In the name of god

Quera talks

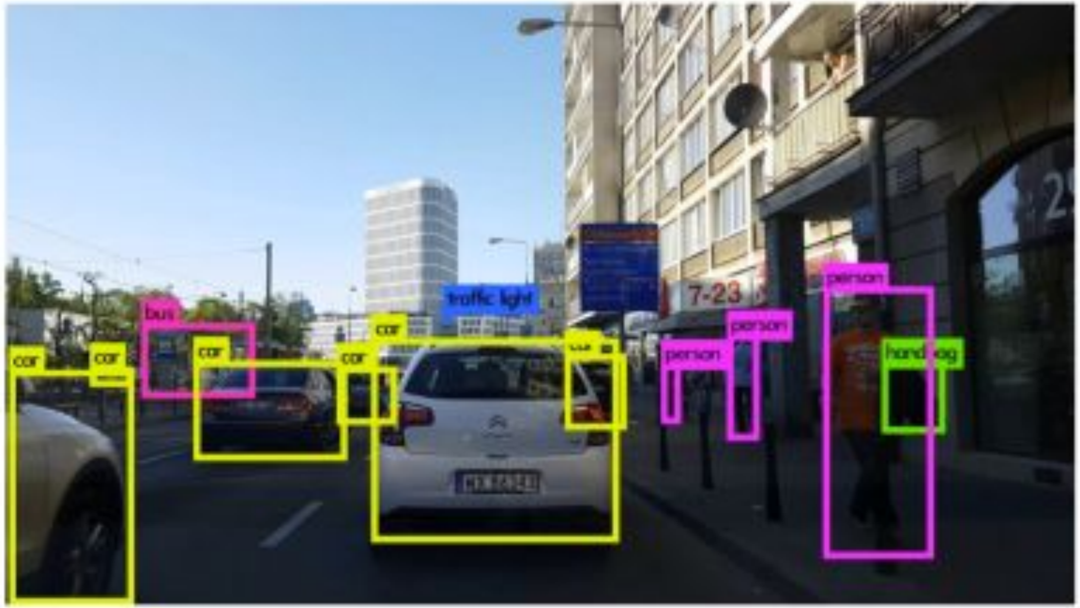
Reinforcement Learning

What is AI?

- Building computers that can act intelligently, like human! (simple definition)
- The science of building machine that

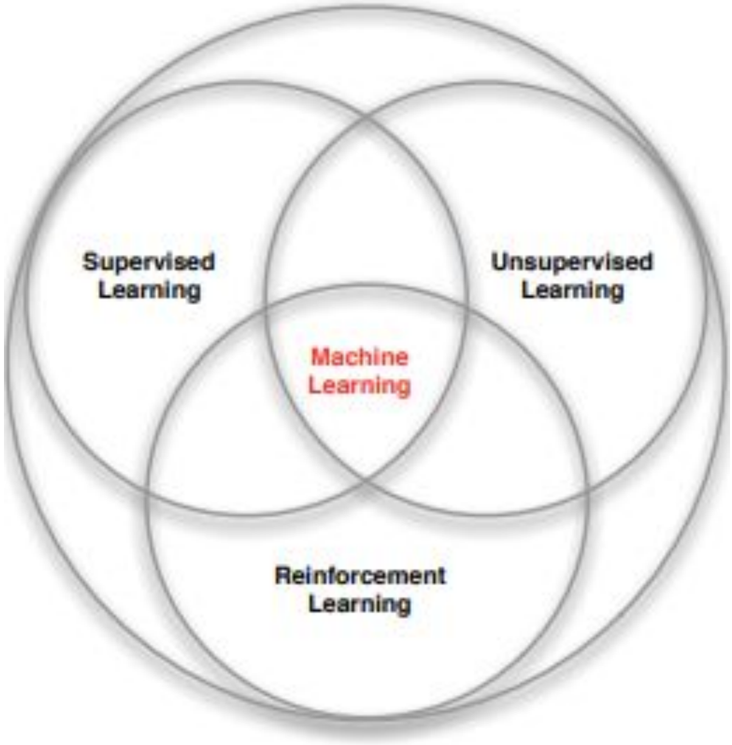
1. Think like human | Think intellectually
2. Act like human | Act intellectually

