Reinforcement Learning

Motivation

Create Myanmar Reinforcement Learning Research Community

Great Expectations

WILL YOU BE

an expert at RL after the class - NO

able to train world champion chess AI - NO

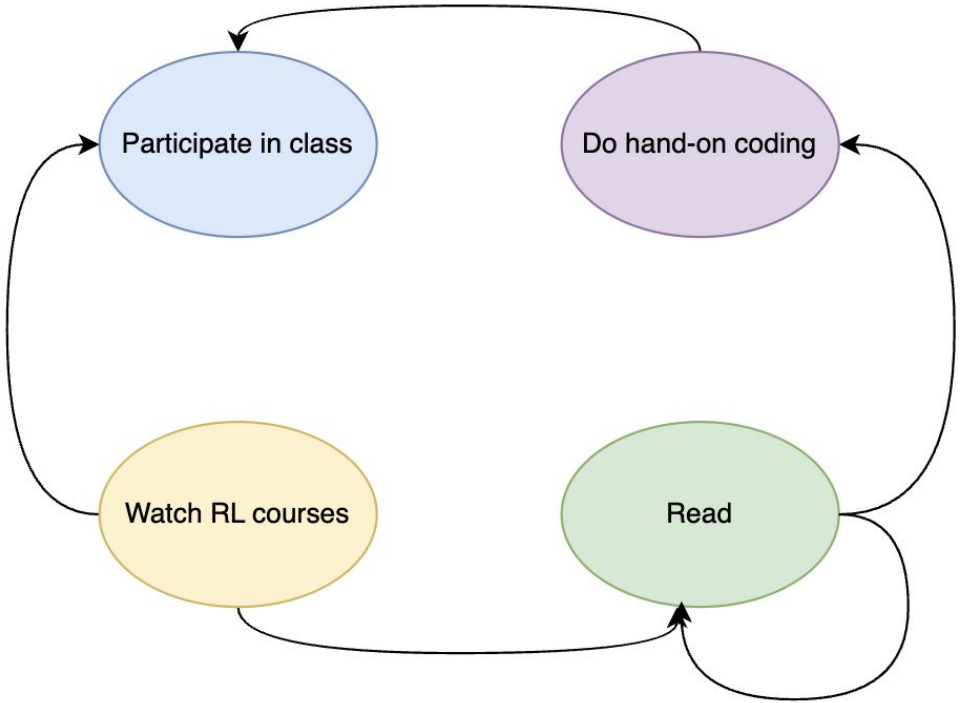
able to implement RL algorithms on your own - MAYBE

able to understand what RL is - YES

Student Profile

	Student 1	Student 2	Student 3
Machine Learning	Well-versed, have done deep learning problems	Beginner, have done toy problems	Very Basic
Reinforcement Learning (RL)	Know definitions and fundamental concept	Know the terms, but No Idea Yet	No Idea Yet
Motivation	Pursue RL	Understand RL	Get to know RL

