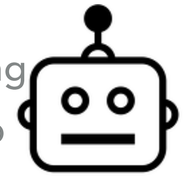
Machine Learning
Reading Group



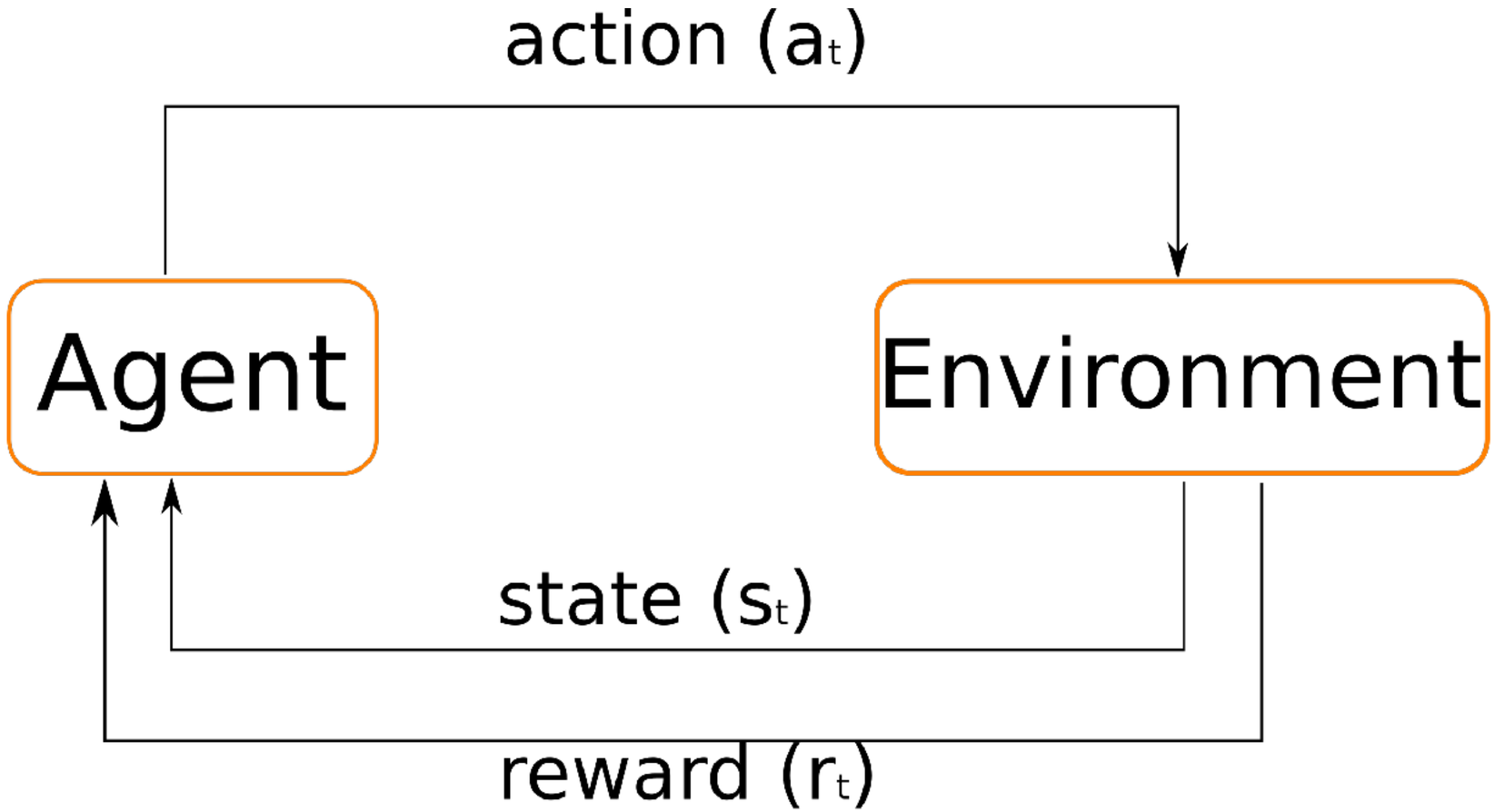
A SHORT PRESENTATION ON

REINFORCEMENT LEARNING