What is AI? (cont’d)

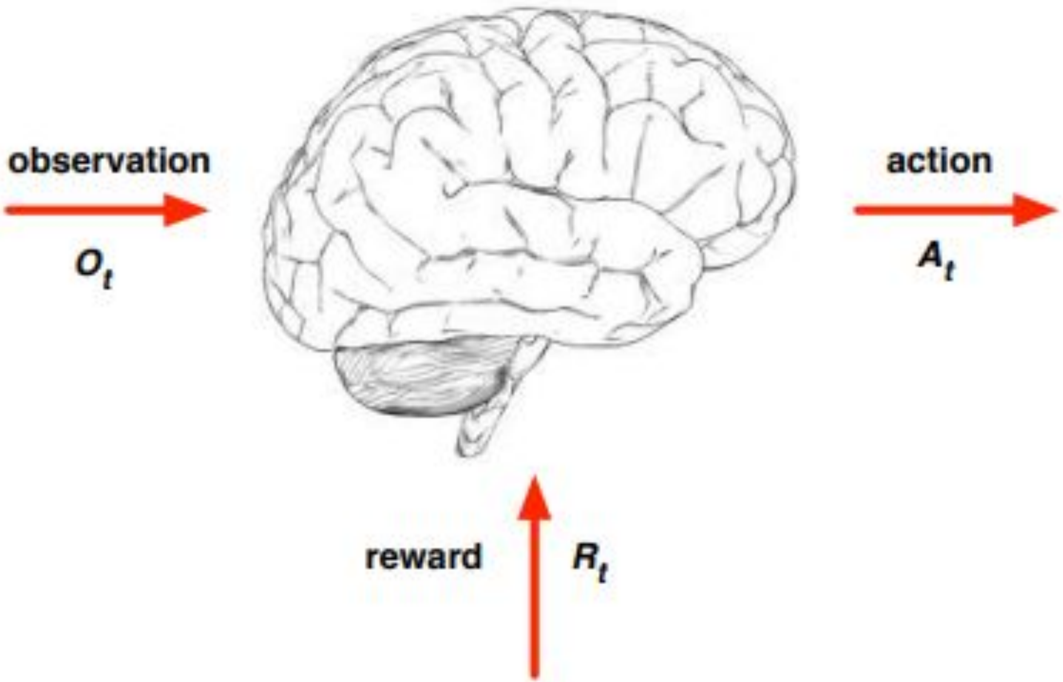


What is AI? (cont'd)

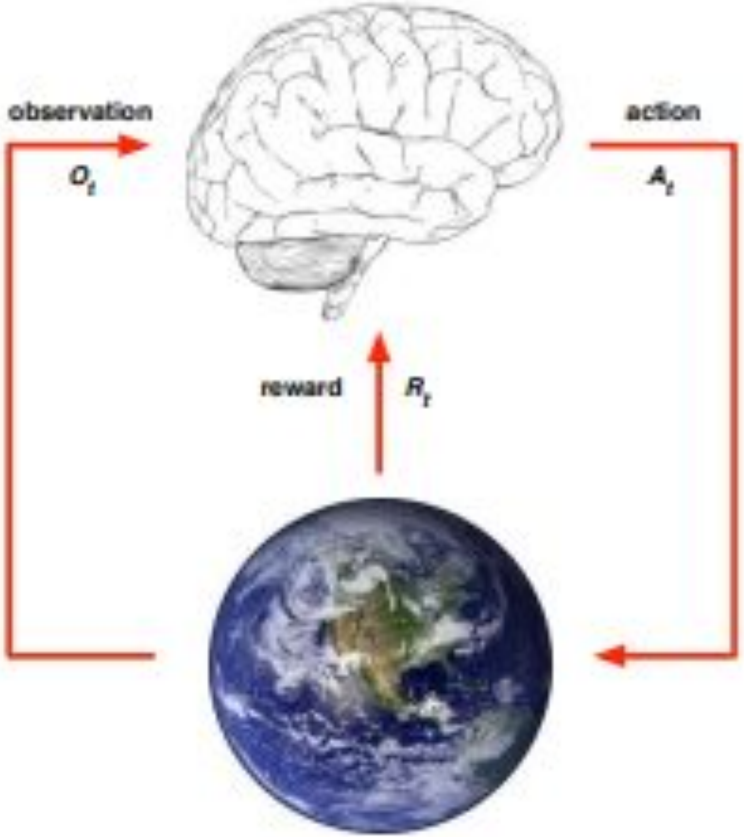
What is reinforcement learning?



