REINFORCEMENT LEARNING OF PACMAN FINAL PROJECT

陳亭禎、郭冠宏、李宗穎

授課教師：曾士桓 博士

TEAM09

電腦與通訊工程系

2022.01.10

BACKGROUND OR TREND

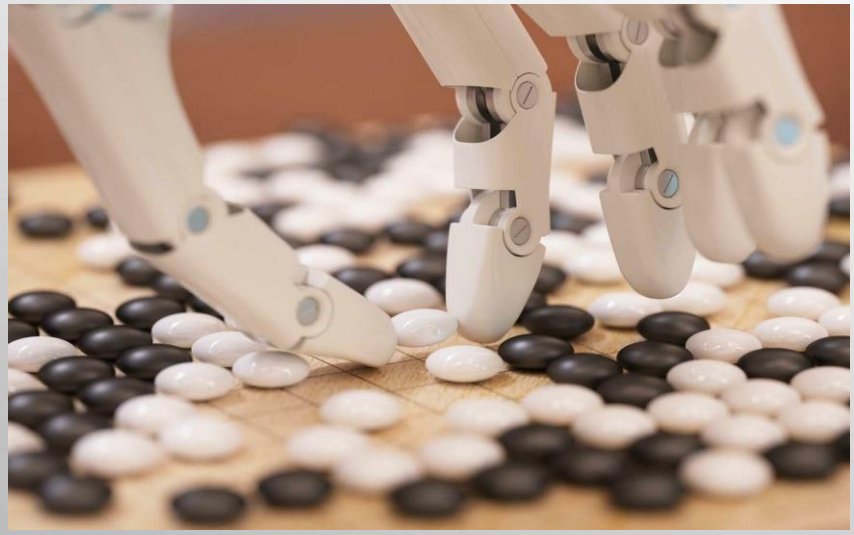
- **發現問題:如何使電腦擁有自我學習的能力？**
  - 於期中前所學：
    - ✓ Informed Search
    - ✓ UnInformed Search

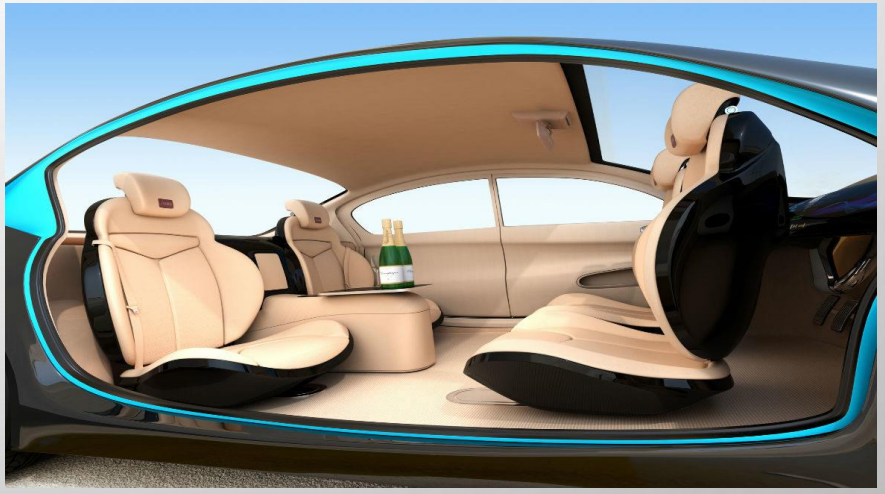
  還不算是讓電腦擁有自我學習的能力

    - ✓ BerkeleyCS188 Pacman程式中Project3 Reinforcement learning的部份當成我們期末專題的實作內容。

MOTIVATION

- **根據前頁的問題，說明解問題的方向**
  - 近些年因人工智慧(ALPHAGO、DRIVERLESS CAR)應用的風行，許多人工智慧相關的應用也接連不斷的出現於生活中，而我們團隊在本次的期末專題採用強化學習的動機主要有兩大重點：



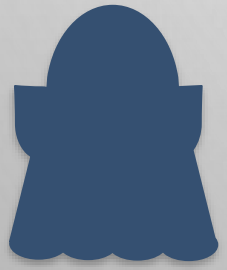


- 圖片來源：https://reurl.cc/g0WRKX、https://reurl.cc/MbA3rk

MOTIVATION

- **根據前頁的問題，說明解問題的方向**
  - 近些年因人工智慧(ALPHAGO、DRIVERLESS CAR)應用的風行，許多人工智慧相關的應用也接連不斷的出現於生活中，而我們團隊在本次的期末專題採用強化學習的動機主要有兩大重點：