How to Make the Most out of this class



Types of Machine Learning

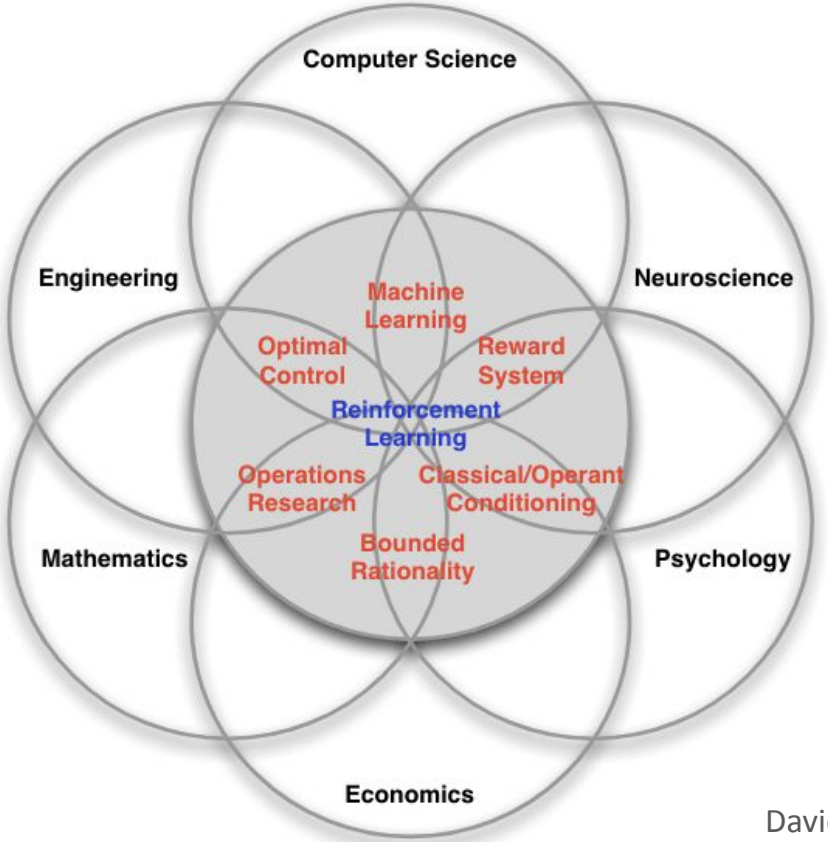
Supervised Learning

Unsupervised Learning

Reinforcement Learning

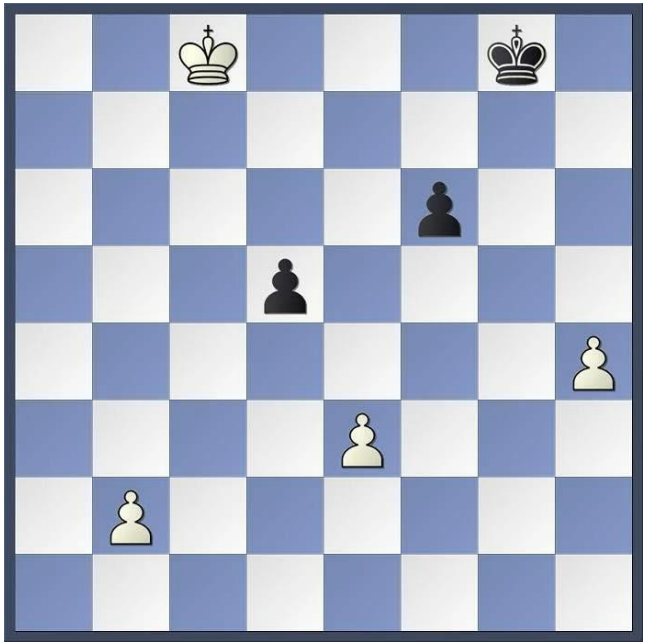
What is Reinforcement Learning (RL)?

Many Faces of Reinforcement Learning



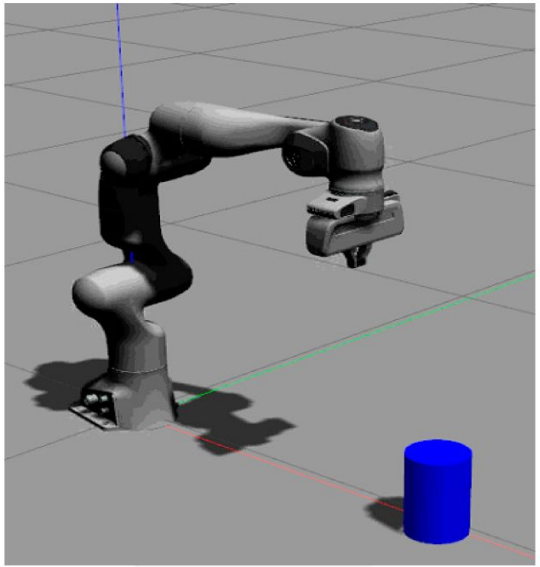
David Silver
https://www.davidsilver.uk/teaching/

Board games

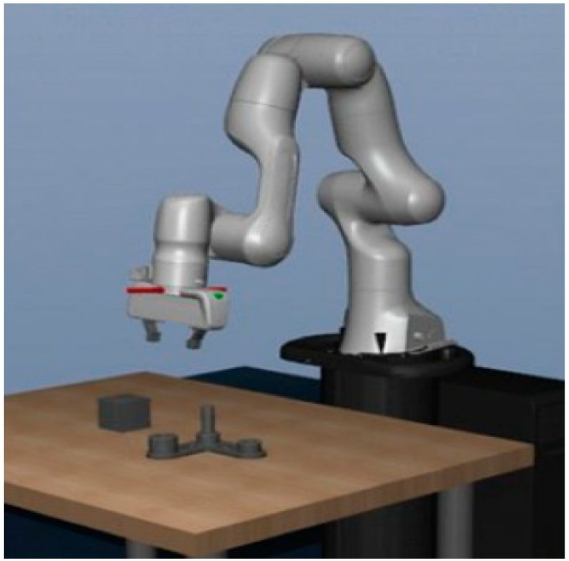


- Chess
- Go
- Connect four
- Basically, all board games

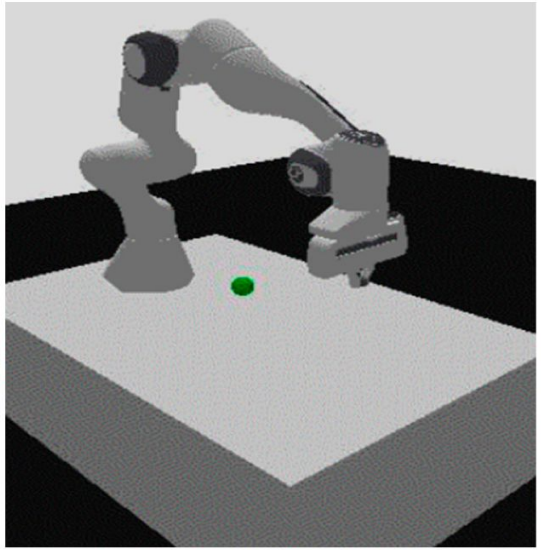
Robotics



(a)



