Reinforcement Learning

-Sarath Chandar

1. Introduction

1. Introduction

Artificial Intelligence (AI):

The goal of AI is to understand the general principles behind human intelligence and to simulate them in machines.

Machine Learning (ML): Takes the "learning" approach to solving AI.

ML involves designing algorithms that can learn through experience.

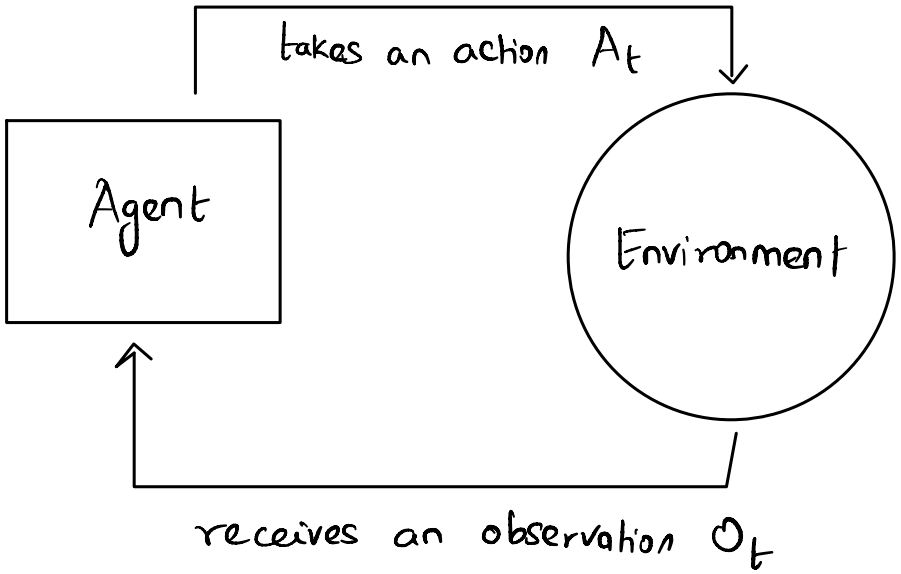
How do humans learn?

① We primarily learn by interacting with the world (Active learning).

② We also learn by reading, listening, observing (Passive learning).

While Supervised learning (SL) focuses on learning from data (passive), the field

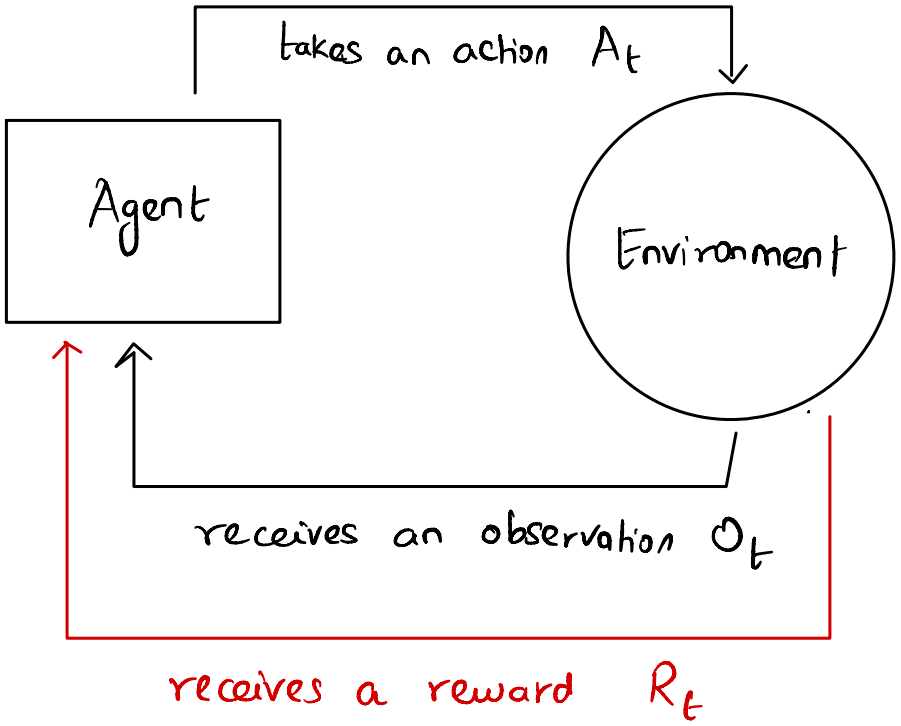
of Reinforcement Learning (RL) focuses on learning through interaction.