1. 利用強化學習作出行為(action)與環境(Environment)互動接收不同的反饋(獎勵、懲罰)來對未知環境做出最好(optimal)的決策。



怪物抓到吃豆人，給予獎勵
怪物沒抓到吃豆人，給予懲罰

吃豆人吃到豆子，給予獎勵
吃豆人被怪物抓到，給予懲罰

OBJECTIVE

- **根據動機，說明解問題的方式**

2. 大家可能會問? 那為何不用Supervised learning? Unsupervised learning?

監督式學習---有特定答案讓電腦學習

身高	體重	性別
162	60	女
172	60	男
172	55	男

身高	體重	性別
152	48	?

非監督式學習---沒有特定答案讓電腦學習，從特徵中尋找關聯

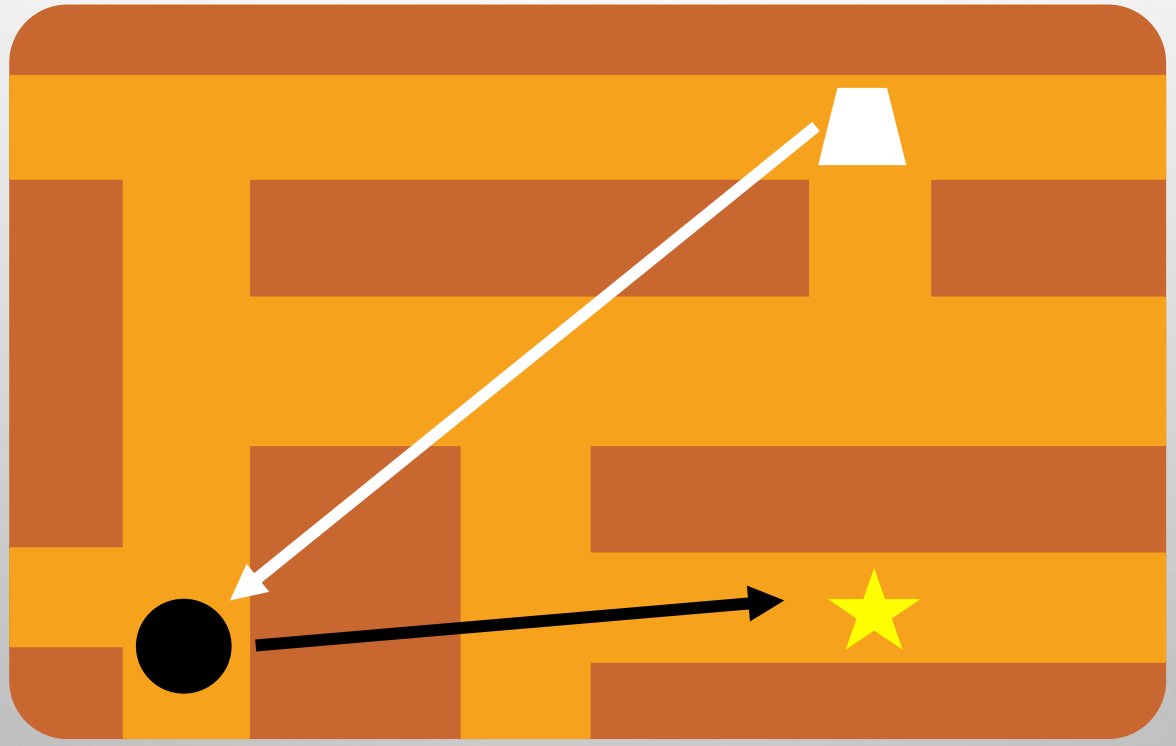
數學	物理	國文	英文
76	89	36	42
88	92	89	70
52	35	25	40
27	15	82	79

數學好　可能　物理好
國文好　可能　英文好

OBJECTIVE

- **根據動機，說明解問題的方式**

強化式學習---根據所處環境學習，無特定特徵，只設定所需達成目標，並建立獎懲機制以利學習



CHALLENGE

- **這類問題有什麼普遍的挑戰或解這問題方式的挑戰**
  - 如何定義獎懲機制？

Agent	Event	Reward
吃豆人	吃到豆子	Score加10分
	吃完豆子	Score加500分
	被怪物抓到	Score減500分
	每經過一個Action	Score減1分
怪物	抓到吃豆人	怪物Score加500分
	每經過一個Action，且沒有抓到吃豆人	怪物Score減1分