(b)



(c)

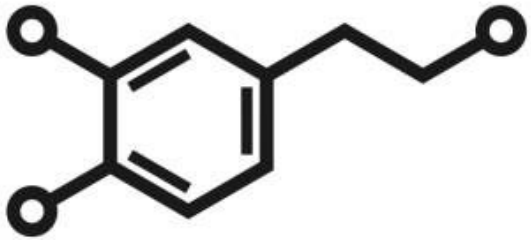
Other Examples

Recommender systems

Stock trading

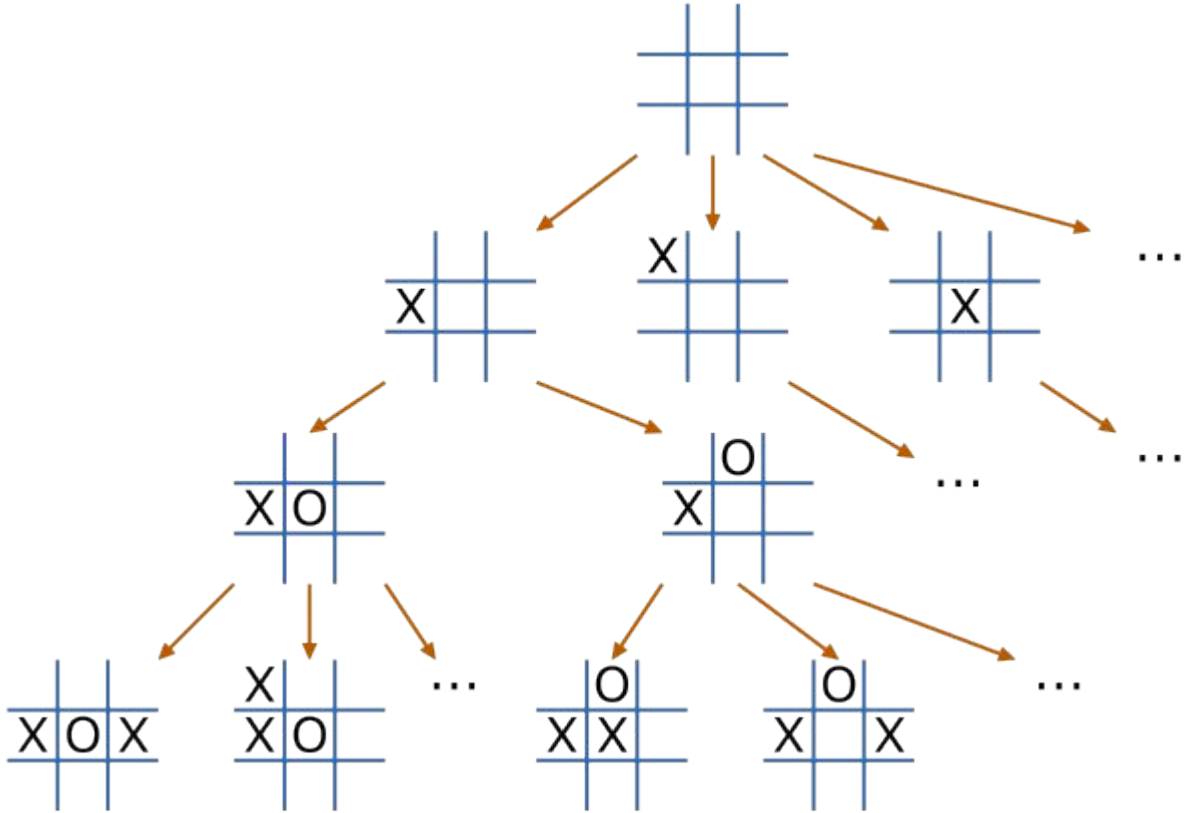