What is reinforcement learning? Agent



What is reinforcement learning? Agent and Environment



Reward

- A reward R(t) is a scalar feedback signal
- Indicates how well agent is doing at step t
- The agent's job is to maximise cumulative reward

Reinforcement learning is based on the reward hypothesis.

Definition (Reward Hypothesis)

All goals can be described by the maximisation of expected cumulative reward.

History and state

The history is the sequence of observations, actions, rewards:

$H(t) = O1, R1, A1, ..., A(t-1), O(t) , R(t)$

- Environment state:
  - Data that environment uses to pick the next observation and reward
  - Usually not completely visible to agent
- Agent state:
  - $S(t) = f(H(t))$
  - Information used by RL agent to choose action

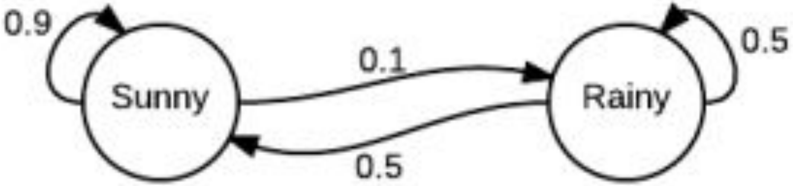
Markov state

- An information state that contains all useful information from history
- A state $S(t)$ is Markov if and only if

$$P[S(t+1) \mid S(t)] = P[S(t+1) \mid S1, ..., S(t)]$$

- The environment state in Markov

Markov process



$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

The weather on day 1 is known to be sunny. $\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 0 \end{bmatrix}$

The weather on day 2 can be predicted by: $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} P = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$

On day 3: $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} P = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$