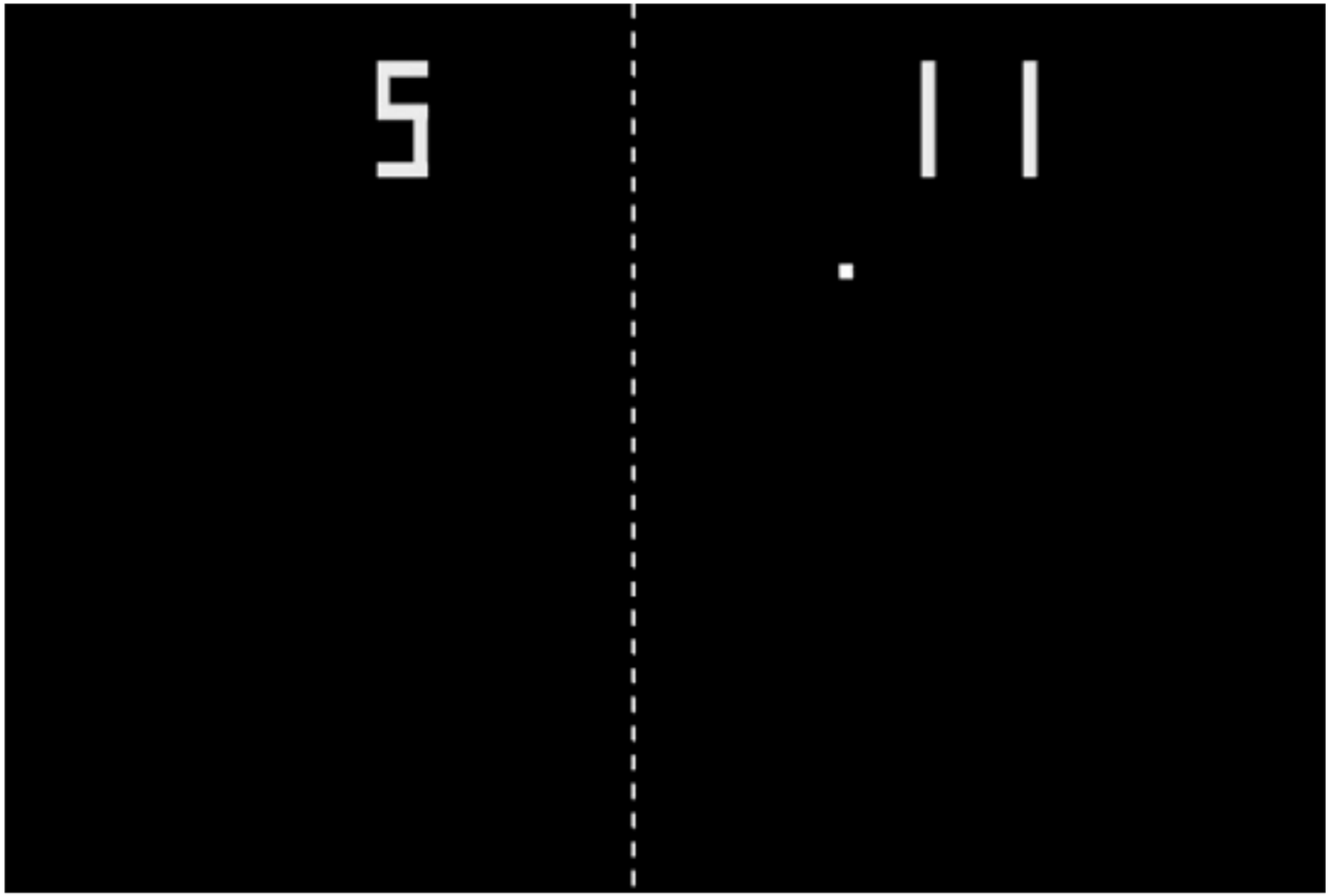
12.06.2018

Alina Miron

REINFORCEMENT LEARNING

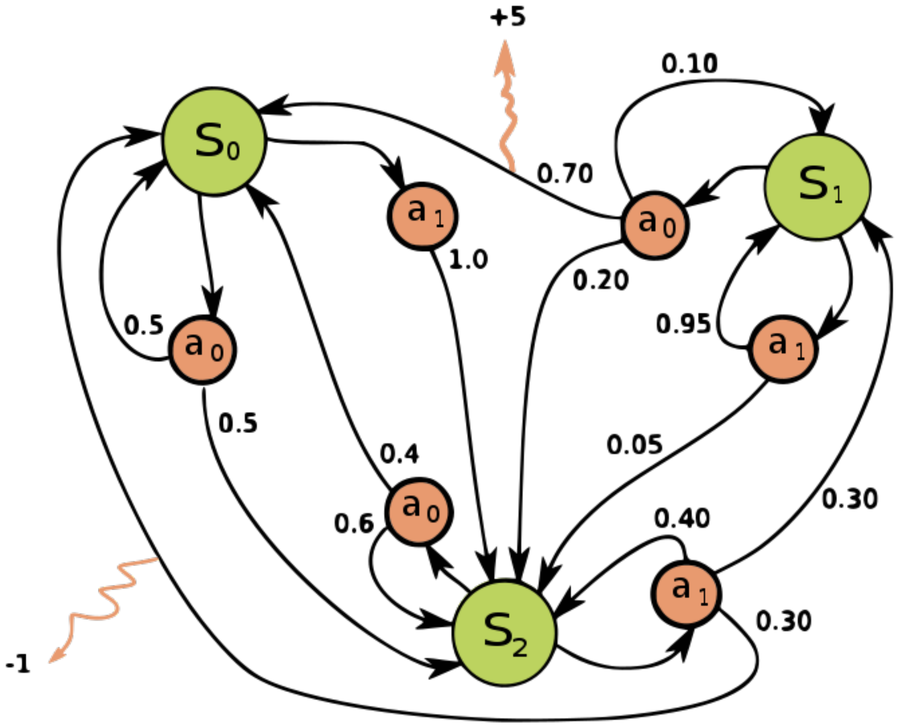


PONG

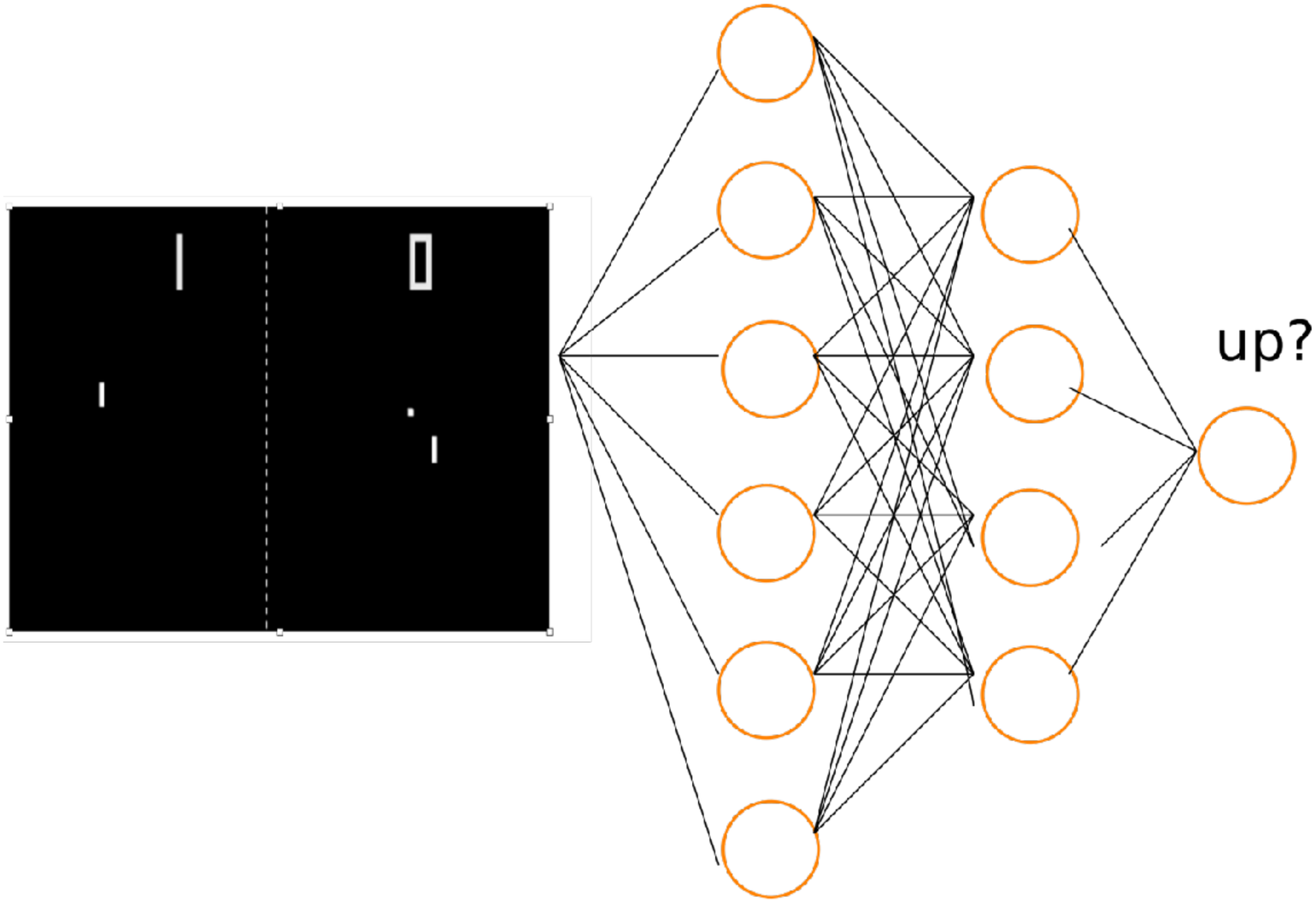