Language model alignment

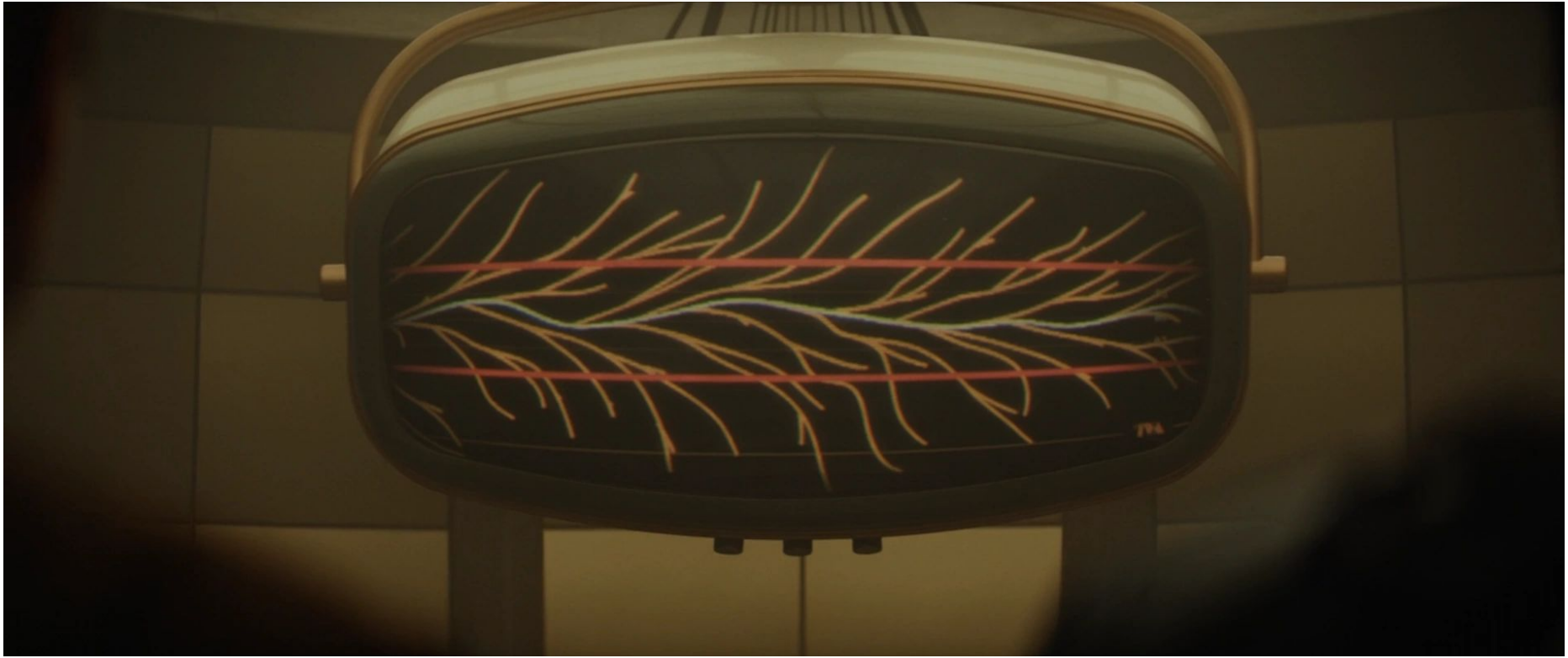
Reward



DOPAMINE

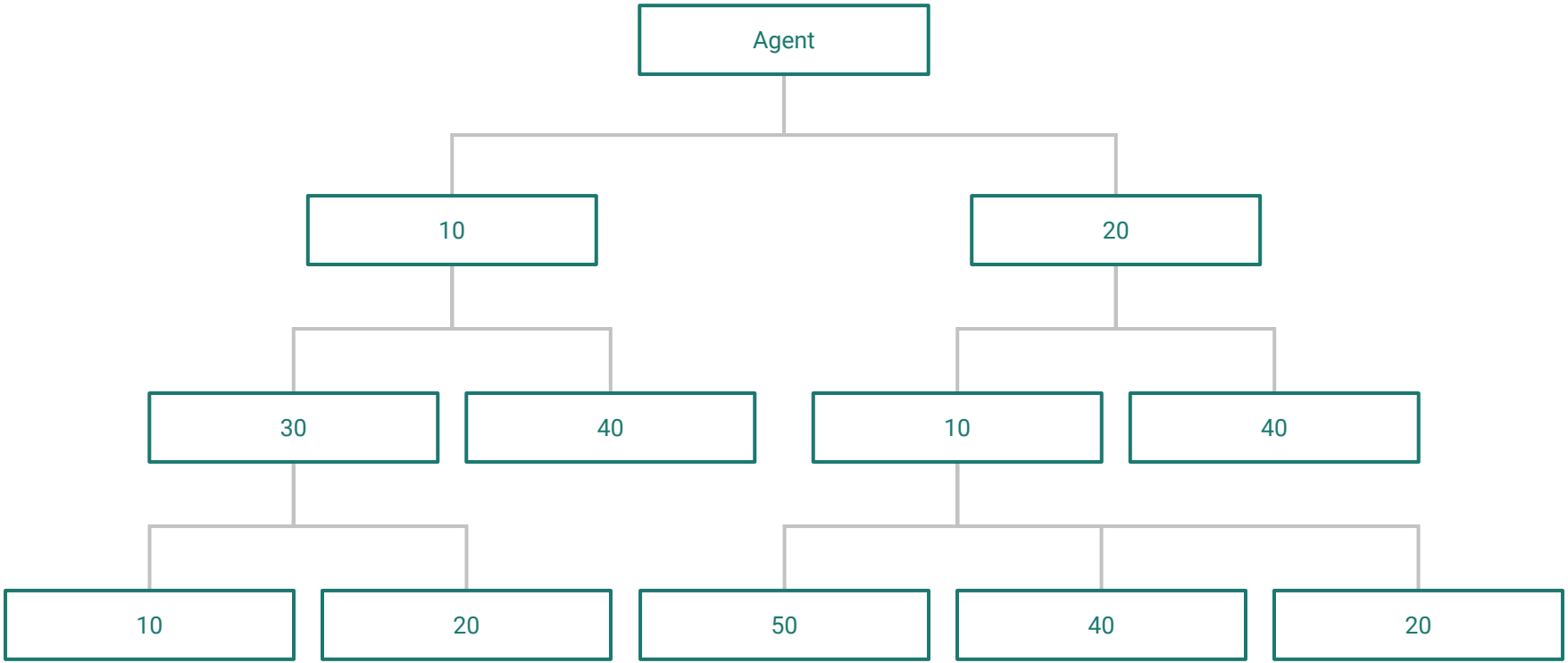
Do we always know the reward?


