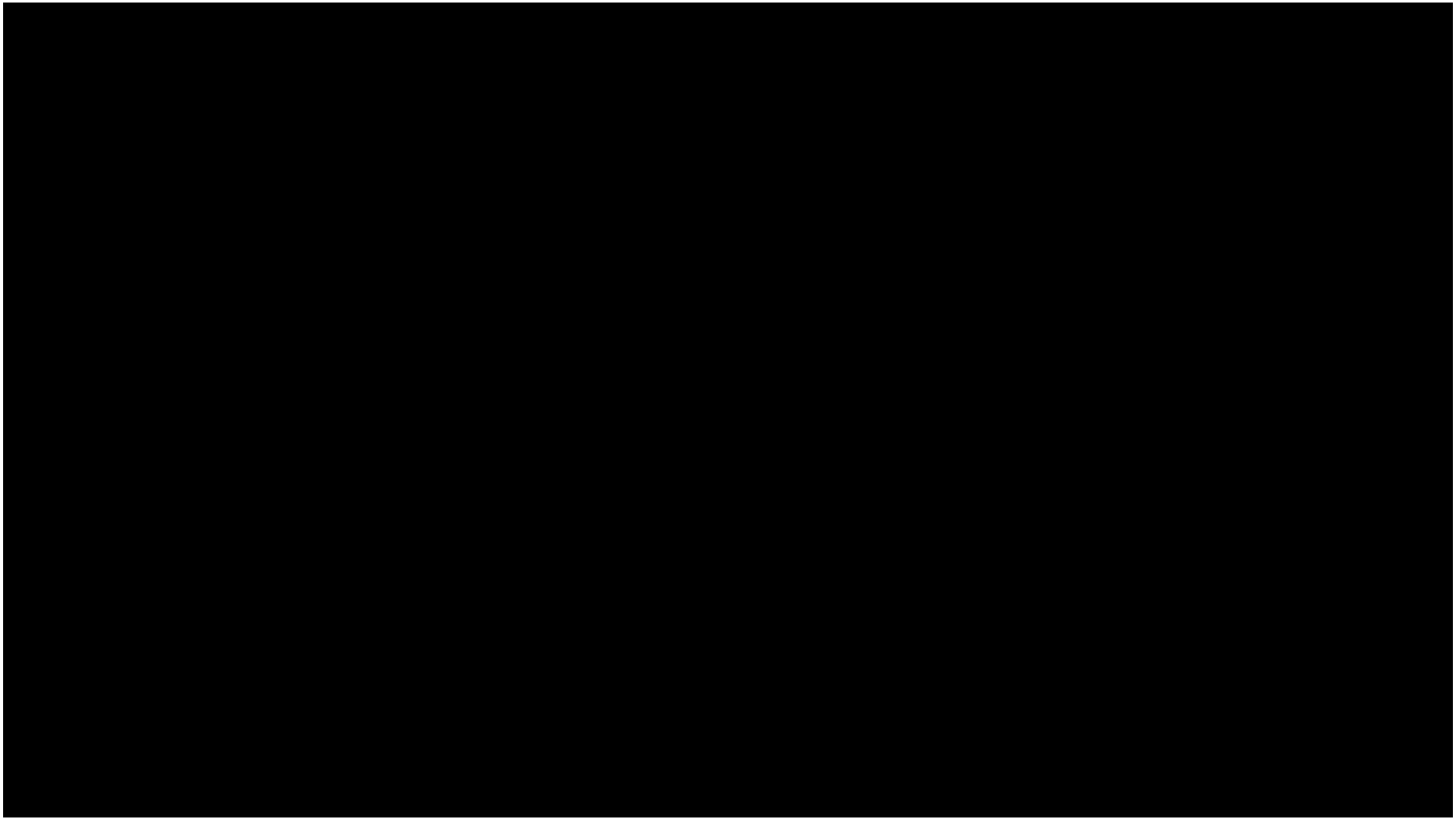
→FTW: the emergence of complex cooperative agents

Human-level performance in first-person multiplayer games with population-based deep reinforcement learning

https://www.doi.org/10.1126/science.aau6249

https://deepmind.com/blog/article/capture-the-flag-science

Human-level performance in first-person… (2018)



https://www.youtube.com/watch?v=dltN4MxV1RI

FTW (2018)

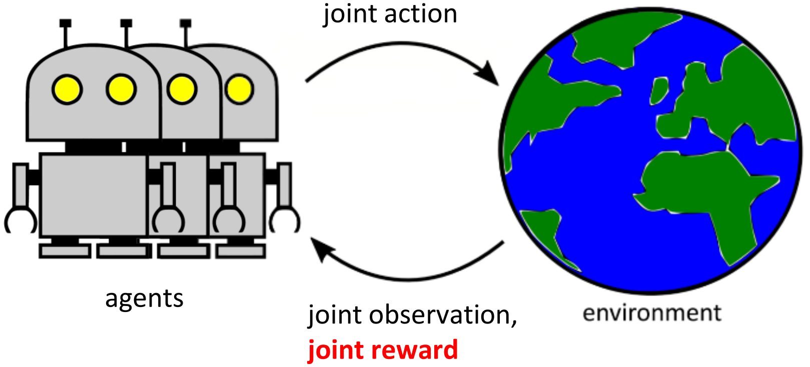
- **agents must learn how to**
  - see
  - act
  - cooperate
  - compete

  in unseen environments

- **the environment provides a single reward signal per match**
  - whether a team won or not

Agents? Cooperation? Teams?
We‘ve only talked about single-agent scenarios so far!