POLICY

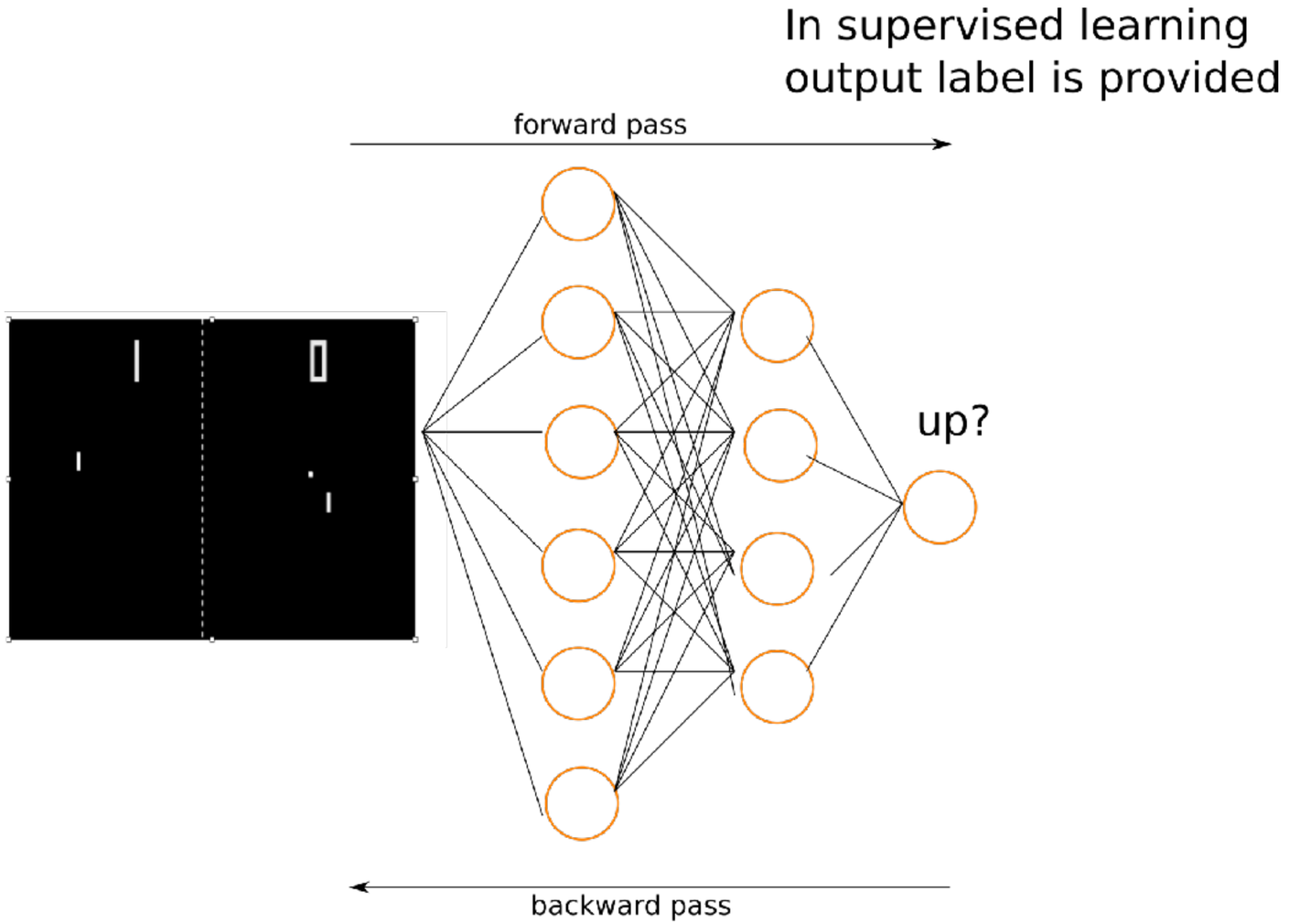

5
11


+5
0.10
S0
0.70
a1
S1
a0
1.0
0.20
0.5
a0
0.95
a1
0.5
0.05
0.4
0.6
a0
0.40
0.30
a1
S2
0.30
-1