CHALLENGE

- **這類問題有什麼普遍的挑戰或解這問題方式的挑戰**
  - 達成目標後，如何判斷是不是最佳的？

    ✓利用各別Agent的score判斷， score越高，訓練效果越好。
  - 如何互換怪物以及吃豆人的學習成果，並加以訓練？

    ✓怪物和吃豆人的學習是同步的，故無交換之可能。
  - 是否好收斂？

    ✓原先使用Q-Learning發現成效不彰，故使用Approximate Q-Learning。

POTENTIAL SOLUTIONS

- **你會採取的演算法、技術等等**
  - 演算法
    - 吃豆人--- Q-LEARNING
    - 怪物--- SARSA

SOLUTIONS

- **你會採取的演算法、技術等等**
  - 演算法
    - 吃豆人---APPROXIMATE Q-LEARNING
    - 怪物--- APPROXIMATE SARSA
  - 繪圖技術
    - 吃豆人、豆子---CIRCLE( )函式繪製
    - 怪物---POLYGON( )函式繪製
    - 初始畫面---PYGAME函式庫、LABEL( )建立MENU

ALGORITHM

- **Q-LEARNING / SARSA**

  演算法方面我們將會使用Q-LEARNING 及SARSA來實作。

  首先兩演算法內部基礎概念來源於TD(TEMPORAL DIFFERENCE)-UPDATE RULE，由下圖所示：

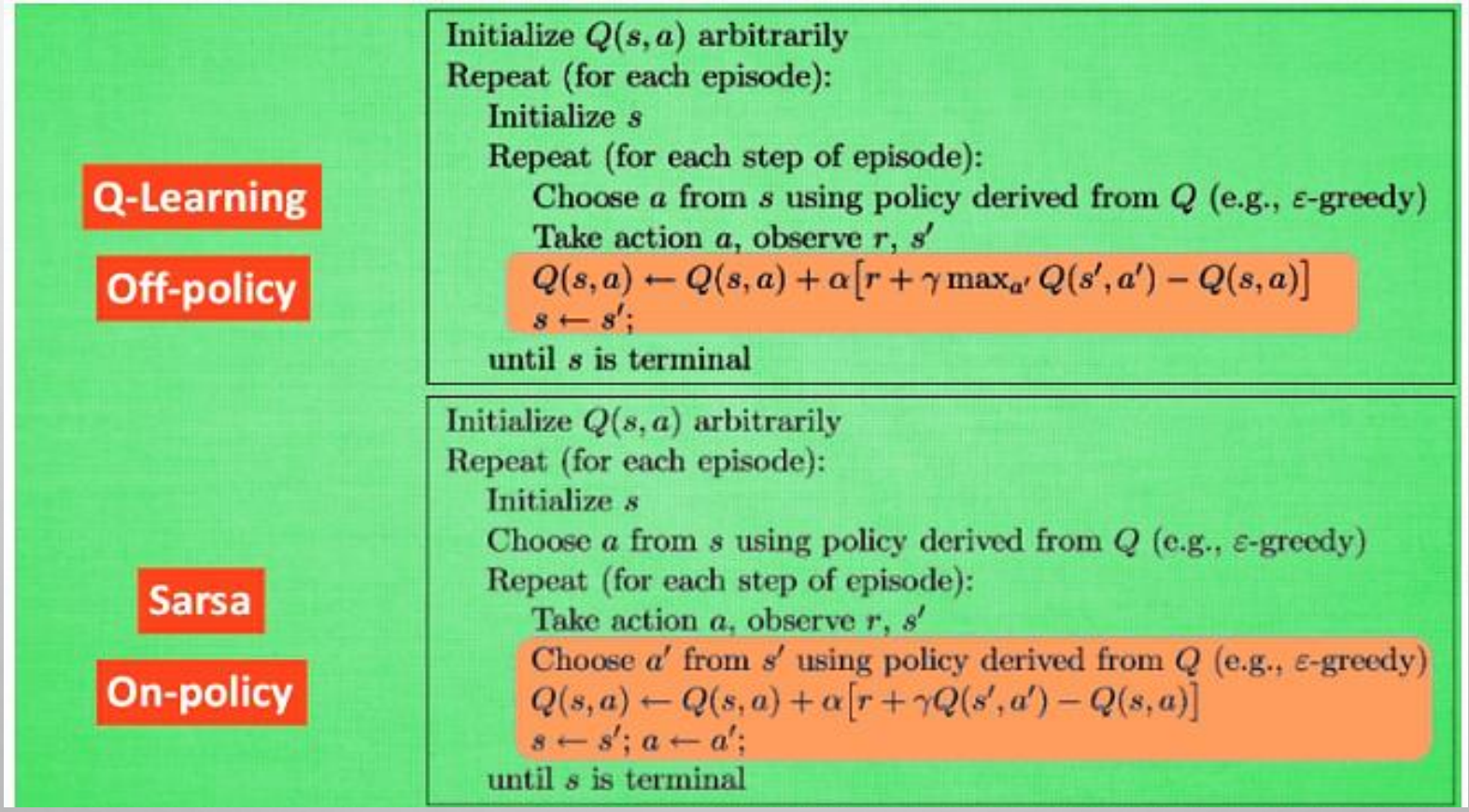
$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$$