Exkursion: Multi Agent Scenarios



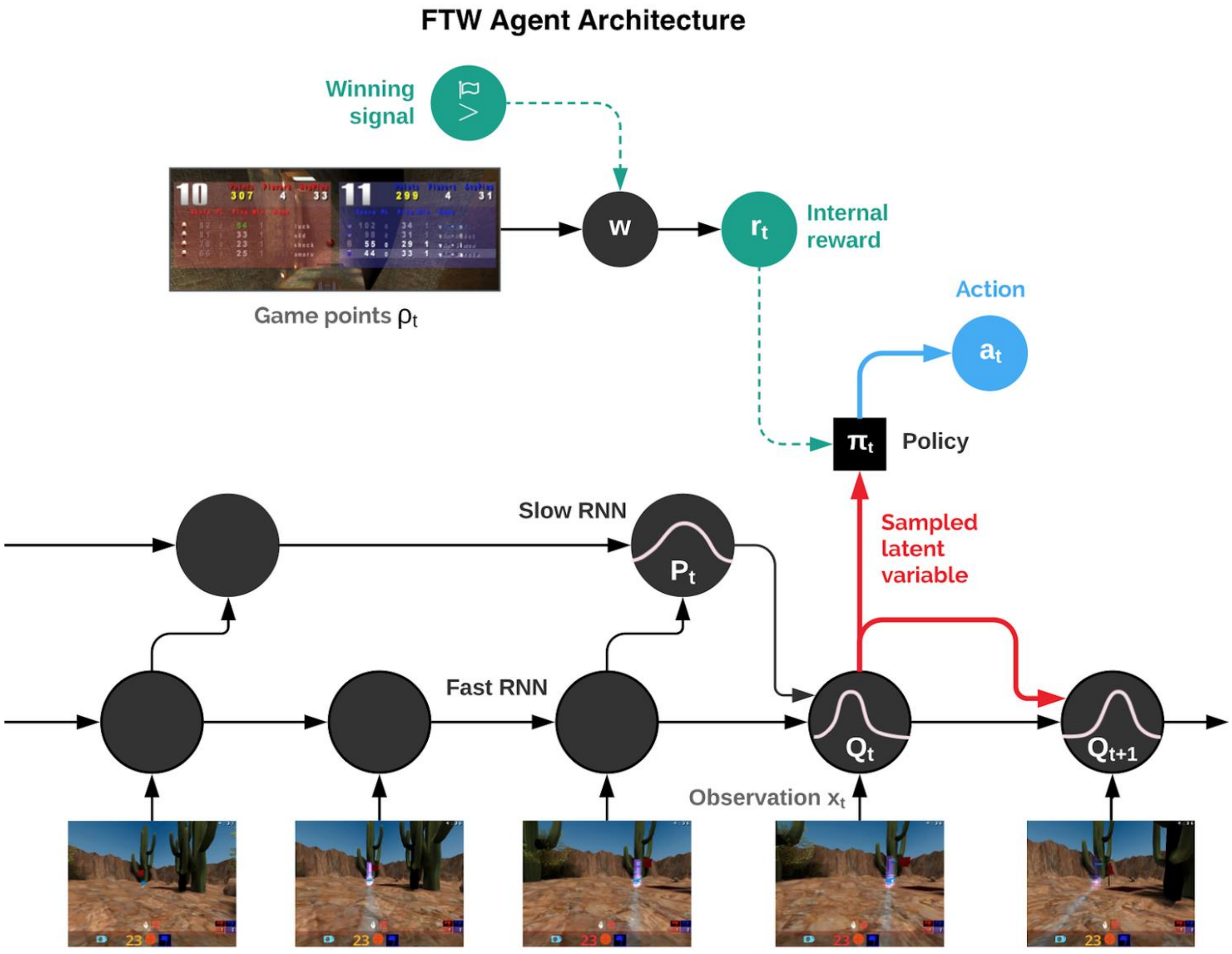
- **many open questions**
  - uncertainty w.r.t. other agents
  - non-stationarity
  - partial observability
  - credit assignment problem

exploding
state-/action space

FTW (2018)

- **FTW agent(s)**
  - population based approach
  - trained by self-play
  - internal (dense) and external (sparse) rewards
- **two-tier process**
  - learn RL policies on individual, internal rewards
  - optimize internal rewards w.r.t. the global goal (winning)
- **two-tier timescale** (combination of a fast and a slow RNN)
  - improves the memory usage
  - generates consistent action sequences

FTW (2018)

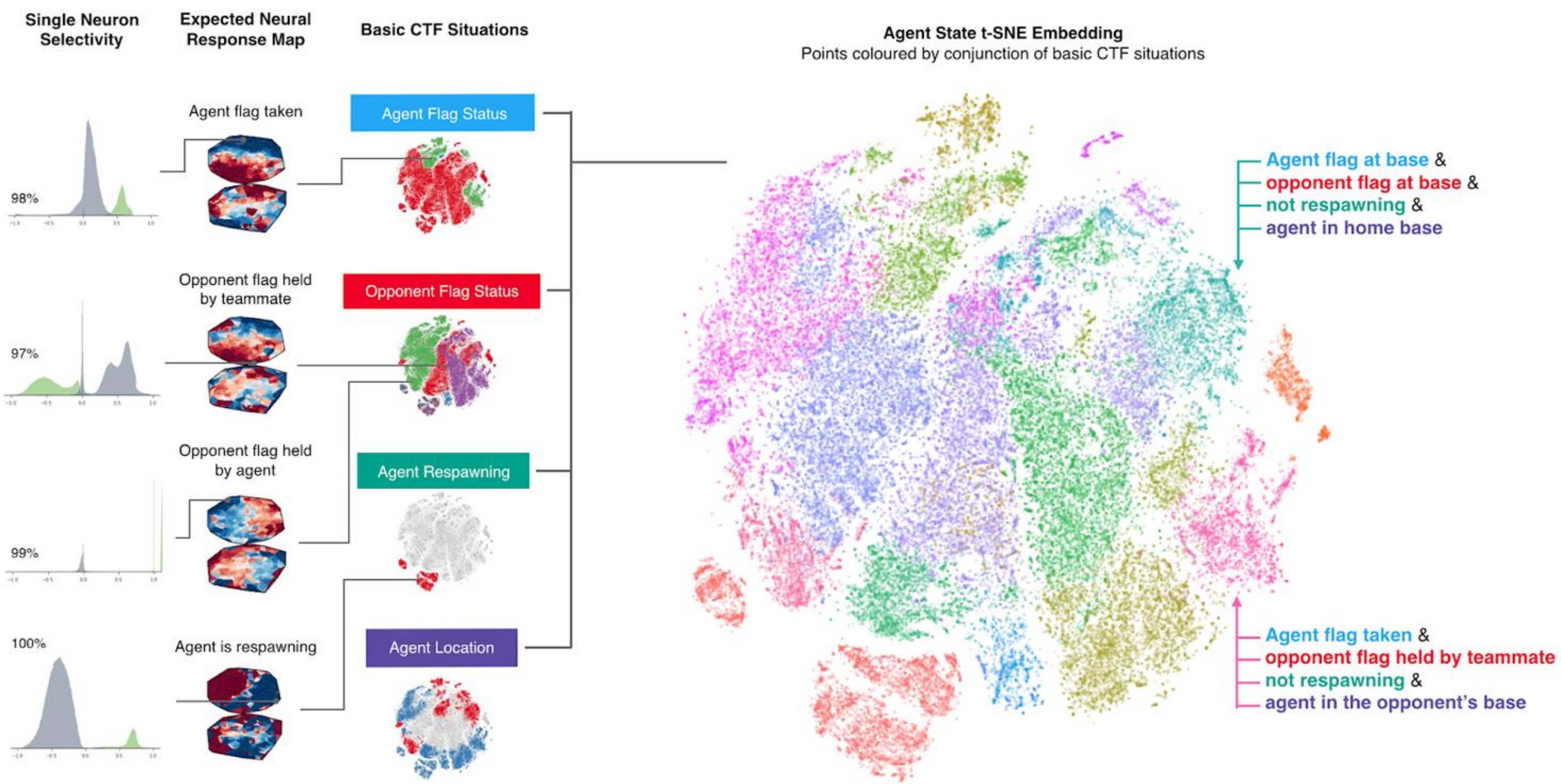


https://deepmind.com/blog/article/capture-the-flag-science

FTW (2018)

- **situational awareness**
  - the agent’s room
  - status of the flags
  - visible teammates and opponents
- **advanced strategies**
  - home base defence
  - opponent base camping
  - teammate following

FTW (2018)