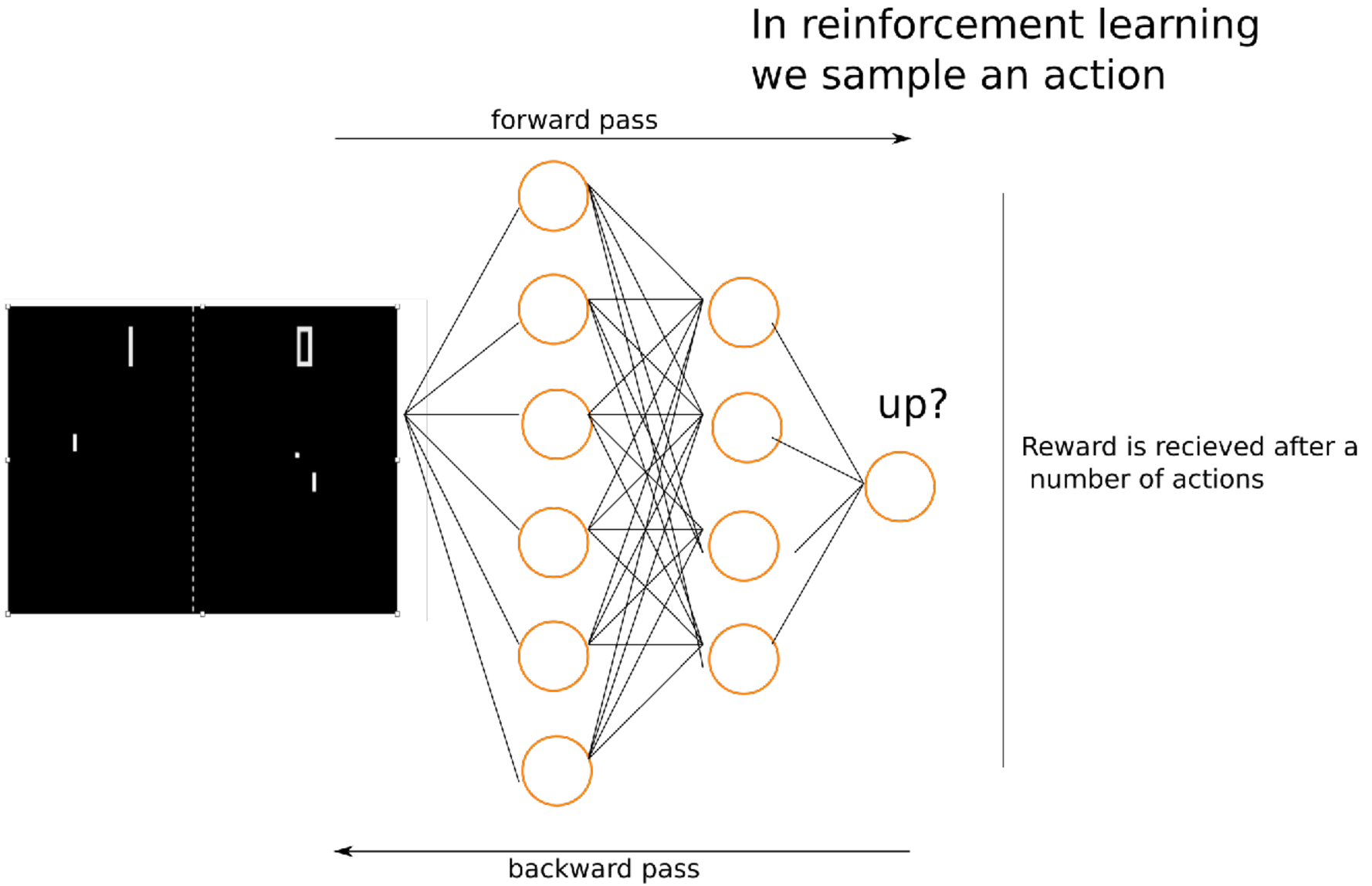
POLICY NETWORK



REINFORCEMENT LEARNING VS SUPERVISED LEARNING

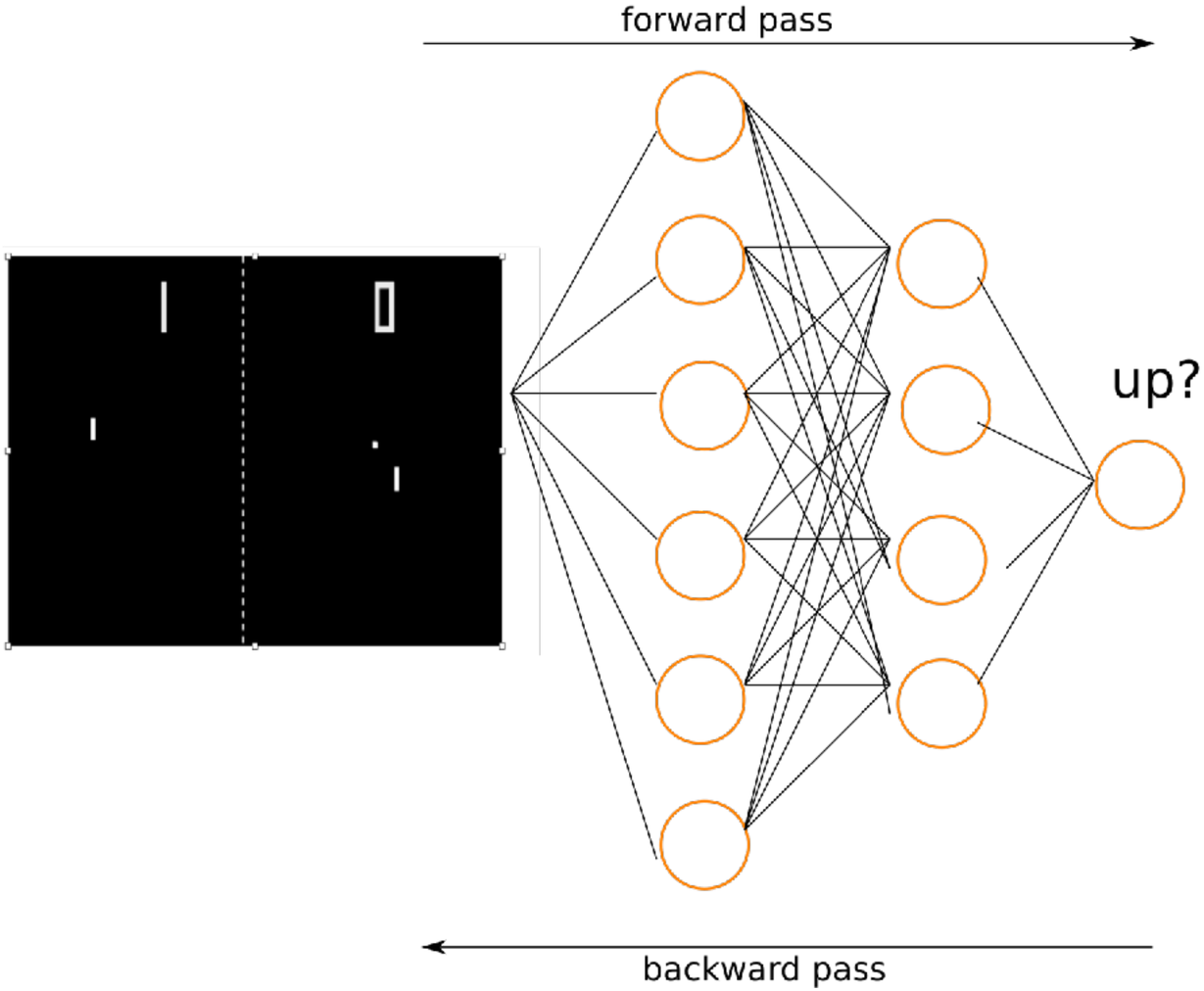


REINFORCEMENT LEARNING VS SUPERVISED LEARNING



