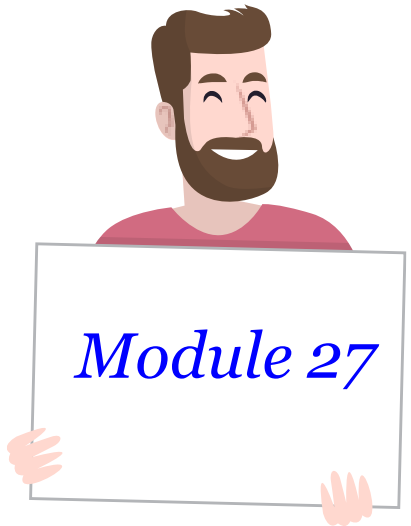
# 強化學習介紹




designed by freepik

Estimated time:
45 min.

# 學習目標

- 27-1:強化學習
- 27-2:價值函數
- 27-3:強化學習的種類



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 27-1:強化學習

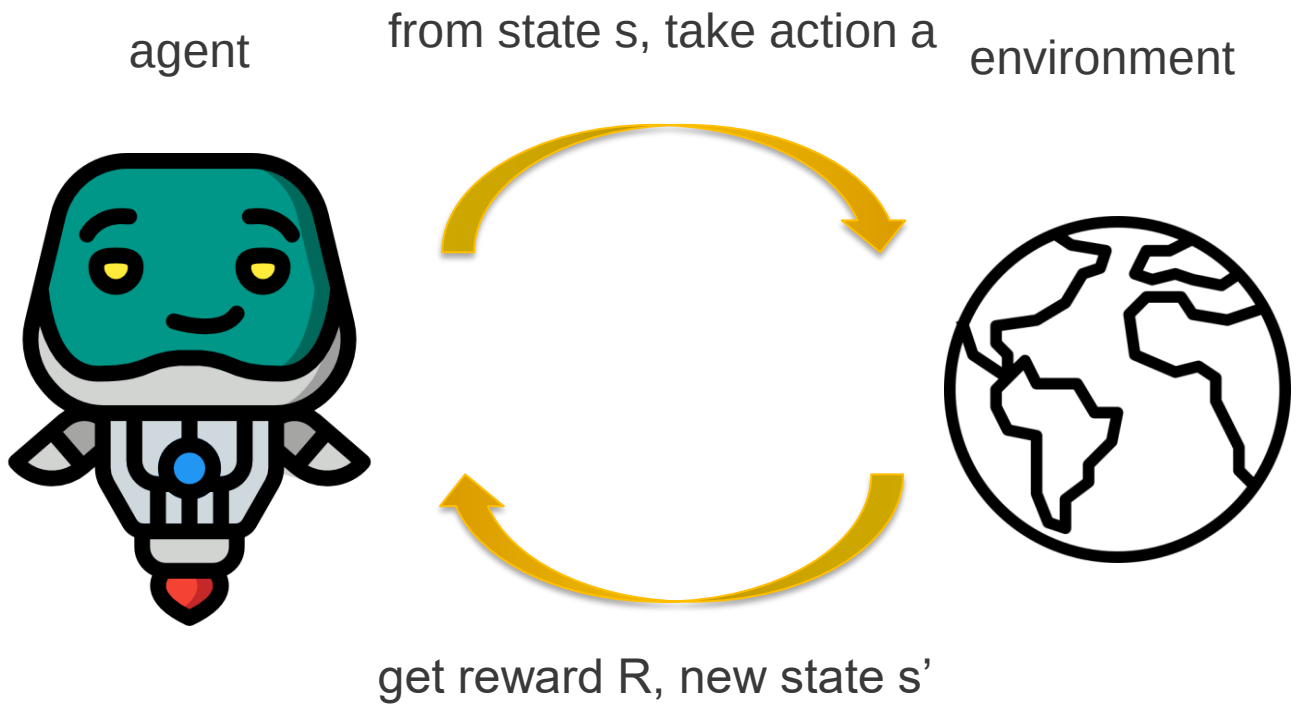
- 強化學習介紹
- 強化學習目標




財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 強化學習介紹

- 強化學習是機器學習的一個分支
  - 主要在探討如何讓**agents**去採取動作與環境互動並極大化未來累積賞籌






財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 強化學習介紹

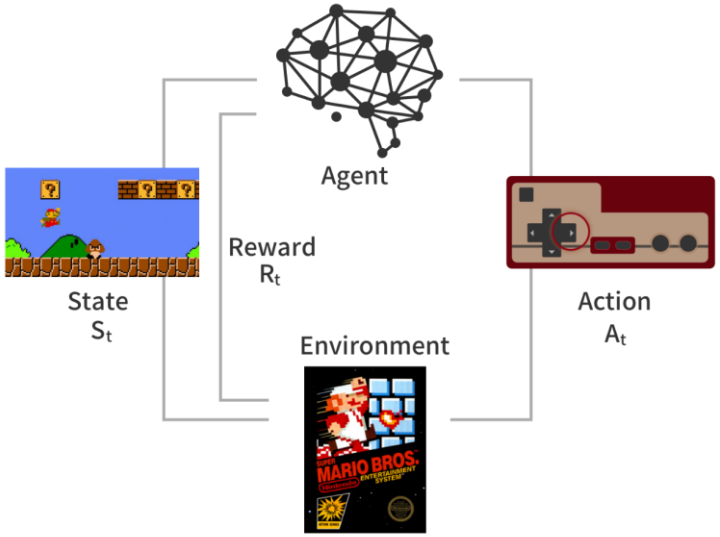
- 在強化學習裡面，Agent智能體會做動作，而環境會給Agent新的狀態state以及賞酬reward




財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 強化學習目標

- 強化學習的目標就是要找到一個policy使得未來累積賞酬最大
  - Policy是一個函數，其輸入為狀態，輸出為動作