X
X
X
...
X O
O
X
...
...
X O X
X
X O
...
O
X X
O
X
X
...



If we know it, it’s called supervised learning. Or it becomes a search problem.

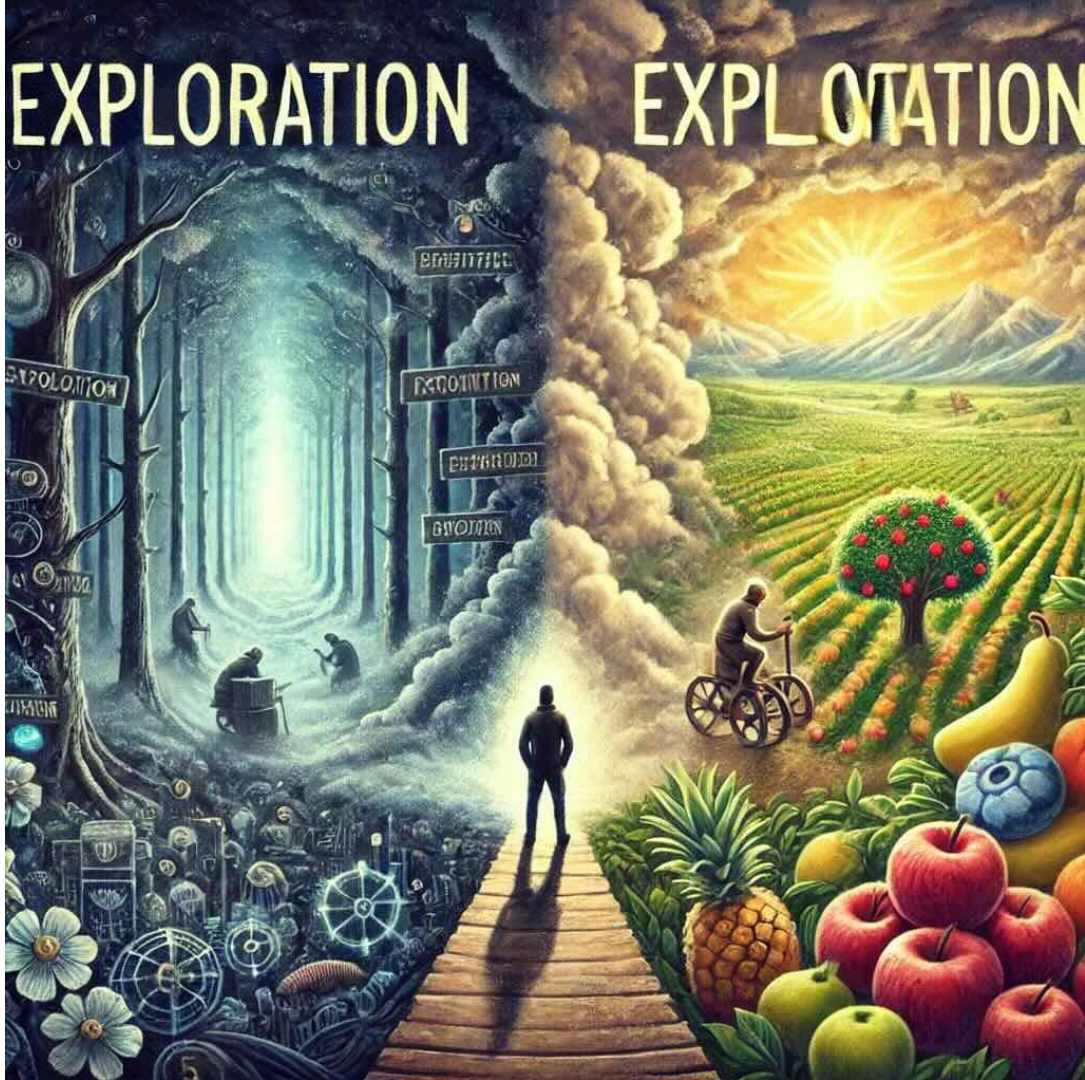
Chess Engines vs RL


Agent
10
20
30
40
10
40
10
20
50
40
20

Exploration

Vs

Exploitation



Life Lesson

- Explore
- But how?
- How do we know the future?
- How can we learn to explore?