REINFORCEMENT LEARNING: POLICY GRADIENT

In reinforcement learning
we sample an action

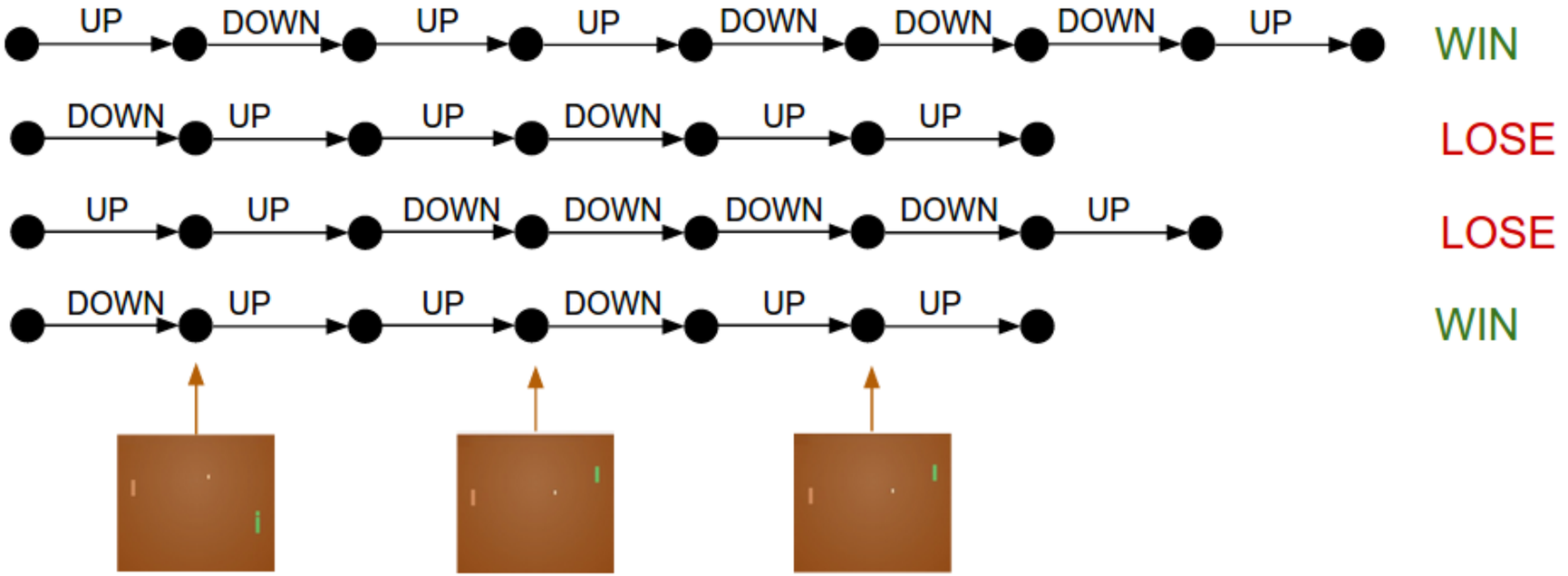


UP: 30%

DOWN: 70%

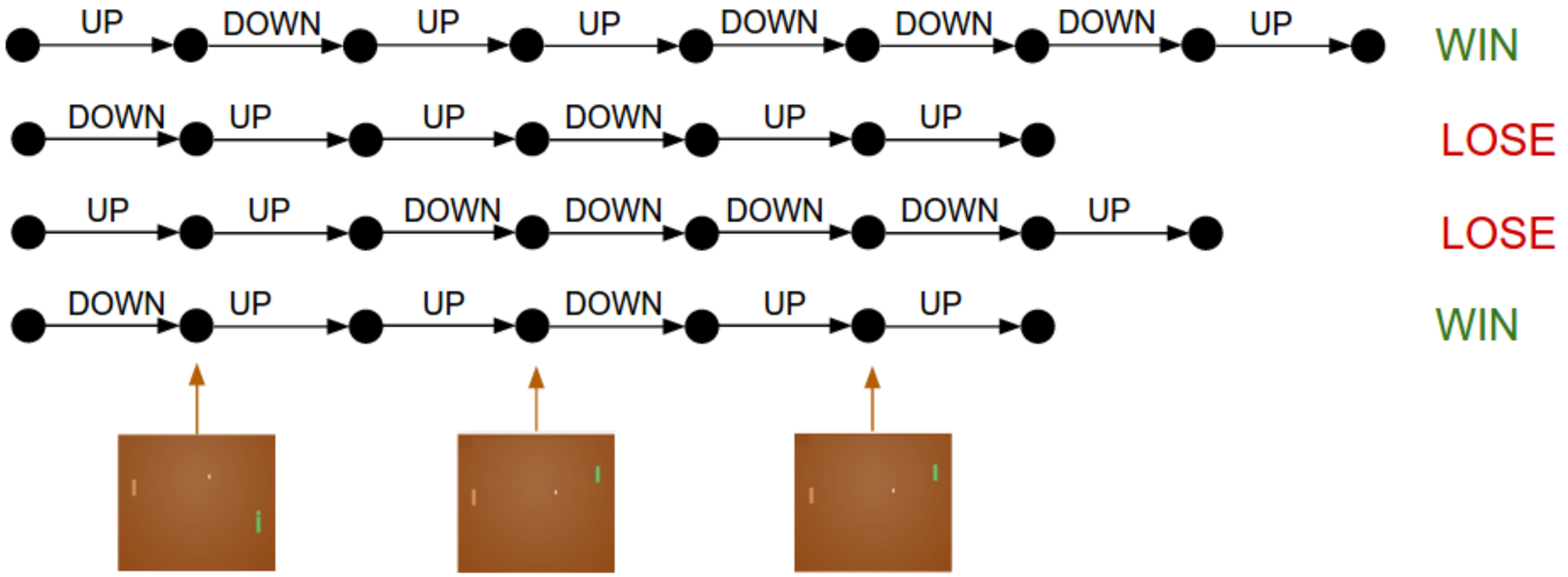
Reward is recieved after a
number of actions