  藉由應用該公式，我們能透過程式在每一次的STATE-TRANSITION中進行Q-VALUE的更新，使得Q-VALUE能夠被逐漸訓練成在PACMAN遊戲中能表現出最好(OPTIMAL)成效的Q-VALUE。而實作過程中，我們也會應用到EPSILON GREEDY以機率決定AGENT在STATE上擁有隨機選擇ACTION的能力能夠來探索地圖世界。

ALGORITHM

- **Q-LEARNING / SARSA**
  - 圖片來源：https://reurl.cc/5GrRgG

對各別AGENT(怪獸及吃豆人的)演算法差異說明



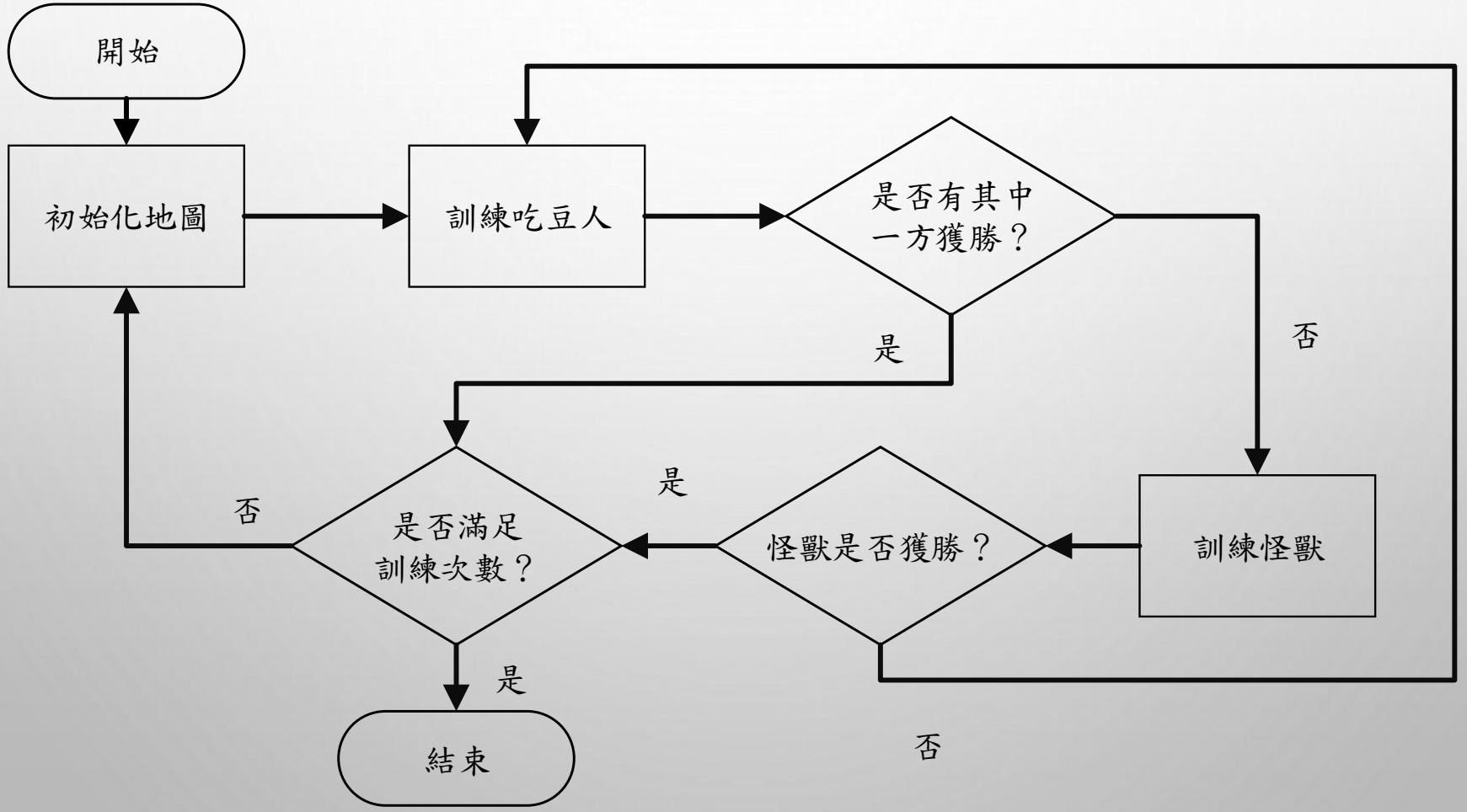
ALGORITHM

- **Q-LEARNING / SARSA**

  而我們將吃豆人與怪獸的 AGENT 分別對應至 Q-LEARNING 與 SARSA，其原因為，吃豆人的主要目標是將所有豆子吃光，因此會選擇能立刻吃到豆子的路線（亦即更新公式中的MAX ACTION）進行更新；而怪獸則是只要能在遊戲結束前將吃豆人吃掉，即可獲得勝利，因此不急於一時，所以採用 SARSA 作為其 AGENT。