  - Policy常常用$\pi(s)$表示



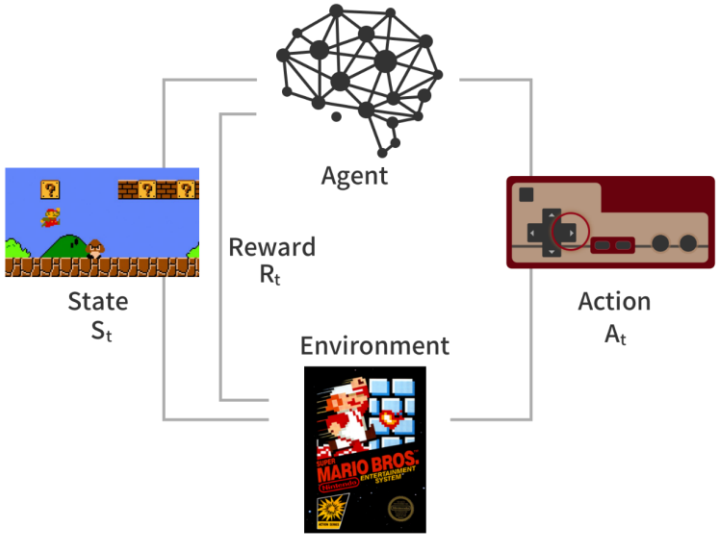

$$S_1, A_1, R_2, S_2, A_2, \ldots, S_T$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 強化學習目標

- 又或者是說，強化學習要找到一個序列S1, A1, R2, S2, A2,……
  - 使得未來累積賞酬最大
  - S代表狀態、A代表動作、R代表賞酬



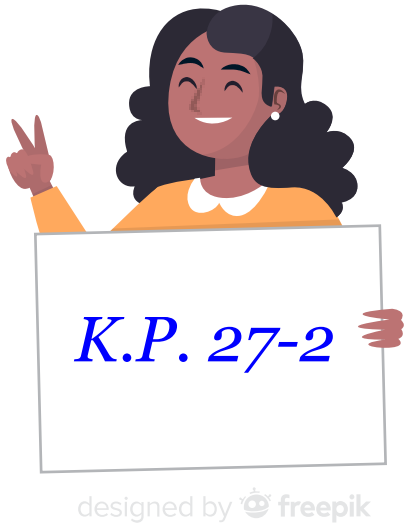

$$S_1, A_1, R_2, S_2, A_2, \ldots, S_T$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 27-2:價值函數

- Policy介紹
- Value Function介紹
- Value Function種類






財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Policy介紹

- Policy $\pi(s)$是一個函數告訴智能體接下來要做什麼動作
  - 輸入是狀態s、輸出是動作a
  - 有分deterministic或是stochastic

Deterministic: $\pi(s) = a$.

Stochastic: $\pi(a|s) = \mathbb{P}_\pi[A = a|S = s]$.