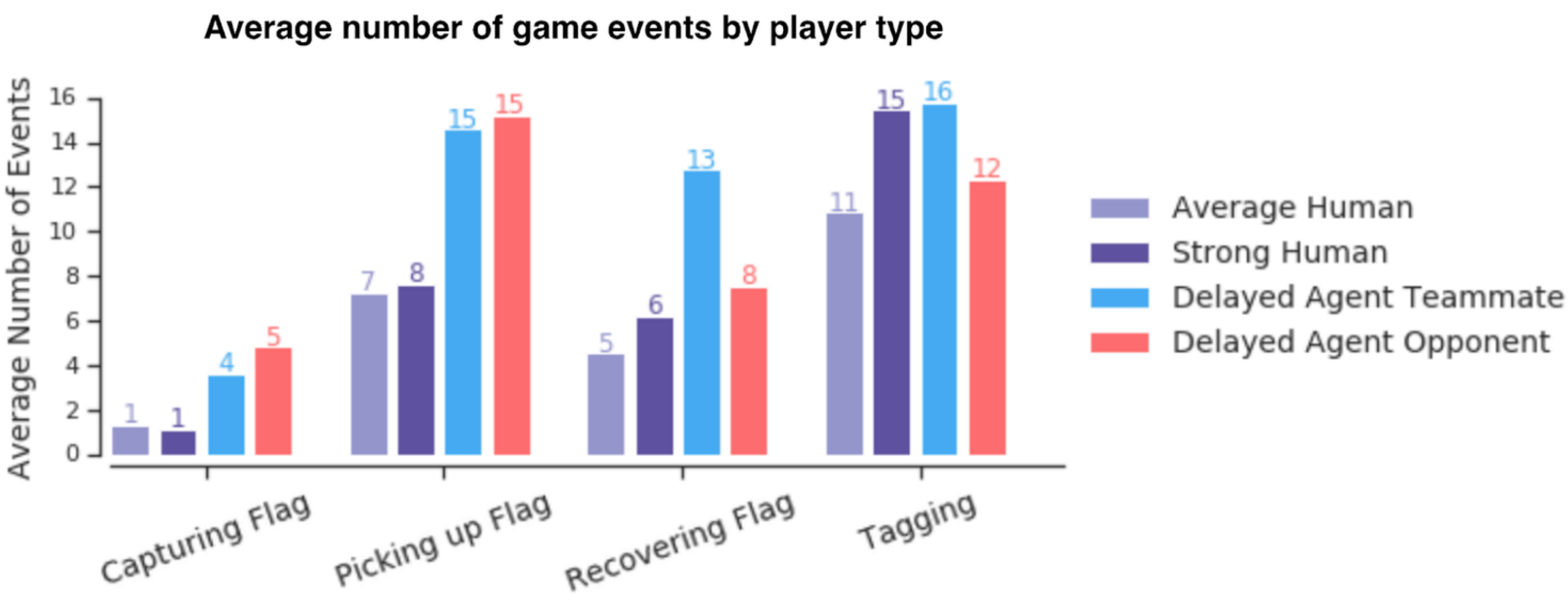


https://deepmind.com/blog/article/capture-the-flag-science

FTW (2018)

267ms response-delayed agent results

Human Game Type	Human Win Rate
Exploitability Trail Games Tester	30%
Strong Human Tournament Participant	21%
Intermediate Human Tournament Participant	12%



https://deepmind.com/blog/article/capture-the-flag-science

FTW (2018)

- **human comparable performance**
  - super-human response time
  - super-human accuracy
  - more collaborative than humans

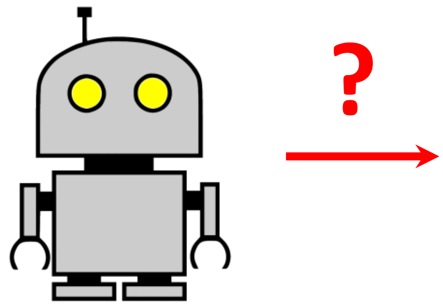
take-away: individual shaped rewards may boost training

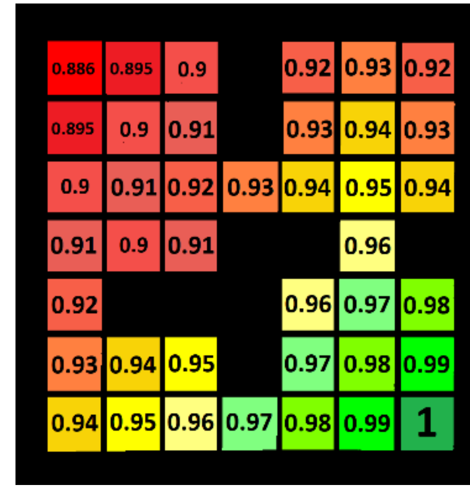
→ Excursion: Reward Shaping

Policy invariance under reward transformations: Theory and application to reward shaping

https://people.eecs.berkeley.edu/~russell/papers/icml99-shaping.pdf

Potential Based Reward Shaping (PBRS)

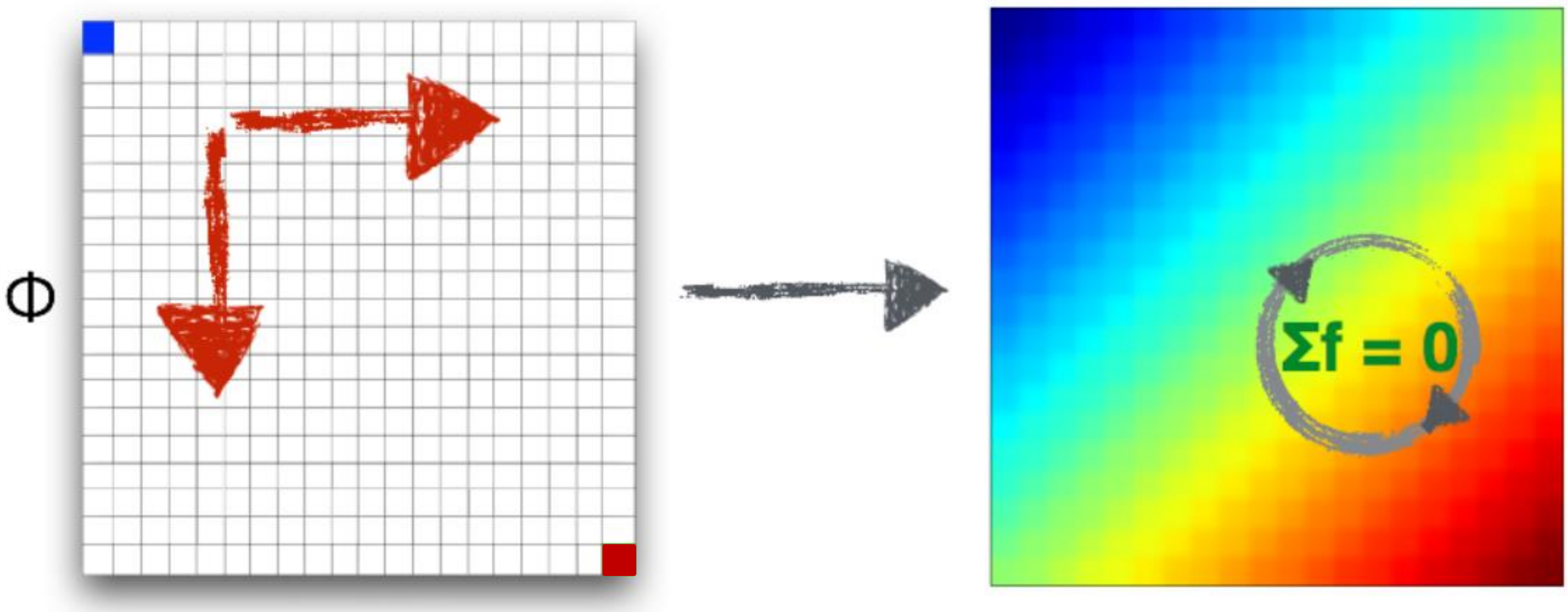




- **In RL, reward is delayed and sparse**
  - may be problematic during exploration
  - definitely is a problem during exploitation

We need to provide additional information without* altering the underlying MDP!