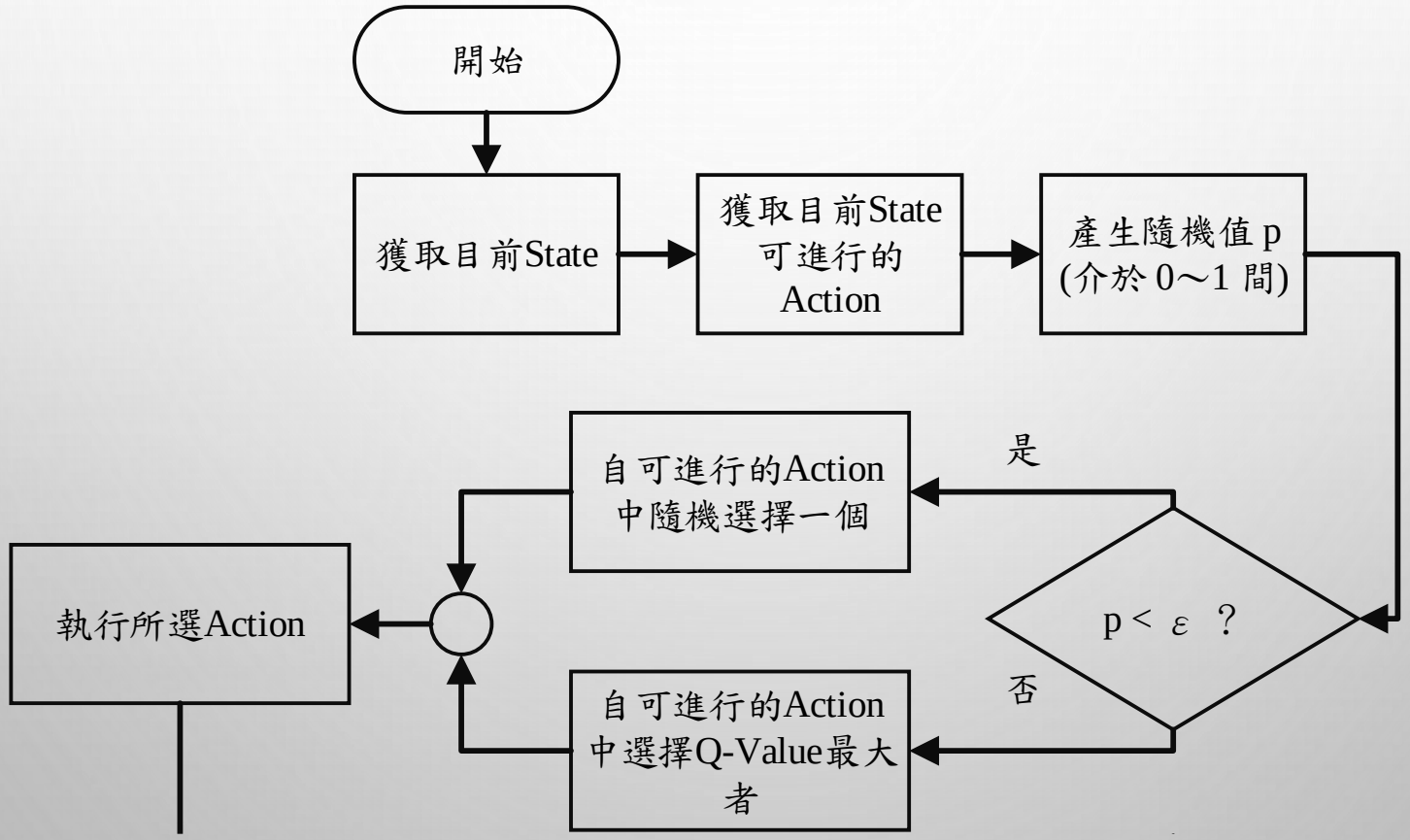
ALGORITHM FLOW CHART

- **整體流程圖**