REINFORCEMENT LEARNING



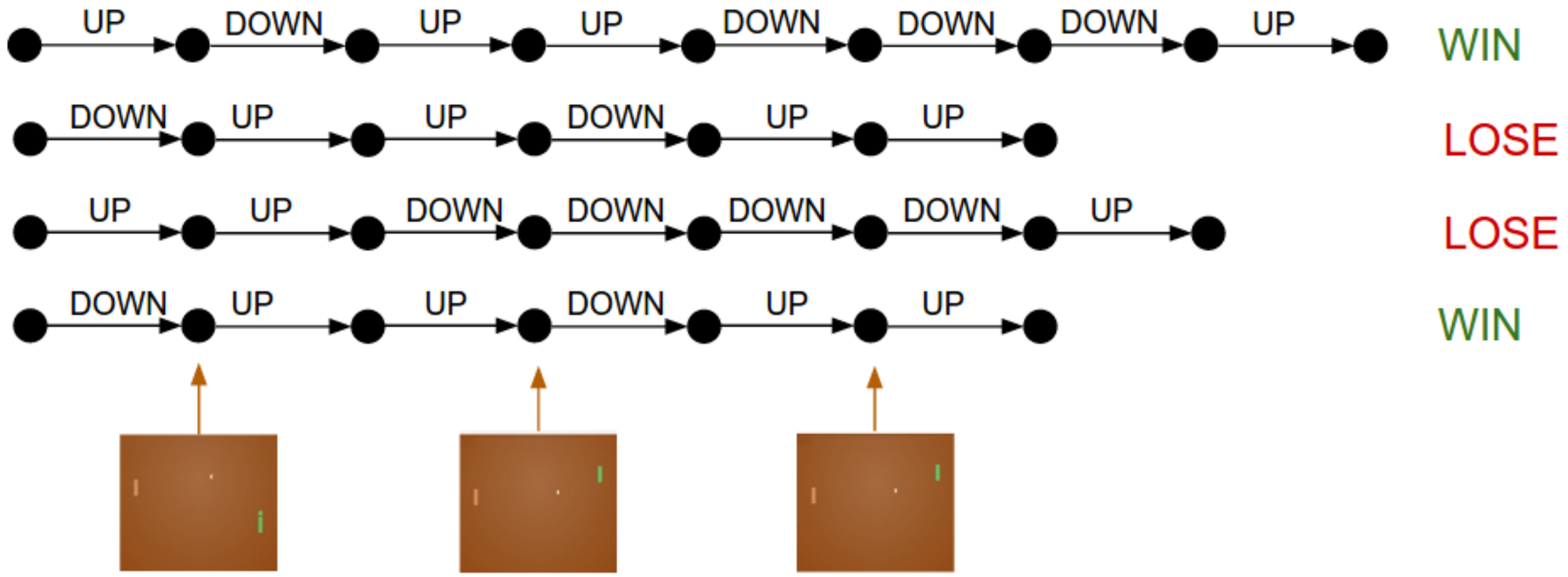
* https://karpathy.github.io/2016/05/31/rl/