Potential Based Reward Shaping (PBRS)



$$f(s_t, s_{t+1}) = \gamma\, \Phi\,(s_{t+1}) - \Phi(s_t)$$

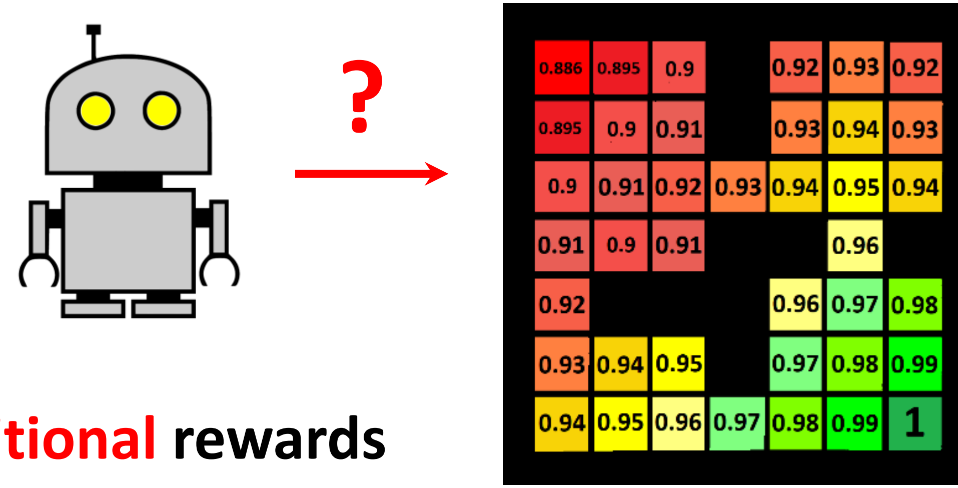
$f(s_t, s_{t+1})$: reward for moving from state s_t to s_{t+1}

$\Phi(s)$: potential of state s

γ: discount factor

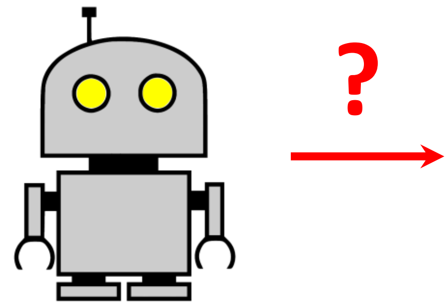
http://anna.harutyunyan.net/wp-content/uploads/2017/08/inria-march-17.pdf

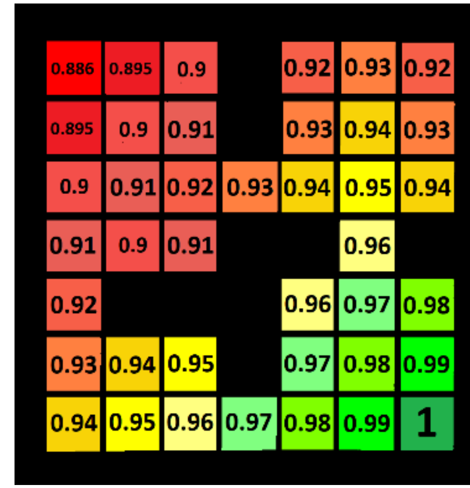
Potential Based Reward Shaping (PBRS)



- **Shaped rewards are additional rewards**
- **For the theoretic guarentees* to hold, there needs to be a shaping compensation**
  - always when moving to a final state
  - typically when ending a training episode

Potential Based Reward Shaping (PBRS)





- **Don‘t: Compare agents with shaped rewards**
  - Better: use the **raw / true*** reward function
  - Much better: keep track of (and plot) meaningfull events
- **Don‘t: Put (too) much attention to the loss**
  - Just because the loss got smaller, your agent(s) must not necessarily have learnt useful behavior
  - And even if it does not, your agent(s) may still improve

Personal, painful experience

https://imgflip.com/s/meme/One-Does-Not-Simply.jpg

→ AlphaStar

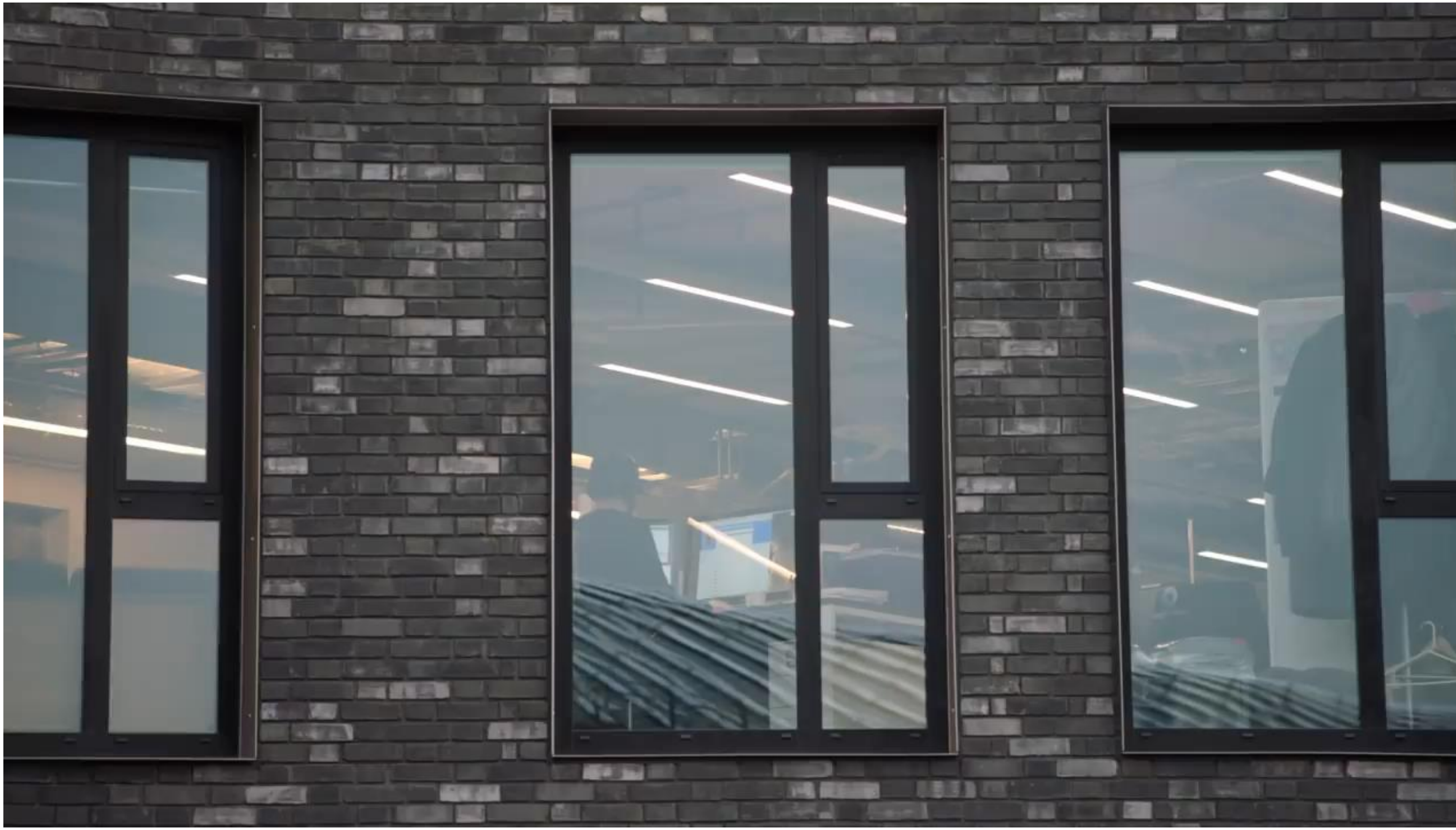
Grandmaster level in StarCraft II using multi-agent reinforcement learning

https://doi.org/10.1038/s41586-019-1724-z

https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning

AlphaStar – The inside story (2019)



https://www.youtube.com/watch?v=UuhECwm31dM

AlphaStar (2019)

