Reinforcement Learning (RL)

1. Khái niệm

Tại sao không dùng Supervised Learning?

- Ý tưởng: dùng dữ liệu trạng thái → hành động từ object để huấn luyện mạng Neural
- Nhưng thực tế:
  - Khó xác định "hành động tốt nhất" trong từng trạng thái.
  - Nhiều hành động đều "hợp lý", nên dữ liệu huấn luyện không rõ ràng.

⇒ **Supervised learning không phù hợp**, cần phương pháp khác

RL giống như huấn luyện chó

- **Không cần dạy chi tiết**, chỉ cần đưa ra phần thưởng hoặc phạt:
  - Làm tốt → khen: "good dog"
  - Làm sai → phạt: "bad dog"
- Mạng Neural sẽ **tự khám phá** ra hành động nào nên làm.

Ý tưởng cốt lõi của RL

- Không cần chỉ rõ **hành động đúng** cho từng đầu vào.
- **Chỉ cần định nghĩa hàm phần thưởng**: làm gì được khen → RL sẽ tự học cách tối ưu hành động để nhận thưởng.

2. Return

- Trong Reinforcement Learning, **Return** là tổng các phần thưởng (rewards) mà tác nhân nhận được sau một chuỗi hành động.
- Tuy nhiên, thông thường thì **phần thưởng gần (về mặt thời gian) thì giá trị hơn** phần thưởng ở tương lai xa.
- Vì vậy, Return được tính bằng cách **giảm trọng số** phần thưởng trong tương lai bằng một **hệ số chiết khấu** (discount factor), ký hiệu là γ (gamma), nằm

trong khoảng [0, 1].

Ý nghĩa của discount factor (γ)

- γ càng gần 1 → hệ thống càng **kiên nhẫn**, coi trọng phần thưởng xa.
- γ càng nhỏ → hệ thống **thiên về phần thưởng gần**.
- Trong tài chính, γ phản ánh **giá trị thời gian của tiền** (time value of money).

Trường hợp phần thưởng âm (negative rewards)

- Discount factor sẽ **khuyến khích hoãn phần thưởng âm càng lâu càng tốt**.

3. Policy

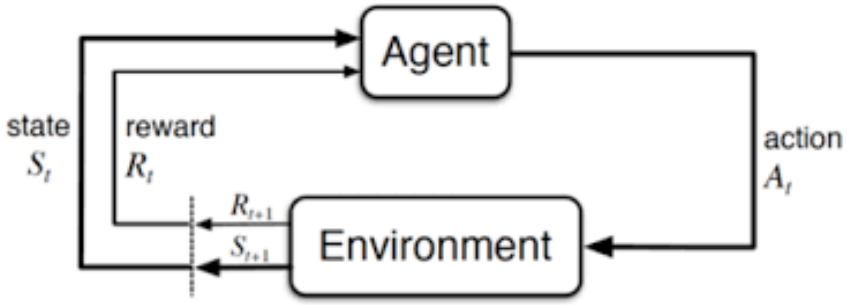
RL (Reinforcement Learning) cần một **hàm quyết định** hành động nên làm trong từng trạng thái cụ thể, hàm này được gọi là **Policy**, ký hiệu là **π (pi)**.

$$\pi(s) = a$$

Trong đó:

- `s` là trạng thái hiện tại.
- `a` là hành động mà policy chọn trong trạng thái đó.
- Mục tiêu của RL là tìm ra một policy **tối ưu**, tức là policy giúp **tối đa hóa tổng phần thưởng (return)** trong suốt quá trình hành động.

4. Tổng kết - Mô hình chung (MDP – Markov Decision Process)



- **State (Trạng thái)**

Mô tả đầy đủ tình hình hiện tại

- **Action (Hành động)**

  Tập các lựa chọn agent có thể thực hiện

- **Reward (Phần thưởng)**

  Giá trị (số thực) phản ánh mức "tốt/xấu" của hành động vừa thực hiện ở trạng thái cũ

- **Discount factor (γ)**

  Hệ số chiết khấu $\in(0, 1)$: làm giảm giá trị các phần thưởng ở tương lai xa hơn so với phần thưởng hiện tại.

- **Return (G_t)**

  Tổng phần thưởng có chiết khấu:

$$G_t = R_{t+1} + \gamma\, R_{t+2} + \gamma^2\, R_{t+3} + \dots$$

- **Policy (π)**

  Hàm ánh xạ từ trạng thái → hành động:

$$\pi(s) = a$$

  Mục tiêu RL: **tìm policy tối ưu** sao cho tổng Return là lớn nhất.