CREDIT ASSIGNMENT PROBLEM



CREDIT ASSIGNMENT PROBLEM

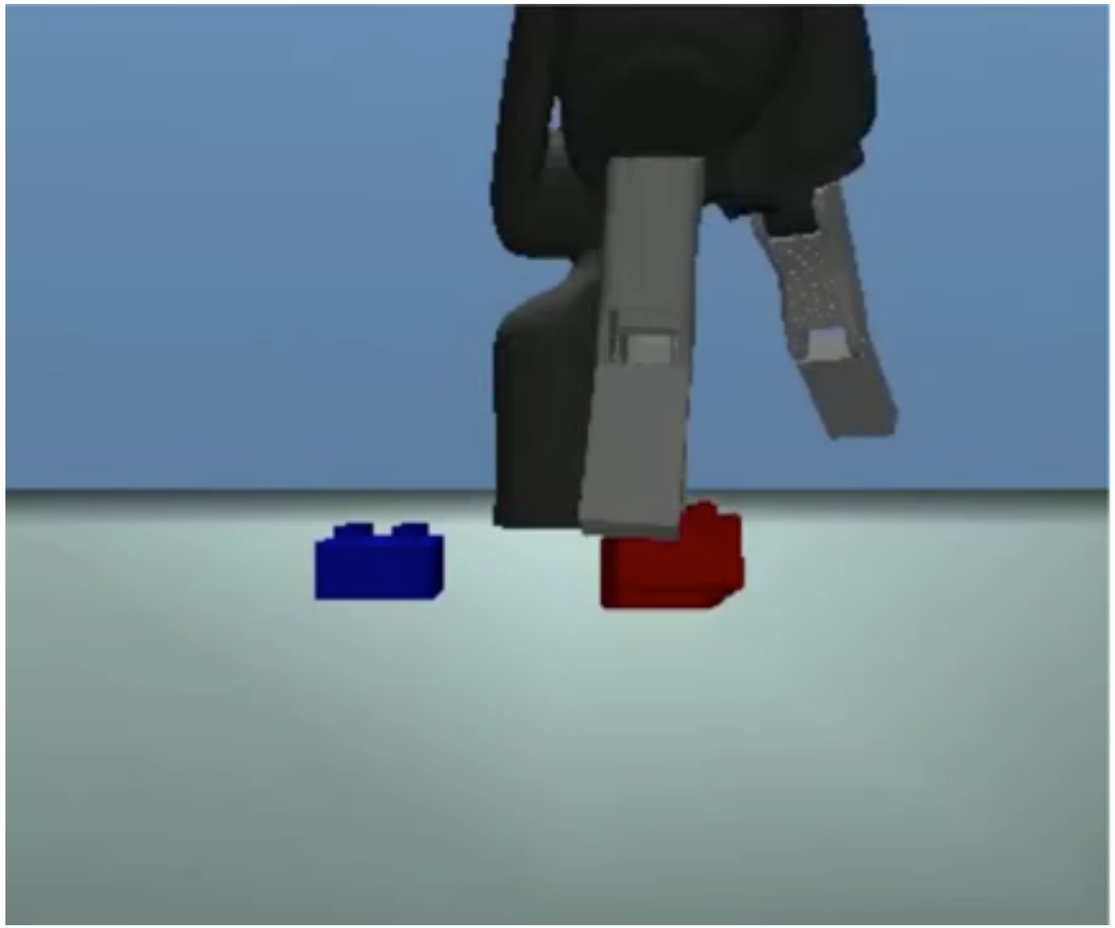
SPARSE REWARD SETTING



REWARD SHAPING

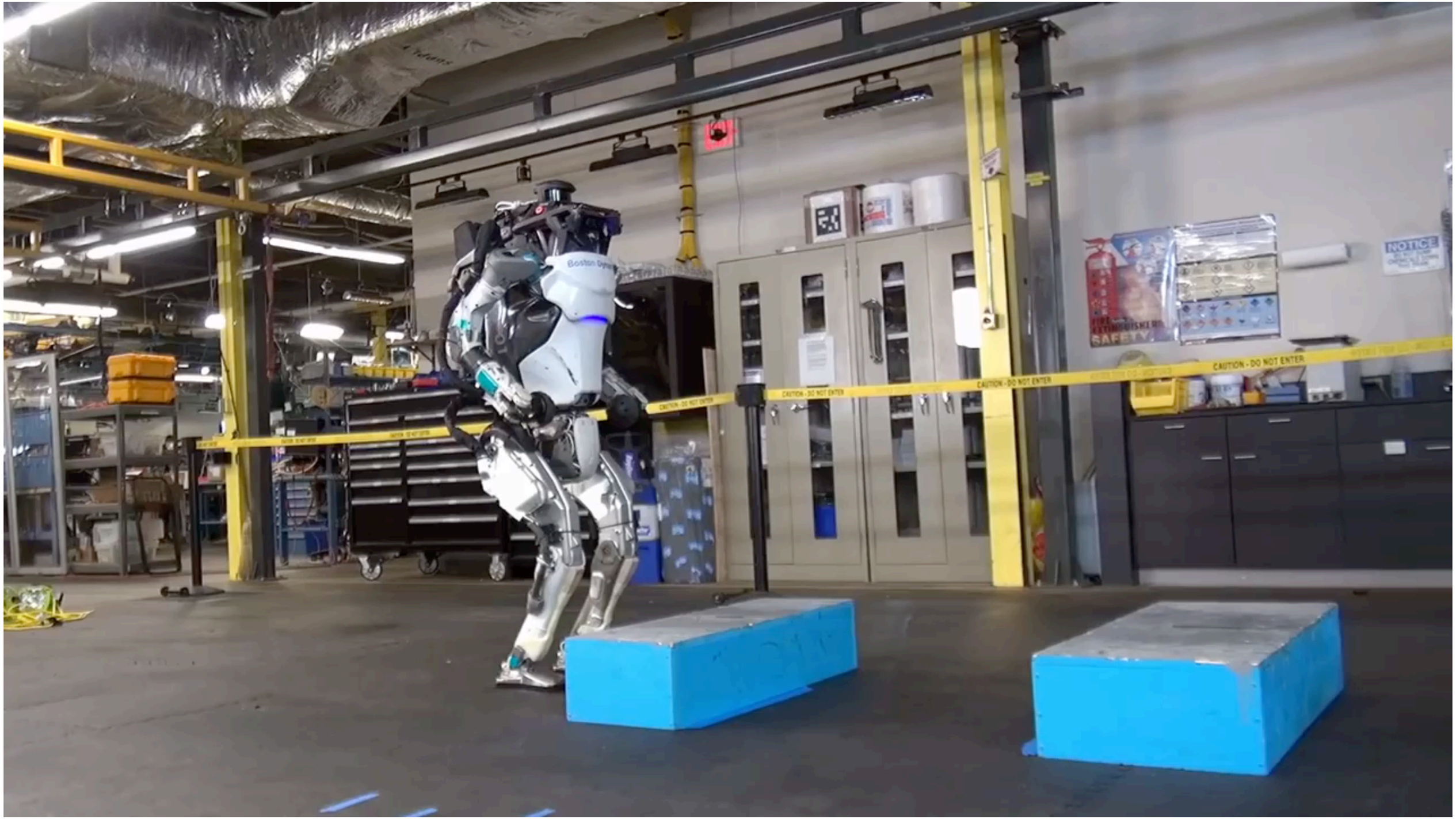
$$r(b_z^{(1)}, s^P, s^{B1}, s^{B2}) = \begin{cases} 1 & \text{if } \text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.25 & \text{if } \neg\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \wedge \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0.125 & \text{if } \neg(\text{stack}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \vee \text{grasp}(b_z^{(1)}, s^P, s^{B1}, s^{B2})) \wedge \text{reach}(b_z^{(1)}, s^P, s^{B1}, s^{B2}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

REWARD SHAPING



"Flipping"

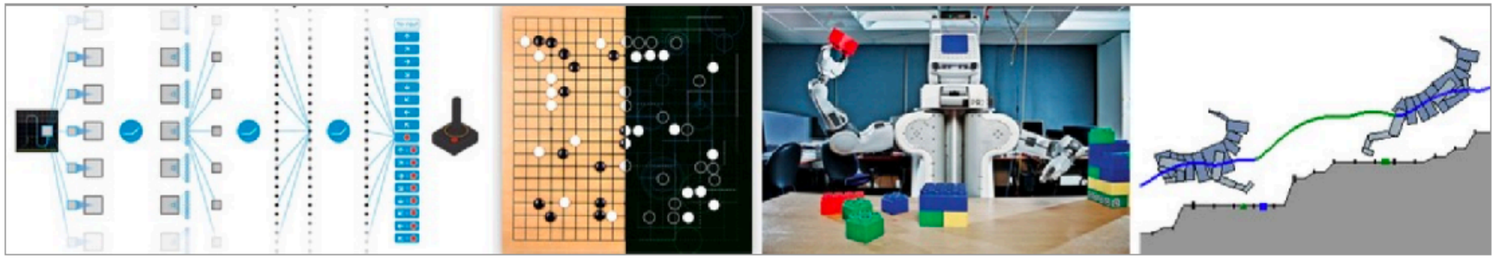
REWARD SHAPING


CAUTION - DO NOT ENTER
NOTICE
SAFETY

Deep Reinforcement Learning: Pong from Pixels

May 31, 2016

This is a long overdue blog post on Reinforcement Learning (RL). RL is hot! You may have noticed that computers can now automatically learn to play ATARI games (from raw game pixels!), they are beating world champions at Go, simulated quadrupeds are learning to run and leap, and robots are learning how to perform complex manipulation tasks that defy explicit programming. It turns out that all of these advances fall under the umbrella of RL research. I also became interested in RL myself over the last ~year: I worked through Richard Sutton's book, read through David Silver's course, watched John Schulmann's lectures, wrote an RL library in Javascript, over the summer interned at DeepMind working in the DeepRL group, and most recently pitched in a little with the design/development of OpenAI Gym, a new RL benchmarking toolkit. So I've certainly been on this funwagon for at least a year but until now I haven't gotten around to writing up a short post on why RL is a big deal, what it's about, how it all developed and where it might be going.