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Value Function介紹

- **Value function是一種可以衡量當下狀態好與壞的函數**
  - **衡量標準是依據預測未來累積賞酬大不大**
  - **未來累積賞酬 = 接下來所有時間點賞酬的總和**

$$G_t = R_{t+1} + R_{t+2} + \dots$$

$$\| \quad \|$$

$$G_t = \sum_{k=0}^{T} R_{t+k+1}$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Value Function介紹

- 離現在越遠的賞酬應該要被打折
  - 因為越遠的事情應該是越不確定的
  - 未來的賞酬無法得到立即回饋
- 基於以上理由，專家門將未來賞酬設計為離越遠就需要乘以越多次折現因子$\gamma$

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots = \sum_{k=0}^{\infty} R_{t+k+1}$$

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Value Function種類

- **Value Function有分成兩個種類**
  - **State value function，輸入是狀態，輸出是衡量狀態好壞的數值**
  - **Action value function ，輸入是狀態及動作，輸出是衡量狀態好壞的數值**

$$V_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

state-value function

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

action-value function



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Value Function種類

- **State value function跟Action value function的關係如下**
  - **這兩個是有辦法做轉換的**

$$V_\pi(s) = \sum_{a \in A} Q_\pi(s,a)\pi(a|s)$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Value Function種類

- 當我們得到一個Value Function，我們可以根據以下關係式子去推導最優Value Function以及最優Policy

最優Value Function

$$V_*(s) = \max_\pi V_\pi(s), Q_*(s,a) = \max_\pi Q_\pi(s,a)$$

最優Policy

$$\pi_* = \arg\max_\pi V_\pi(s), \pi_* = \arg\max_\pi Q_\pi(s,a)$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 27-3:強化學習的種類

- Episodic task與Continuing task
- Monte Carlo與TD
- 不同種類的強化學習




財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Episodic task與Continuing task

- **Episodic task**
  - 有起始也有結束明確點的工作，例如破關遊戲
- **Continuing task**
  - 這個遊戲永遠會持續下去
  - 智能體將持續做決策直到我們將其停止



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Monte Carlo與TD

- 我們可以根據更新智能體的方式分成下面兩類
  - **Monte Carlo**，每次遊戲結束時才會計算一次期望未來總賞酬
  - **TD(Temporal Difference Learning)**，在每一步做決策的時候都會計算一次賞酬



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Monte Carlo與TD

- 可以看到以下更新Value Function的式子
  - Monte Carlo是遊戲結束後才會更新一次
  - TD則是變執行決策邊更新

Monte Carlo $$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

TD Learning $$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 不同種類的強化學習

- 在強化學習裡面，有三種不同解決強化學習的方法
  - **Value based，其目標是找到最優的value function，在反推policy**
  - **Policy based，直接去找尋最優的policy**
  - **Model based**