Credit Assignment Problem

- Why do we win?
- Why do we lose?
- What went well?
- What went wrong?

MATHEMATICS

Expectation

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i$$

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

Expected value of a Fair Die

p(1) = ⅙

p(2) = ⅙

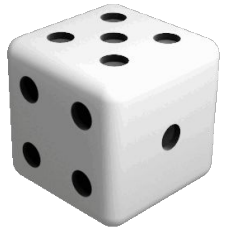
p(3) = ⅙

p(4) = ⅙

p(5) = ⅙

p(6) = ⅙

E(X) = (⅙ * 1) + (⅙ * 2) + (⅙ * 3) + (⅙ * 4) + (⅙ * 5) + (⅙ * 6)



Expected value of unfair die

p(1) = 0.3

p(2) = 0.25

p(3) = 0.1

p(4) = 0.075

p(5) = 0.075

p(6) = ?

E(X) = ?

Expected value of mystery die

Probabilities of rolls are not given?

How would you find the expected value?

Law of Large Numbers

In probability theory, the law of large numbers (LLN) is a mathematical theorem that states that the average of the results obtained from a large number of independent random samples converges to the true value, if it exists

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

$$Q_t(a) = \frac{R_1 + R_2 + \cdots + R_{N_t(a)}}{N_t(a)}$$

Assume Q_t (a) is expected value of a dice

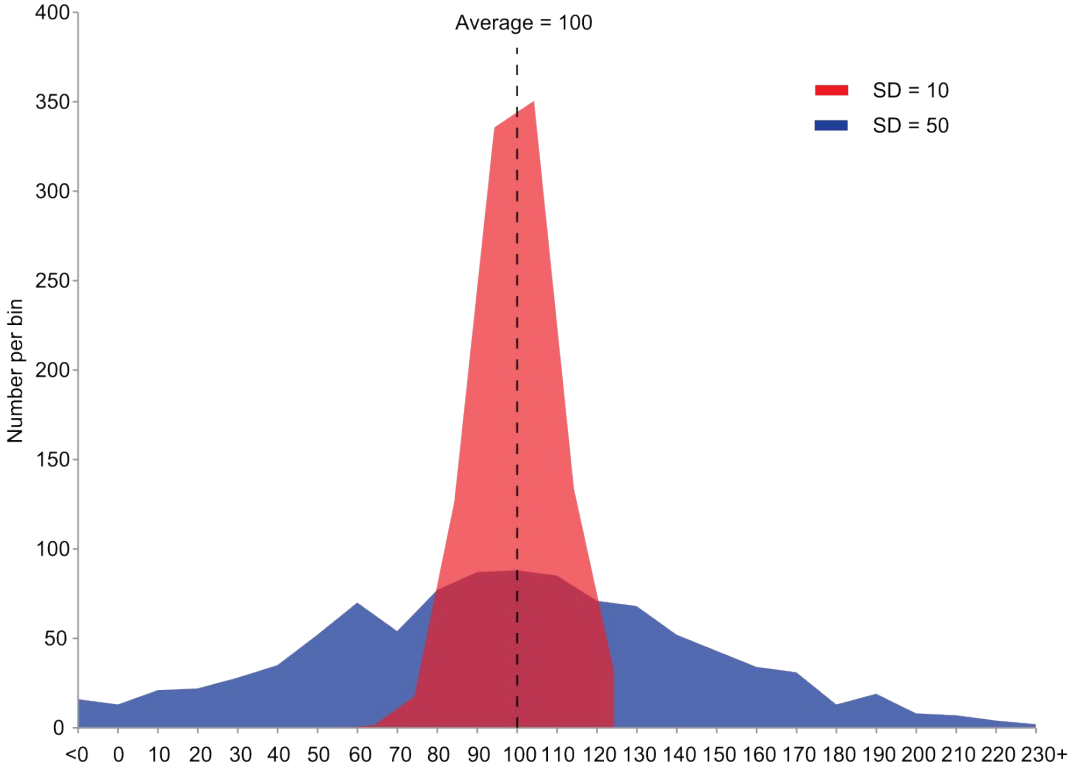
R_1, R_2 ... are values of dice rolls.

N_t is number of dice rolls.

$$
\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}
$$

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \left[Target - OldEstimate \right].$$

Variance


Average = 100
SD = 10
SD = 50
Number per bin
400
350
300
250
200
150
100
50
0
<0 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230+