Examples of RL in the wild. **From left to right**: Deep Q Learning network playing ATARI, AlphaGo, Berkeley robot stacking Legos, physically-simulated quadruped leaping over terrain.

MASTERING THE GAME OF GO WITHOUT HUMAN KNOWLEDGE


AlphaGo Zero
Starting from scratch

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]

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting *tabula rasa*, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.

Much progress towards artificial intelligence has been made using supervised learning systems that are trained to replicate the decisions of human experts[1–4]. However, expert data sets are often expensive, unreliable or simply unavailable. Even when reliable data sets are available, they may impose a ceiling on the performance of systems trained in this manner[5]. By contrast, reinforcement learning systems are trained from their own experience, in principle allowing them to exceed human capabilities, and to operate in domains where human

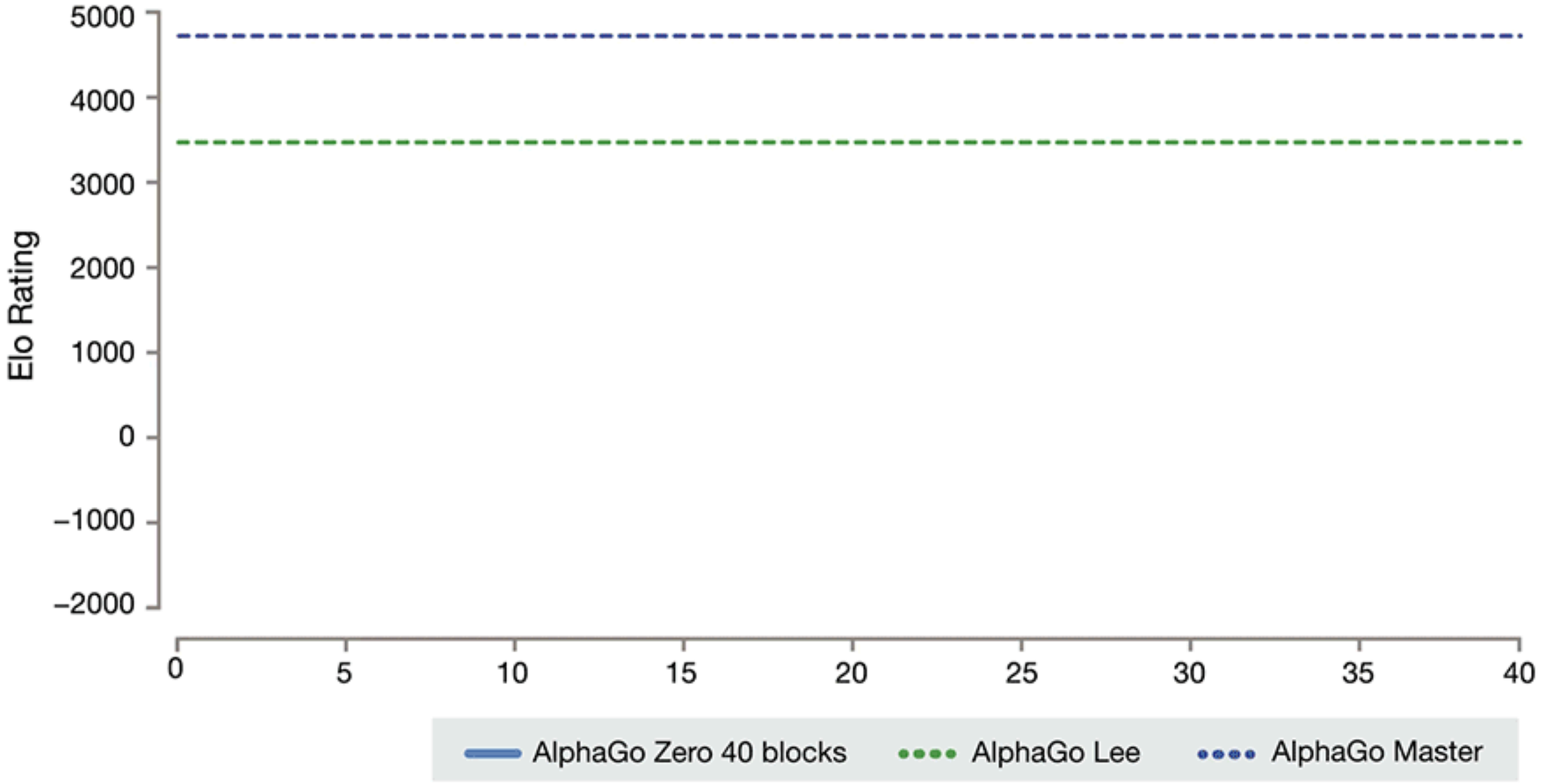
trained solely by self-play reinforcement learning, starting from random play, without any supervision or use of human data. Second, it uses only the black and white stones from the board as input features. Third, it uses a single neural network, rather than separate policy and value networks. Finally, it uses a simpler tree search that relies upon this single neural network to evaluate positions and sample moves, without performing any Monte Carlo rollouts. To achieve these results, we introduce a new reinforcement learning algorithm that incorporates

Search-Based Policy Iteration

- **Search-Based Policy Improvement**
  - Run MCTS search using current network
  - Actions selected by MCTS > actions selected by raw network

- **Search-Based Policy Evaluation**
  - Play self-play games with AlphaGo search
  - Evaluate improved policy by the average outcome

See also: Lagoudakis 03, Scherrer 15, Anthony 17


5000
4000
3000
2000
1000
0
−1000
−2000
Elo Rating
0
5
10
15
20
25
30
35
40
AlphaGo Zero 40 blocks
AlphaGo Lee
AlphaGo Master

It is a narrow AI:

1. fully **deterministic.**
2. fully **observed.**
3. the action space is **discrete**
4. we have access to a perfect **simulator** (the game itself)s.
5. each episode/game is relatively **short,** of approximately 200 actions.
6. the **evaluation** is clear, fast and allows a lot of **trial-and-error** experience

DEEPMIND ALPHAZERO - MASTERING GAMES WITHOUT HUMAN KNOWLEDGE: 2017 NIPS KEYNOTE BY DEEPMIND'S DAVID SILVER

HTTPS://WWW.YOUTUBE.COM/WATCH?V=WUJY7OZVDJK