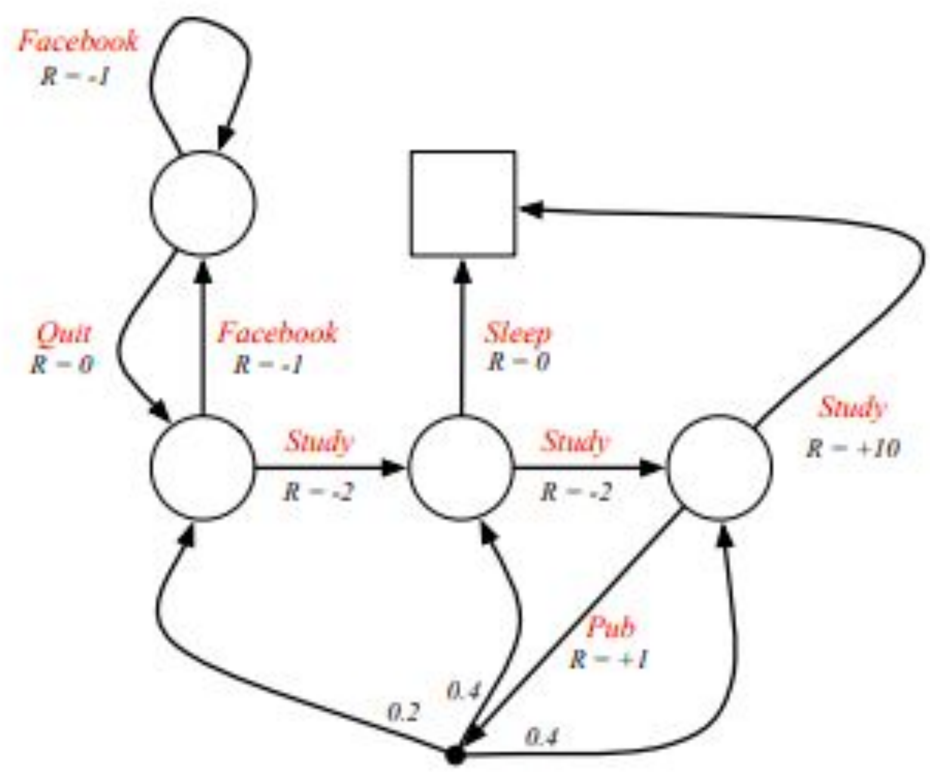
From wikipedia

Markov decision process (MDP)

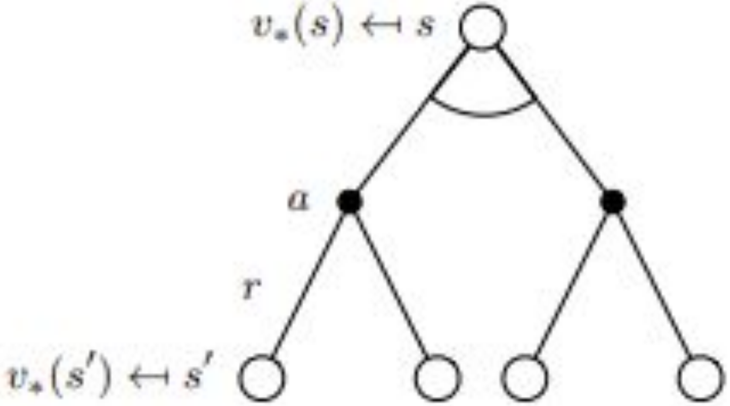
- A Markov Decision Process is a tuple (S, A,P, R, γ)
- S is a finite set of states
- A is a finite set of actions
- P is a state transition probability matrix
- R is a reward function
- γ is a discount factor γ ∈ [0, 1]

★ Almost all RL problems can be formalised as MDPs

Example: student MDP



Optimal state value



$$v_*(s) = \max_a \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_*(s')$$

Solving MDPs: value iteration

1) Start with v(s) = 0 for all s in S

2) For each step k+1: $v_{k+1}(s) = \max_{a \in \mathcal{A}} \left(\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k(s') \right)$

3) Continue to converge v*

4) Extract optimal policy from optimal values

Solving MDPs: policy iteration

1) Supposing a deterministic policy, evaluate that policy (calculate values applying that policy)

2) Improve policy: using calculated values, improve policy

3) Continue until policy converges

Reinforcement Learning and Markov Decision process

Same process but:

1) We don’t already know transition probabilities anymore!

2) We don’t already know rewards anymore!

★ We must explore and learn them ourselves.