Any learning agent should have an objective. What is the objective of an RL agent?

The objective/goal for an RL agent is to learn how to map situations to actions - So as to maximize a numerical reward signal.

At each time step, the agent receives a numerical reward R_t.



The agent's goal is to maximize the total amount of reward it receives over time.

Example: Trash collecting robot. Every time the robot collects a trash, it receives +1. It receives zero reward for the rest of its time. It receives a reward of −100 if it runs out of battery.

An RL agent learning to maximize the total amount of reward it receives over time will learn to pick more trash efficiently

So that it can get a lot of +1 and also learn to go to charging station and recharge itself everytime the battery is low, so that it can avoid getting penalized with a -100 reward.

Look at how complex behaviours emerge based on this simple scalar reward structure!

Reward maximization is a very powerful learning objective.

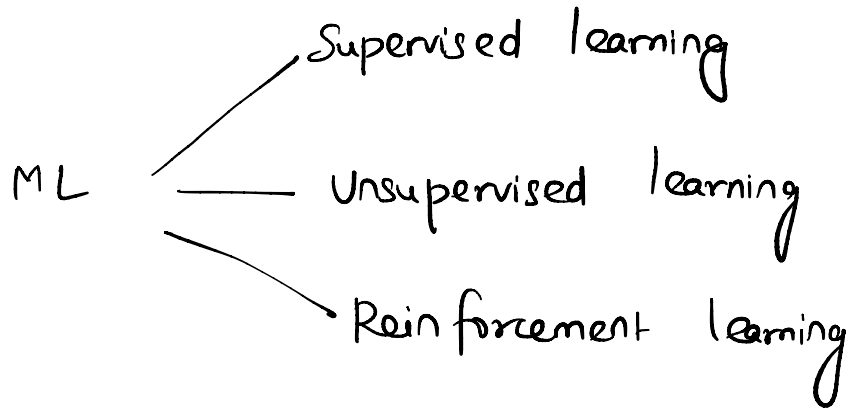
Reward hypothesis!-

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

Reward-is-Enough hypothesis :- (Silver et al. 2021)

Intelligence, and its associated abilities, can be understood as subserving the maximization of reward by an agent acting in its environment.

A quick detour to ML :-



Supervised learning :-

Given a training set $D_{Tr} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ of N examples, learn a function $f : x \to y$ such that given a new x, one can make an accurate prediction of the corresponding y.

ex: Given a movie review, predict whether the sentiment of the review is positive or negative.

x : review

y : +1 for positive review

-1 for negative review.

This is a "sentiment classification" task.

The focus of supervised learning is prediction.

Unsupervised learning :-

Given a dataset $D = \{x^{(i)}\}_{i=1}^{N}$ of N data points, find some structure in this data.

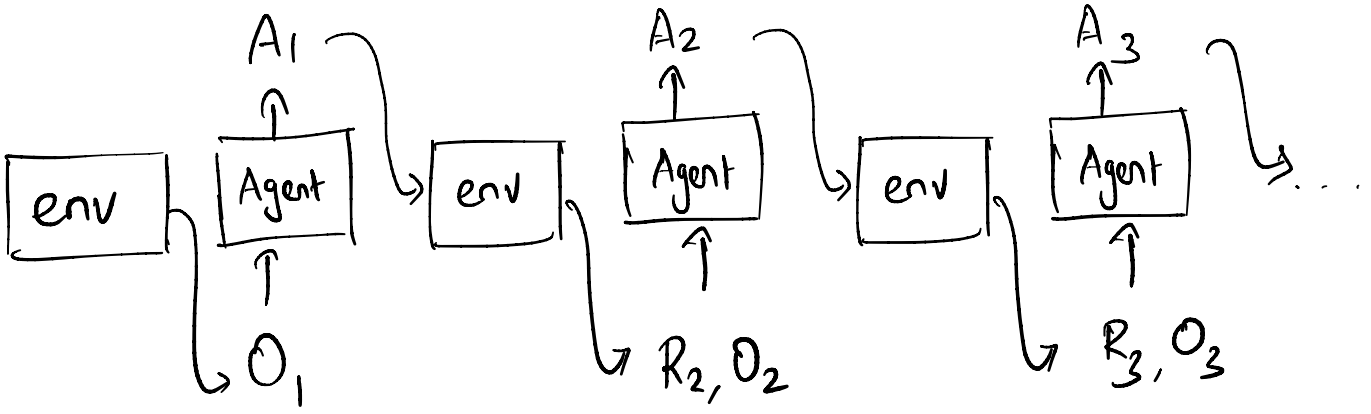
ex: Given a bunch of news articles, cluster them into several groups such that articles in same cluster talk about similar topic and articles in different clusters talk about different topics.