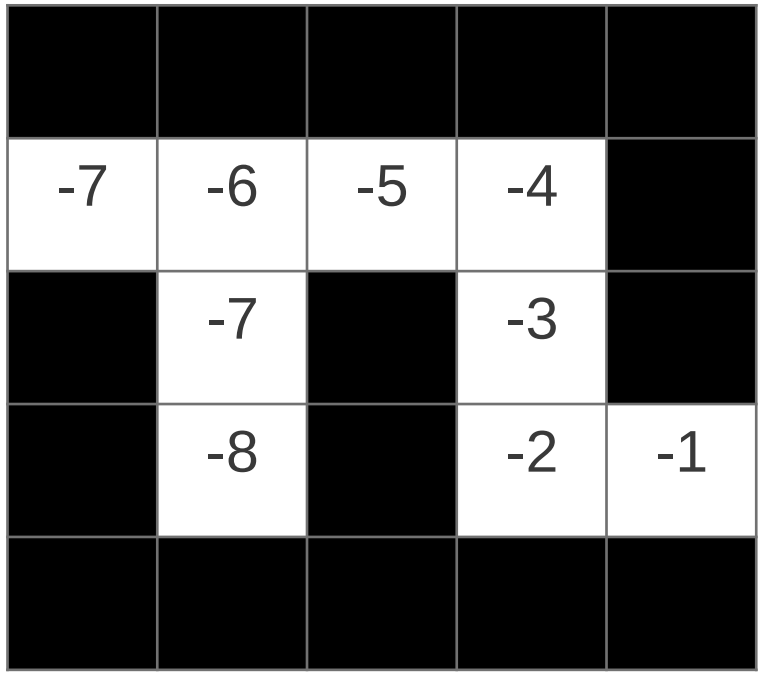


財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 不同種類的強化學習

- 以下是Value based的示意圖
  - 可以看到每個狀態都有一個數值，代表那個狀態的好壞

start



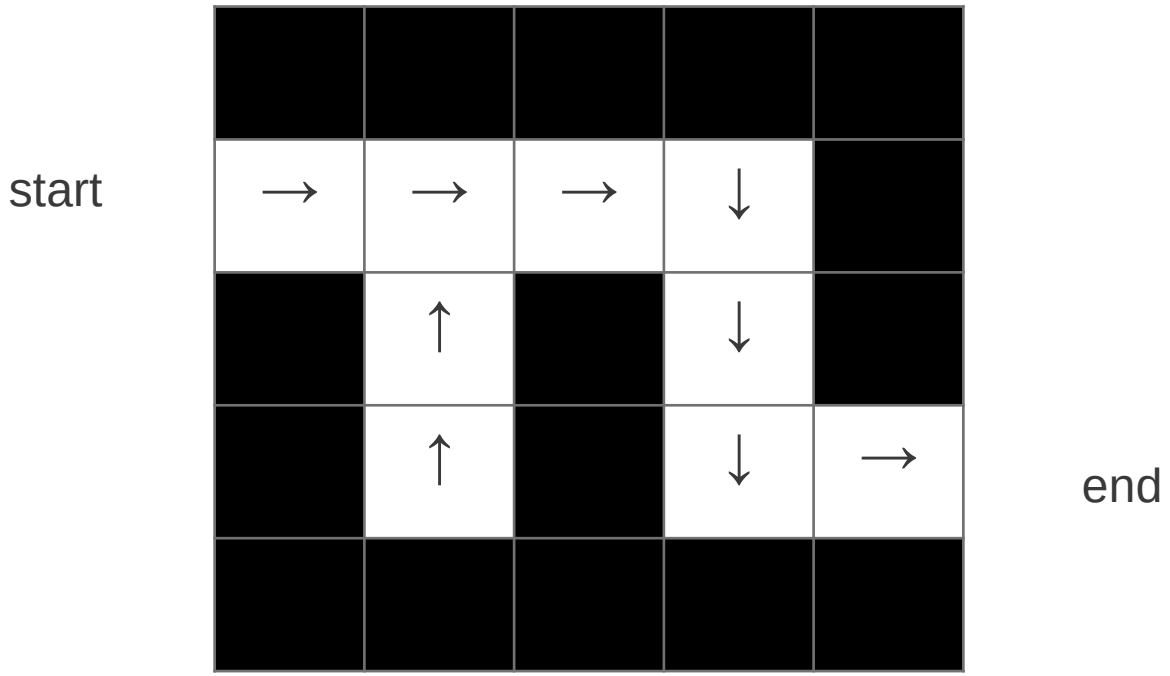

end



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 不同種類的強化學習

- 以下是Policy based的示意圖
  - 可以看到每個狀態都直接對應一個動作

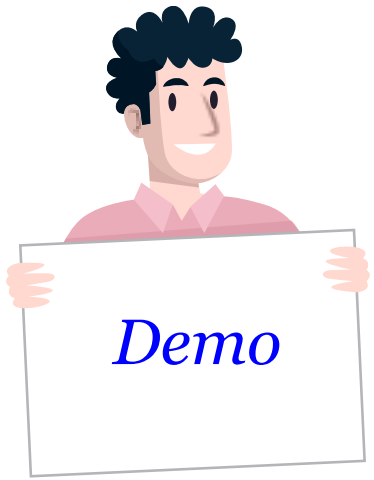





財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Demo 27-3

- 計算未來累積賞酬
- 調高衰減因子
- 調低衰減因子




designed by freepik

# 線上Corelab

- 題目1：計算未來累積賞酬
  - 給予一個未來賞酬的list，計算當下的賞酬數值
- 題目2：計算未來累積賞酬
  - 給予一個未來賞酬的list並設定衰減值為0.8，請計算當下的賞酬數值
- 題目3：計算未來累積賞酬
  - 給予一個未來賞酬的list並設定衰減值為0.4，請計算當下的賞酬數值



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# 本章重點精華回顧

- 強化學習的概念
- 價值函數
- 強化學習的種類



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY

# Lab:Python 簡介

- **Lab01:計算未來累積賞酬**
- **Lab02:調高衰減因子**
- **Lab03:調低衰減因子**

Estimated time:
20 minutes



財團法人資訊工業策進會
INSTITUTE FOR INFORMATION INDUSTRY


THANK YOU