- **key challenges**
  - game theory: no single best strategy
  - imperfect information
  - long term planning
  - real time
  - large action space

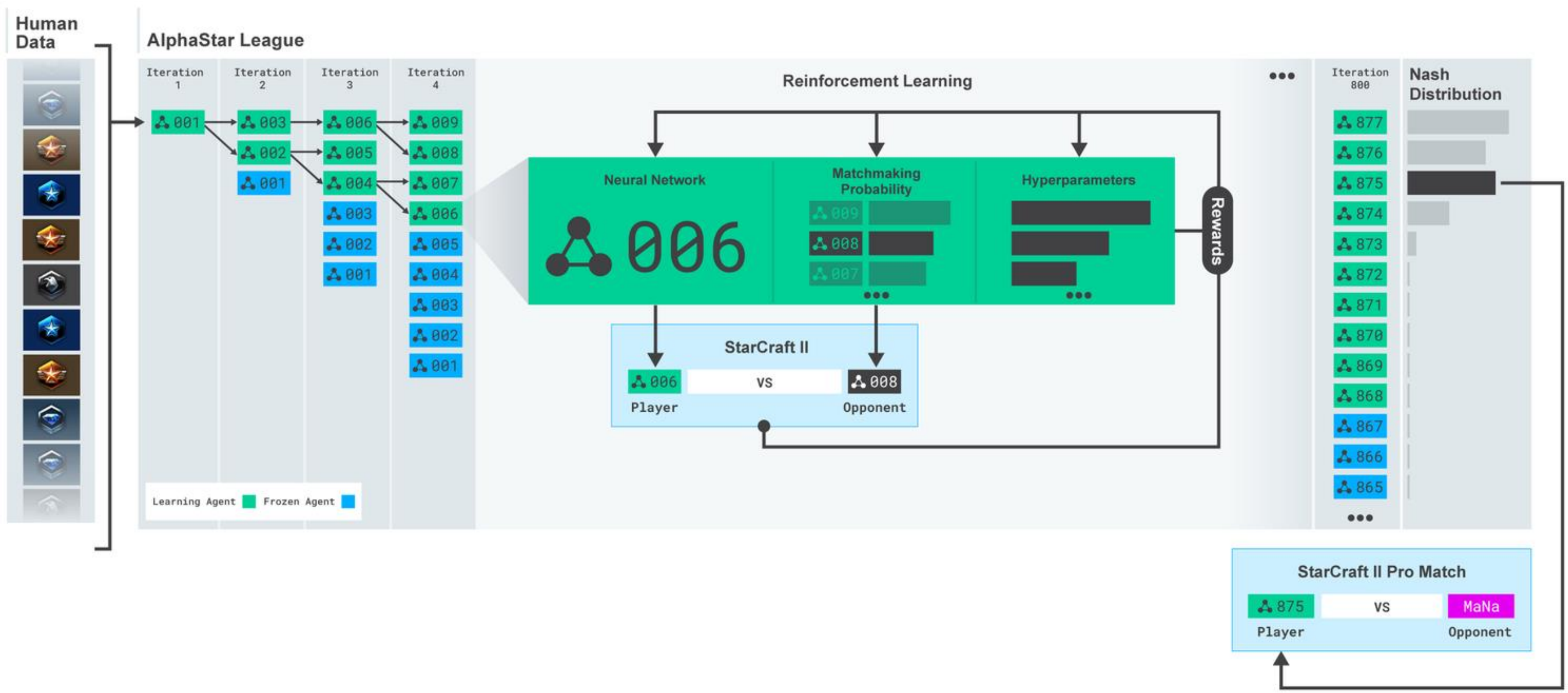
Click where?
Build/train what?
Estimated reward?

- **initial conditions**
  - single map
  - single race
  - global view
  - unlimited APM

AlphaStar (2019)

- **AlphaStar (League) agents**
  - initially trained with SL (human game replays)
  - further trained with RL (**IMPALA**, off-policy actor-critic)
- **population based approach (FTW+)**
  - original agents are kept when new agents branch
  - matchmaking probabilities and hyperparameters determine branched agent’s learning objective
  - difficulty increases iteratively while diversity is preserved
- **agents for a specific target (FTW++)**
  - sampled from the total Nash distribution of the league

AlphaStar (2019)