This is an example of clustering.

Reinforcement Learning :-

Reinforcement learning focuses on learning how to act. Specifically, given some

observation O_t, we are interested in finding the optimal action A_t such that the cumulative reward in the future is maximized. Note that the action A_t changes the environment and hence also affects the next observation O_{t+1}.



$O_1 \xrightarrow{\text{agent}} A_1 \xrightarrow{\text{env}} R_2, O_2 \xrightarrow{\text{agent}} A_2 \xrightarrow{\text{env}} R_3, O_3$

$\downarrow$ agent

A_3

$\downarrow$ env.

R_4, O_4

$\downarrow$

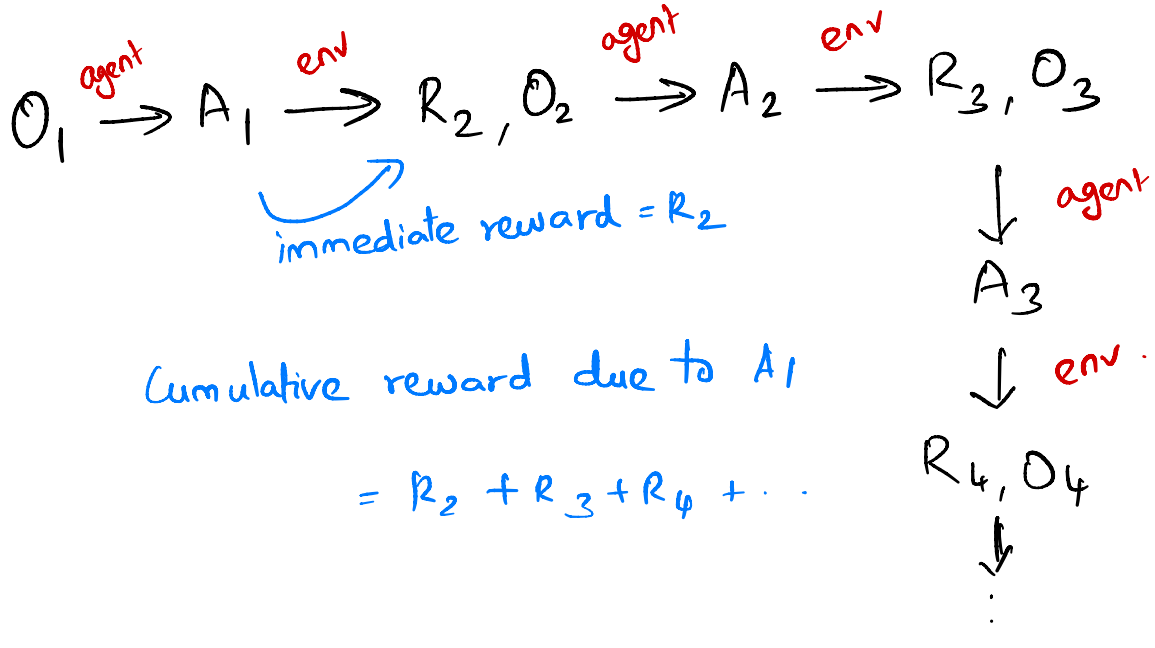
This is a sequential decision making process where time matters.

The agent's action A_t not only influences the immediate reward R_t but also influences the next observation O_{t+1} and hence all the future rewards $R_{t+1}, R_{t+2}, \ldots$

What are the implications of this closed loop interaction?

In Forward View:

The goal of the agent is to maximize its cumulative reward. So sometimes it needs to take an action that results in lower immediate reward but puts the agent in a state that it can receive higher rewards in the future.