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2,$$

where μ is the expected value. That is,

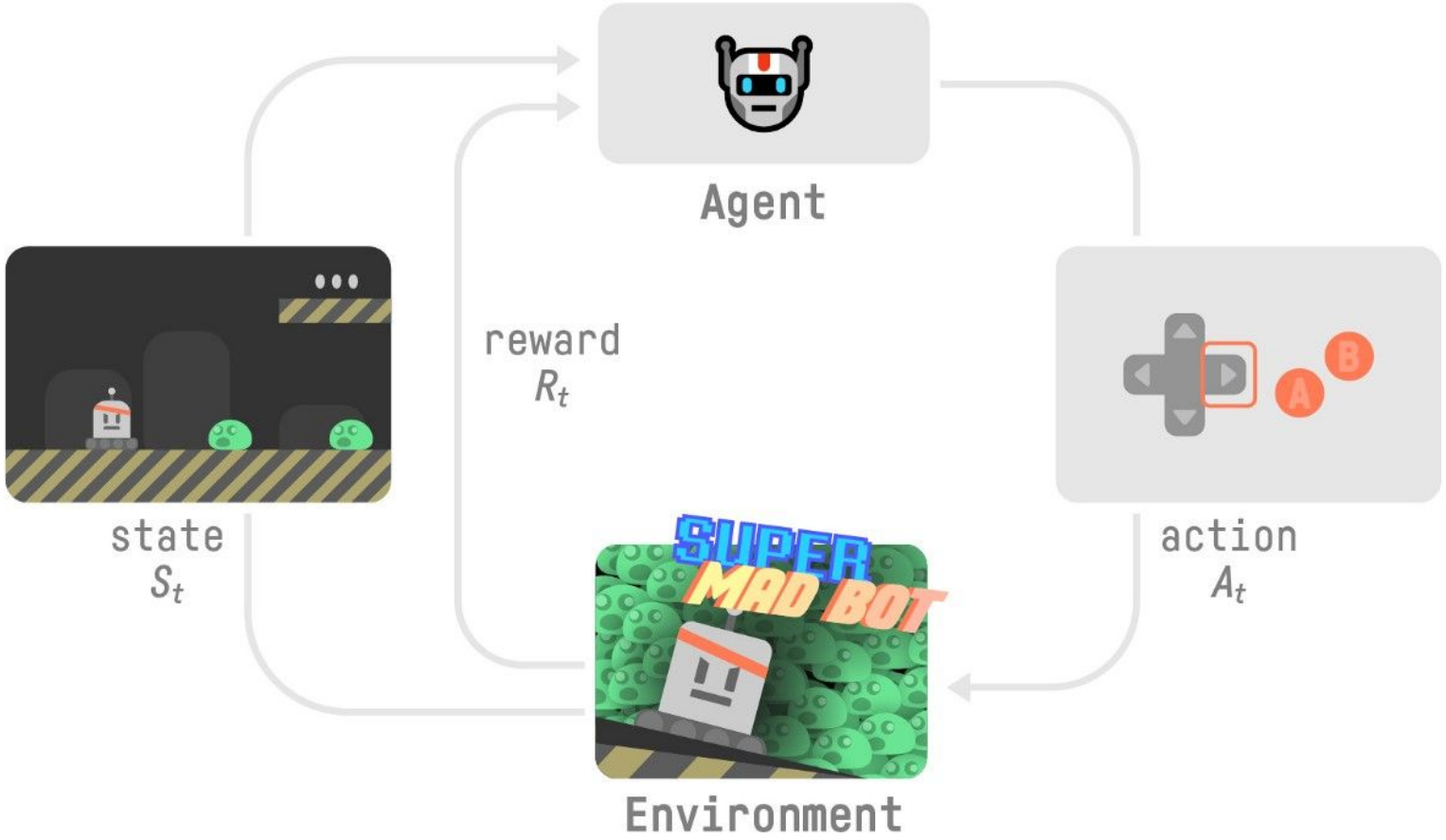
$$\mu = \sum_{i=1}^{n} p_i x_i \,.$$

You are given six dice.

Assume you cannot see the die faces. And the probability is not uniform.

Find the strategy to get maximum sum in 1000 rolls or fewer.

BASIC TERMINOLOGIES


Agent
reward
R_t
state
S_t
action
A_t
SUPER
MAD BOT
Environment

Terminologies

State (S) - The current situation or configuration of the environment.

Action (A) - A decision or move that the agent can take from a given state.

Policy (π) - An agent's behavior, specifying which action to take in each state. It can be deterministic (always choosing the same action for a given state) or stochastic (random).

Reward (R) - A scalar feedback received by the agent after taking an action in a state and transitioning into a new state.

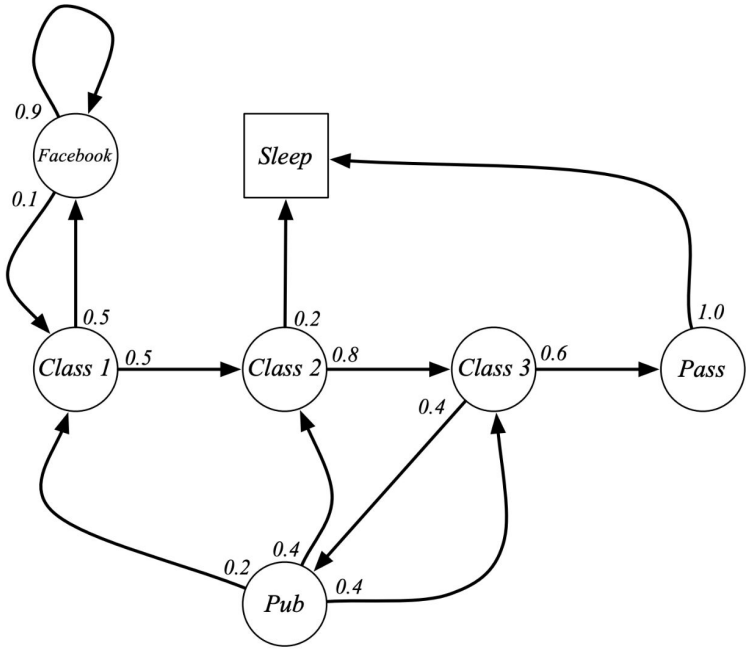
Transition Probability (P) - the probability of moving from current state s to new state s'.