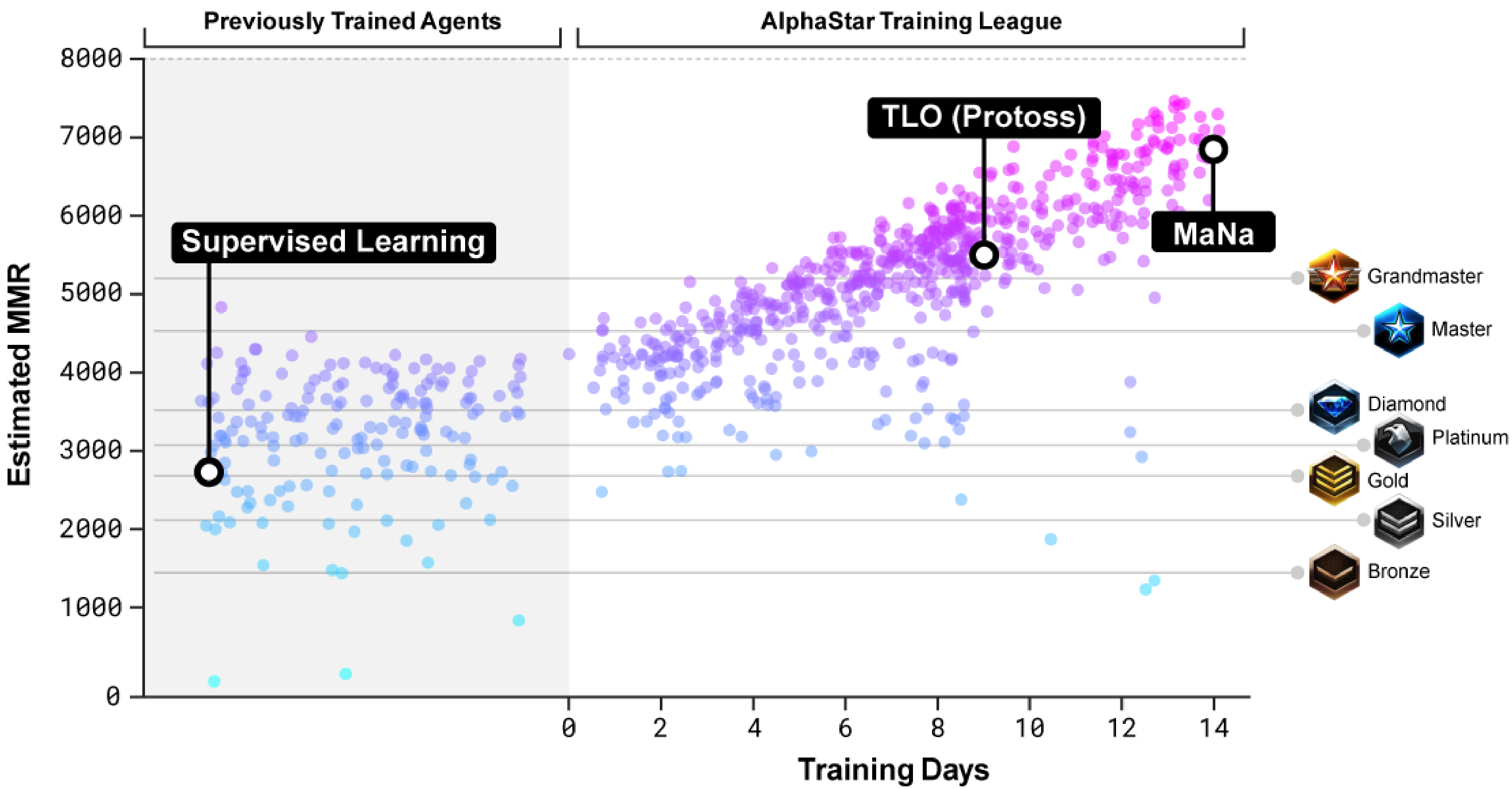


https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar (2019)

- **training a league of agents (in early 2019)**
  - took 3 (SL) + 14 (RL) days
  - each agent played ~200 years real time (-> **IMPALA**)
  - each agent utilizes ~50 GPUs for training

AlphaStar (2019)

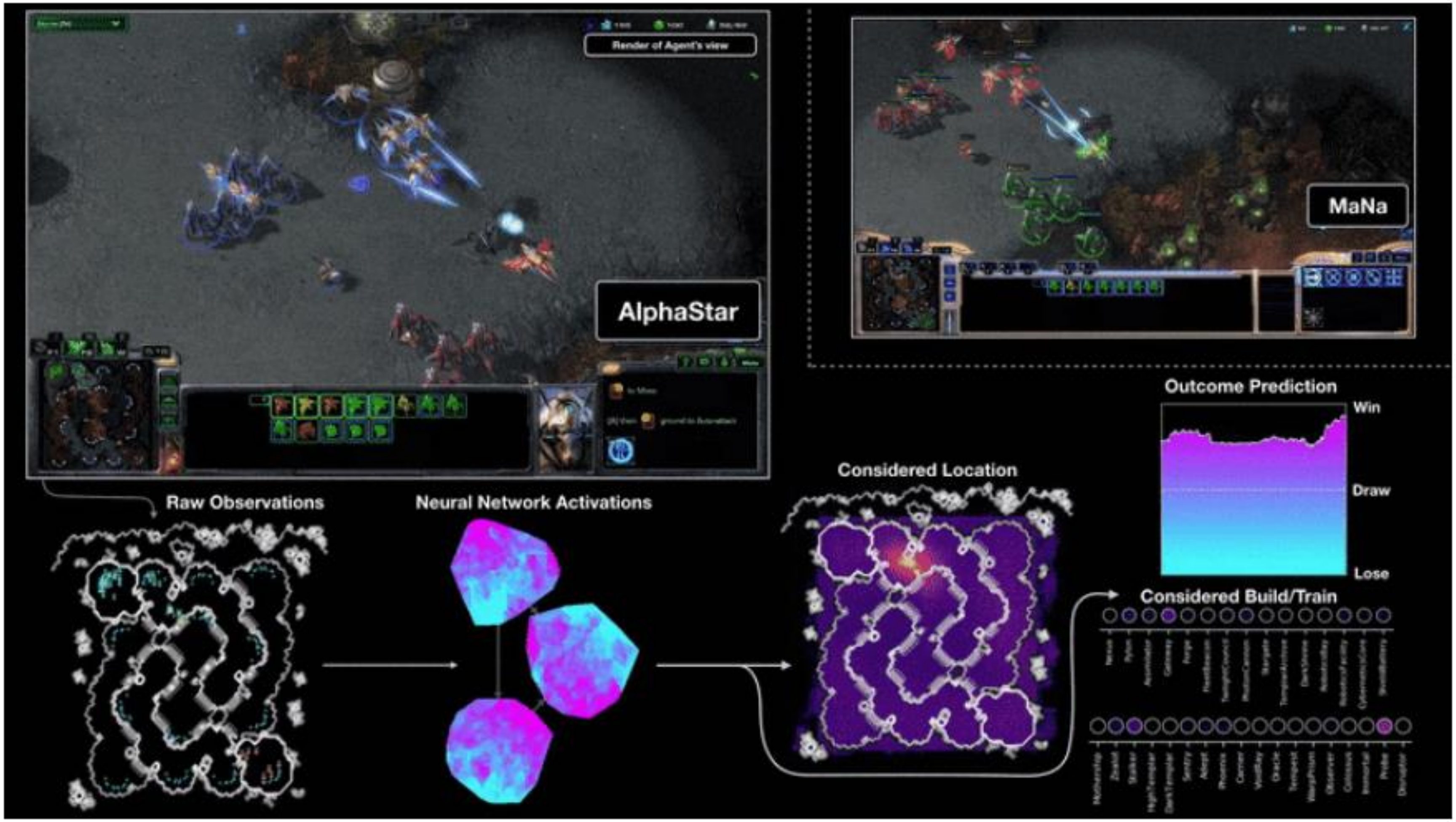


https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar (2019)

- **model architecture (in early 2019)**
  - transformer (self-attention)
  - relational deep RL (relations between objetcs)
  - deep LSTM core (combinated attention layers)
  - auto-regressive policy head (multi-dim action predictions)
  - pointer network (variable input and output lengths)
  - centralised value baseline (COMA)

AlphaStar (2019)