Example 1: Studying every day after the lecture might not be attractive in terms of the immediate reward, but will be very useful to get a high cumulative reward (an A+ in the end of the semester).

Example 2: In the game of chess, sometimes players have to sacrifice an important piece to win the game.

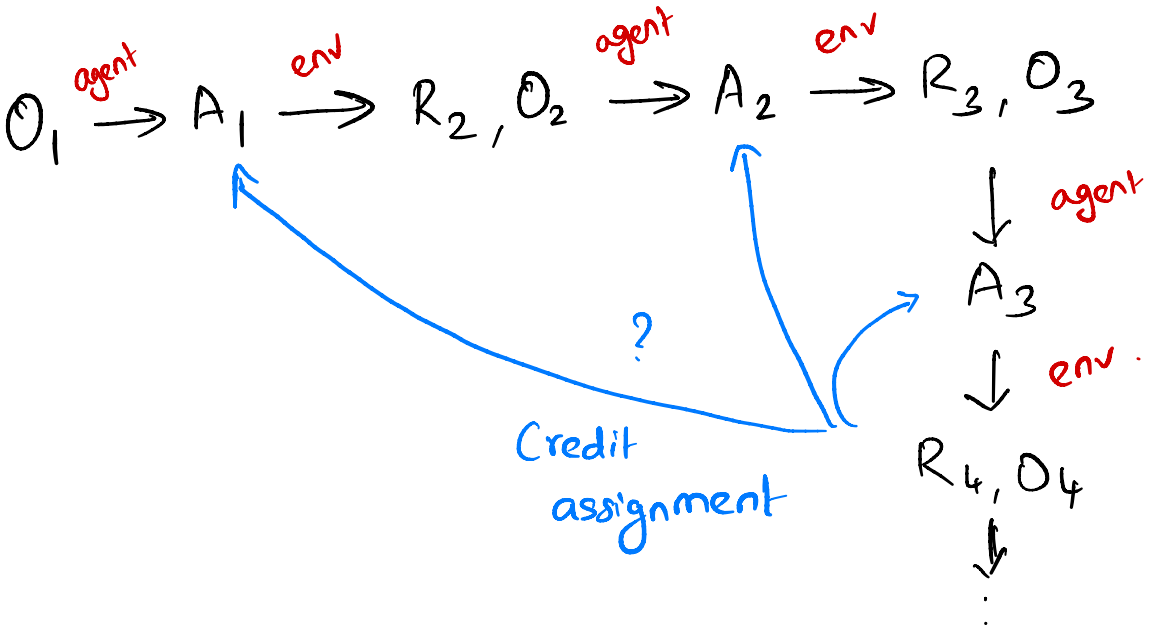
In backward view:-

One can also think about the implications in the backward view.

This means that the current reward R_t could be the result of any of the past actions $A_1, \ldots A_{t-1}$, not just the previous action A_{t-1}.

R_t could be a delayed reward for a past action.

Assigning credits to previous actions for the current reward is known as the problem of "credit assignment" — one of the central problems in RL.



Example: While playing chess, let us say that you won the match. Now which actions that you took in the entire match resulted in this win? This is a credit assignment problem.

Reinforcement learning (RL) Vs. Supervised learning (SL):-

① RL is learning through interaction while SL is learning from data.

② In supervised learning, the agent knows the right y for the x in the training data. But in RL, the agent is never told what is the right

action to take. The agent has to figure out the right or optimal action by trial and error.

(3) SL is all about prediction. But in RL, Prediction is just the means to Control the environment and hence achieve maximum Cumulative reward.

Reinforcement Learning vs. Unsupervised learning:-

(RL) (USL)

In USL, there is absolutely no supervision to the agent. However, in RL there is feedback to the agent in the form of scalar rewards.

Exploration vs. Exploitation dilemma:-

To obtain a lot of rewards, an RL agent must prefer actions that it has tried in the past and found to be effective in producing reward.

But to discover such actions, it has to try actions that it has not selected before.

The agent has to

→ exploit what it has already experienced in order to obtain reward.

→ also explore in order to make better action selections in the future.

It is not sufficient to just explore or just exploit. This dilemma of whether to explore or exploit is one of the unique challenges in RL that does not exist in SL/USL.

RL — Summary :-

* Agent view of AI
* Sequential decision making
* Delayed rewards and credit assignment.
* Exploration / Exploitation dilemma.