Major Components of an RL Agent

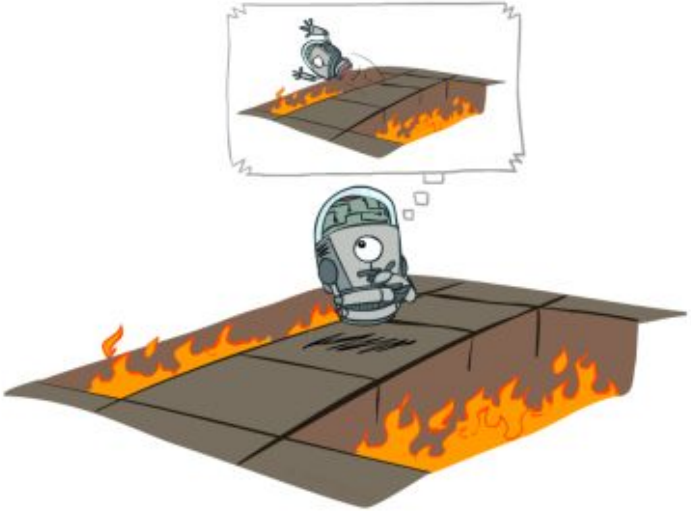
An RL agent may include one or more of these components:

❏ Policy: agent’s behaviour function

❏ Value function: how good is each state and/or action

❏ Model: agent’s representation of the environment

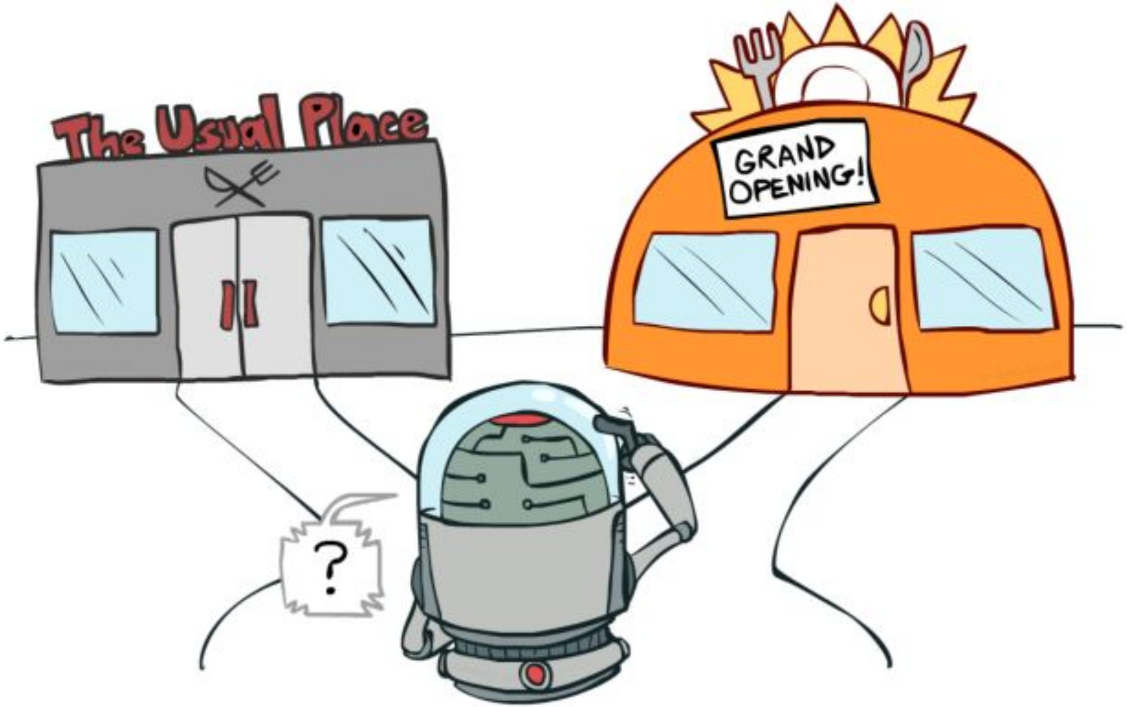
RL (online) vs MDP (offline)



Offline راه‌حل

Online یادگیری

Exploration vs Exploitation



Question?

Confucius:

The man who asks a **question** is a fool for a minute, the man who does not **ask** is a fool for life.

References

- Mr David Silver RL course
- Mr Mohammad Taher PilehVar AI course at IUST

Thanks