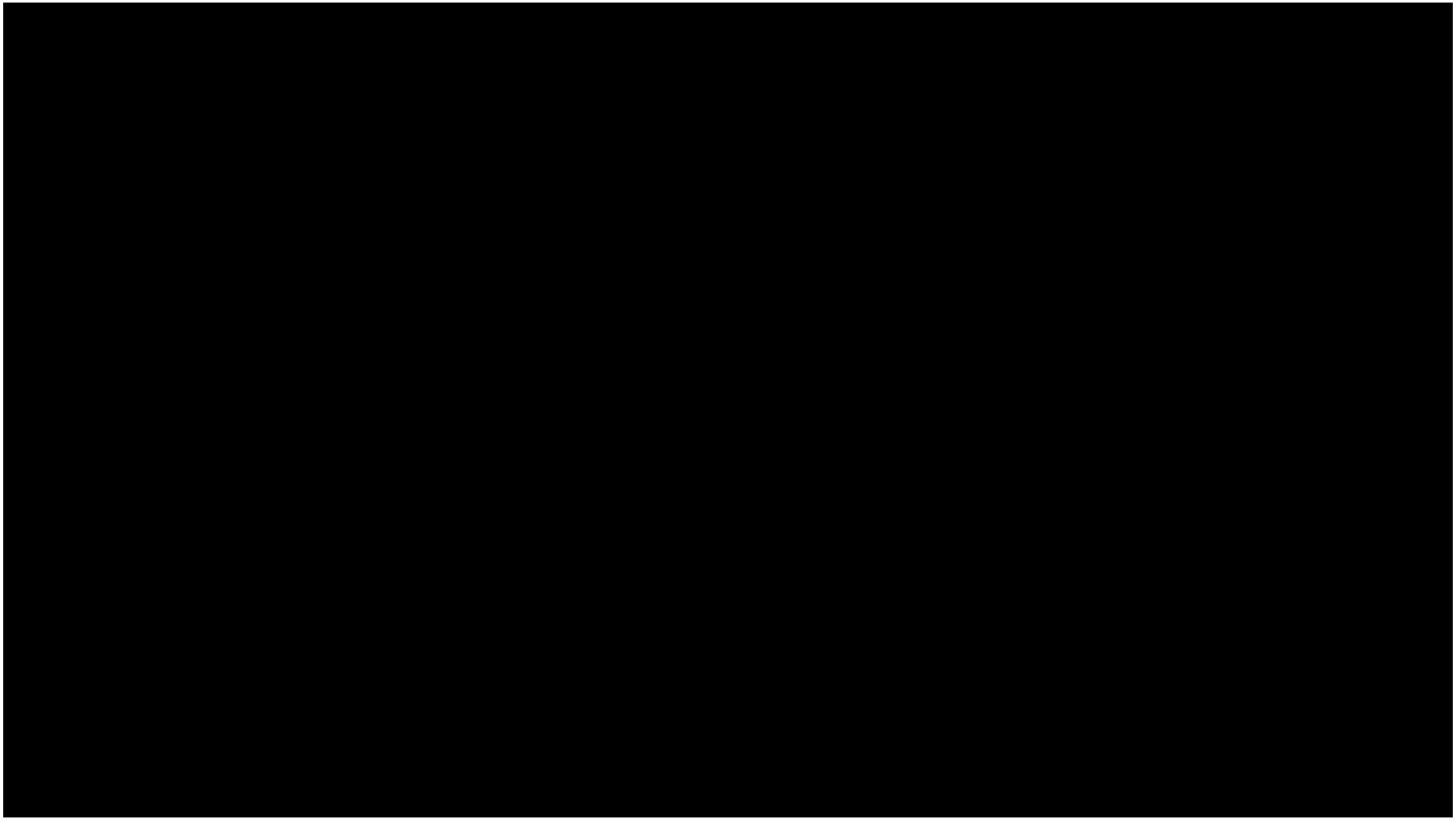


https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar (2019)

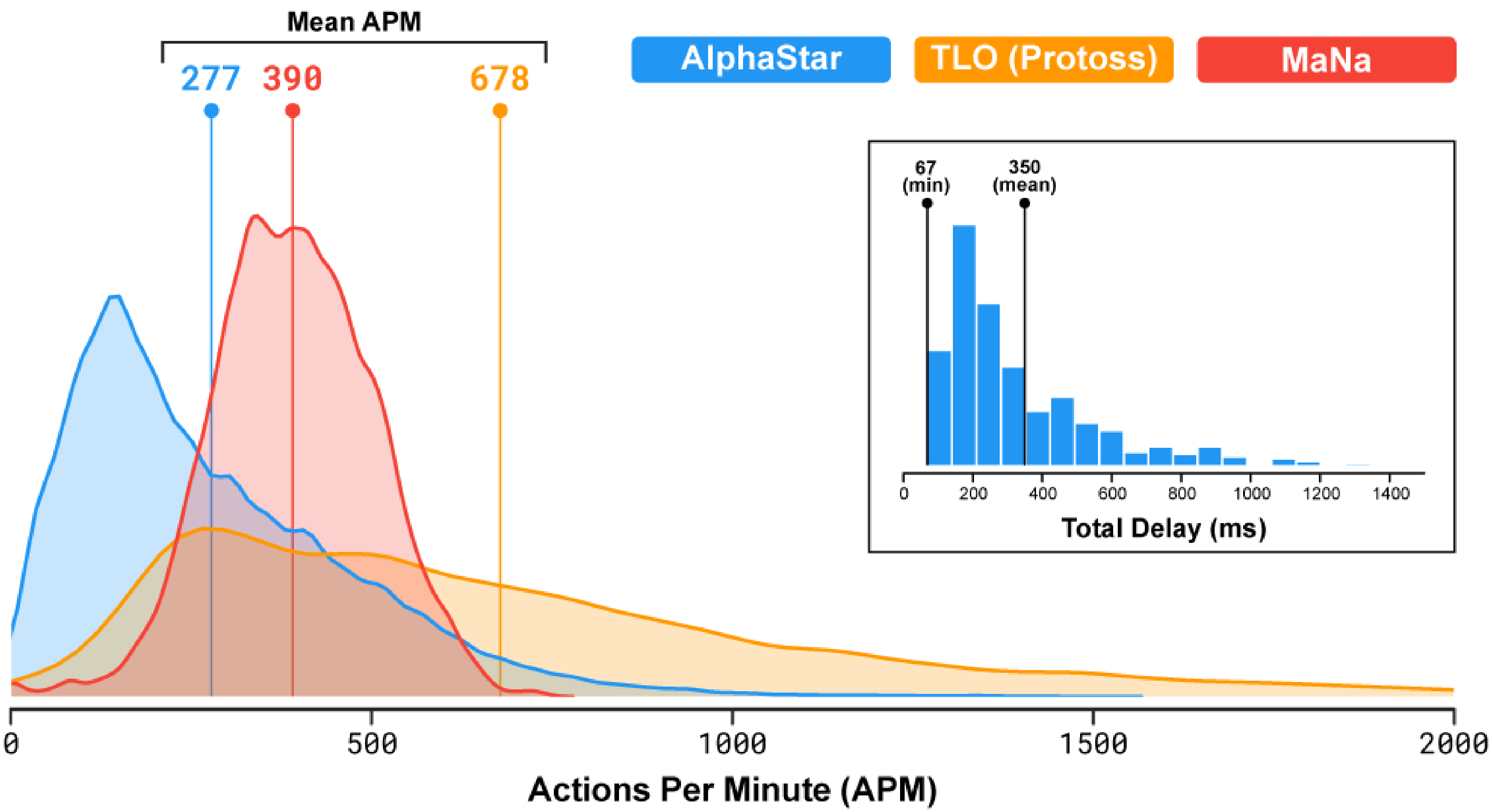
- **early 2019: human professional performance, but…**
  - one race on a single map
  - bursts of super-human APM
  - superior macro- and micro-management
  - raw interface (whole map): 10 wins, 0 losses
  - human „camera-like“ interface: 0 wins, 1 loss

Micro AI – Dodging Splash Damage (2015)