Reward hypothesis

That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)

Markov Decision Process

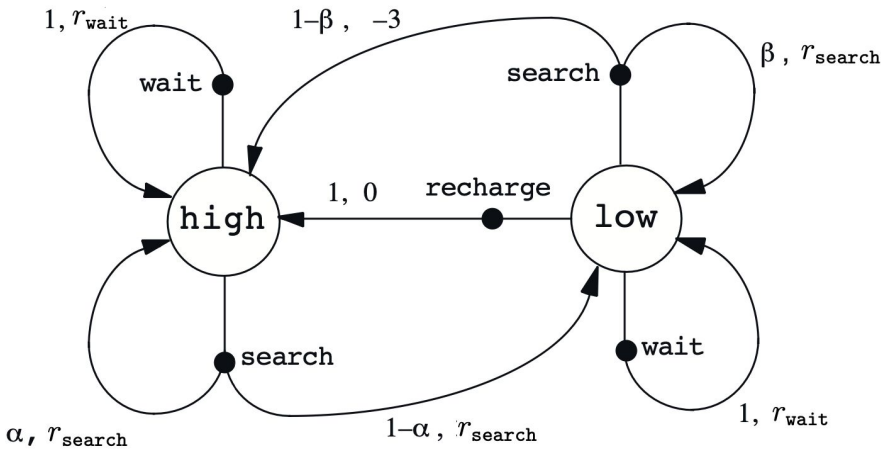
The future state depends only on the current state, not on the sequence of events that preceded it.



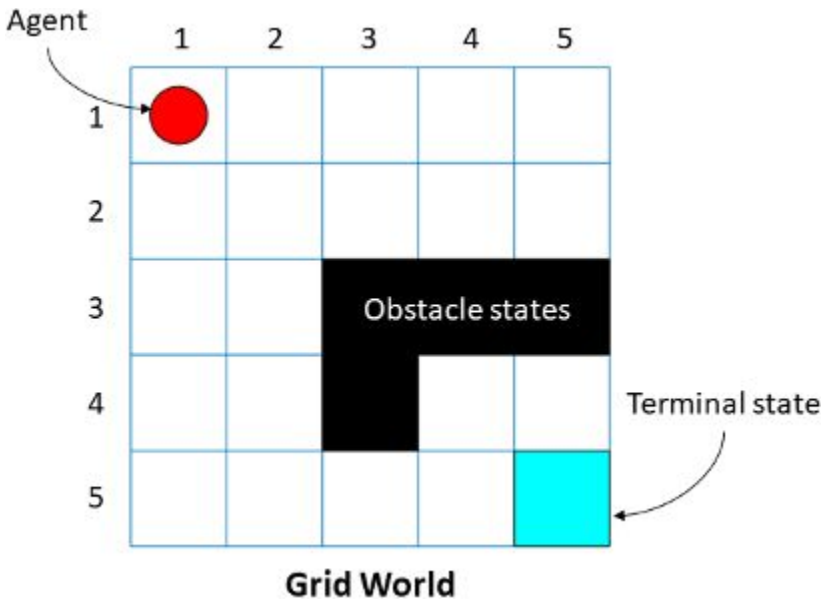
David Silver
https://www.davidsilver.uk/teaching/

Recycling Robot

s	s'	a	$p(s'\|s,a)$	$r(s,a,s')$
high	high	search	α	$r_{\texttt{search}}$
high	low	search	$1-\alpha$	$r_{\texttt{search}}$
low	high	search	$1-\beta$	-3
low	low	search	β	$r_{\texttt{search}}$
high	high	wait	1	$r_{\texttt{wait}}$
high	low	wait	0	$r_{\texttt{wait}}$
low	high	wait	0	$r_{\texttt{wait}}$
low	low	wait	1	$r_{\texttt{wait}}$
low	high	recharge	1	0
low	low	recharge	0	0.



Grid World


Agent
1
2
3
4
5
1
2
3
4
5
Obstacle states
Terminal state
Grid World

Policy

Policy defines the learning agent's way of behaving at a given time

Denoted as **π**

$$\pi(a|s) = p(a \mid s)$$

Probability of taking action a given state s.

Returns

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T,$$

Value function of state s

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s]$$