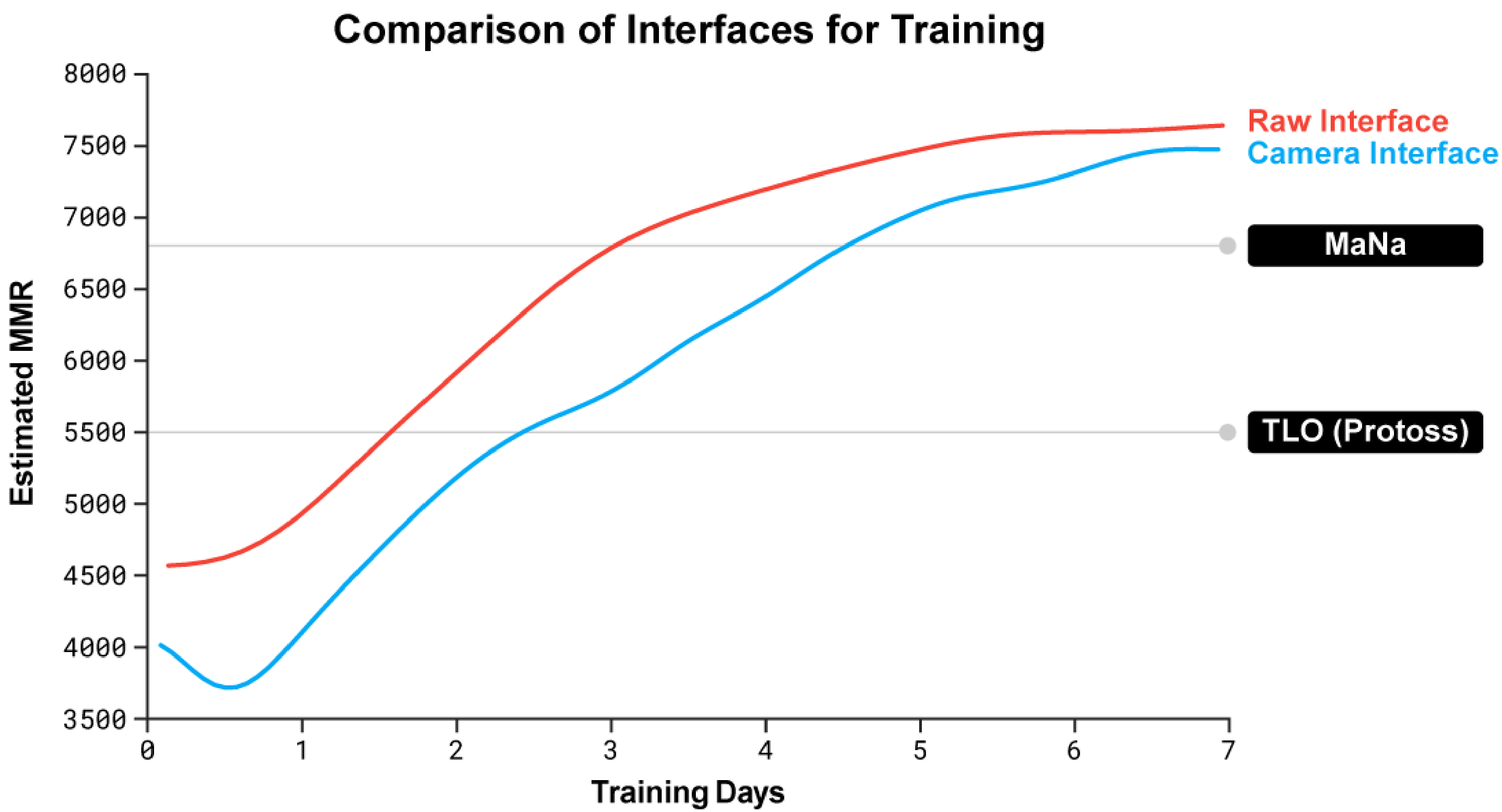
https://www.youtube.com/watch?v=lwxyFxFvi3s

AlphaStar (2019)



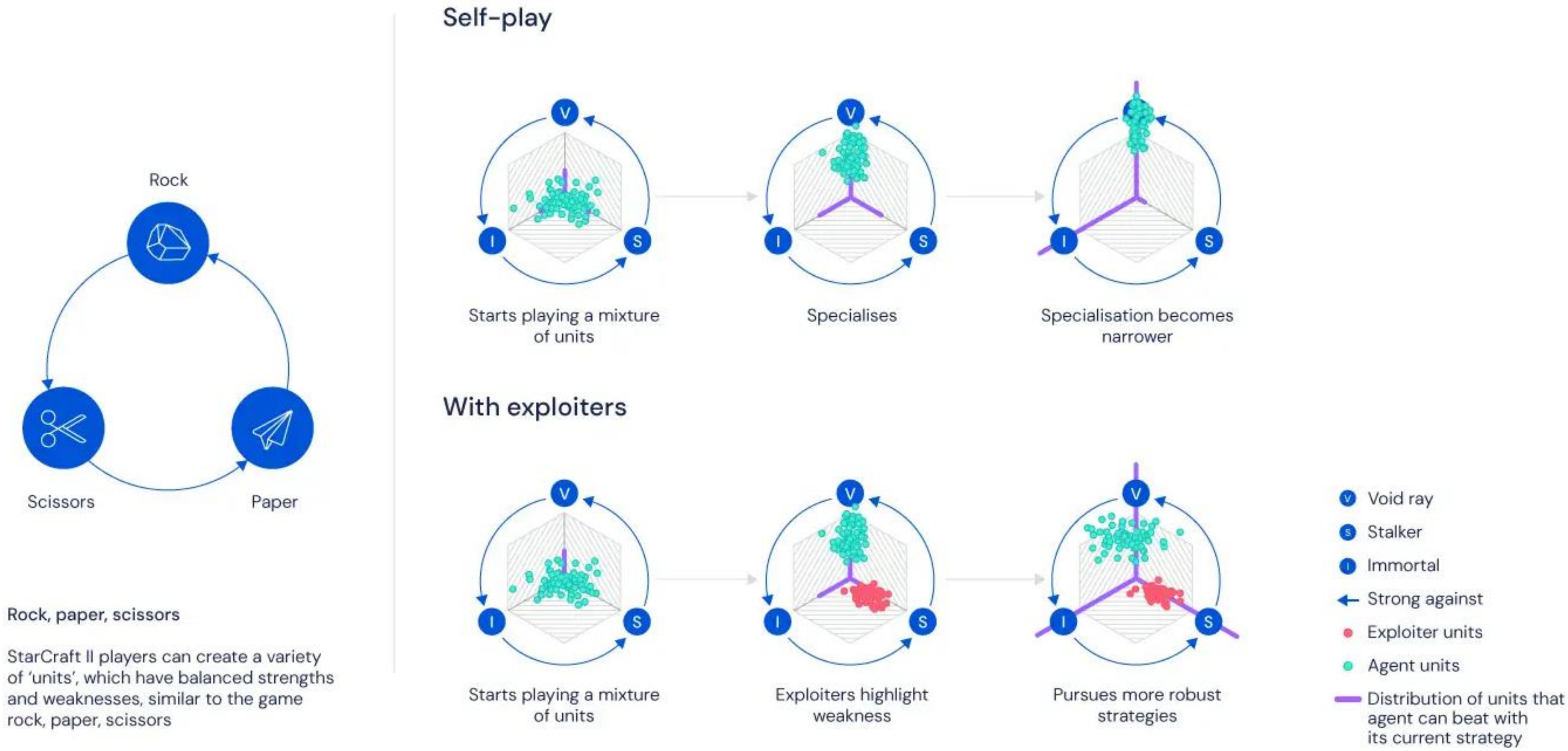
https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar (2019)



https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

AlphaStar (2019)



https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning

AlphaStar (2019)

- **late 2019: „grandmaster“ performance**
  - better than 99.8% of all players with all races on all maps
  - human „camera-like“ interface
  - APM limited to human level
  - human-like delay (30-300 ms)

take-away: pool of human-bootstrapped, diverse agents

Thank you!