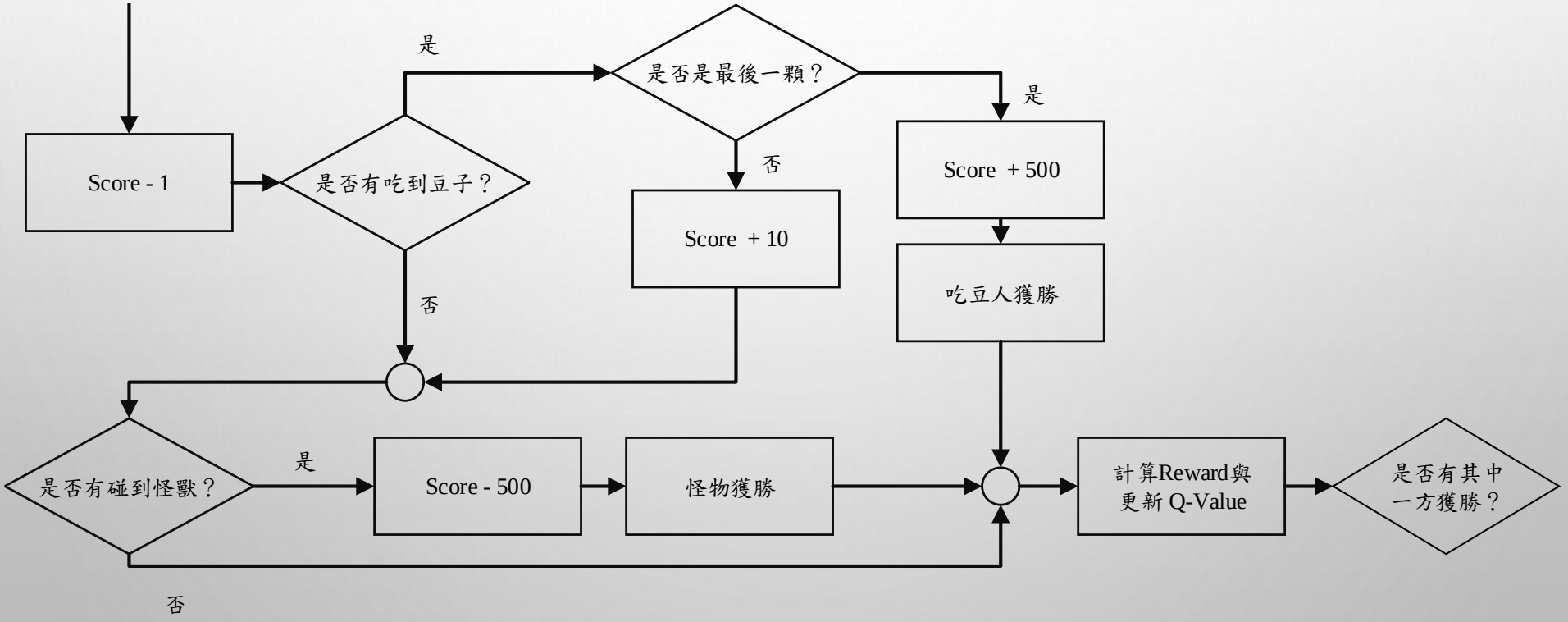
ALGORITHM FLOW CHART

- **吃豆人流程圖**



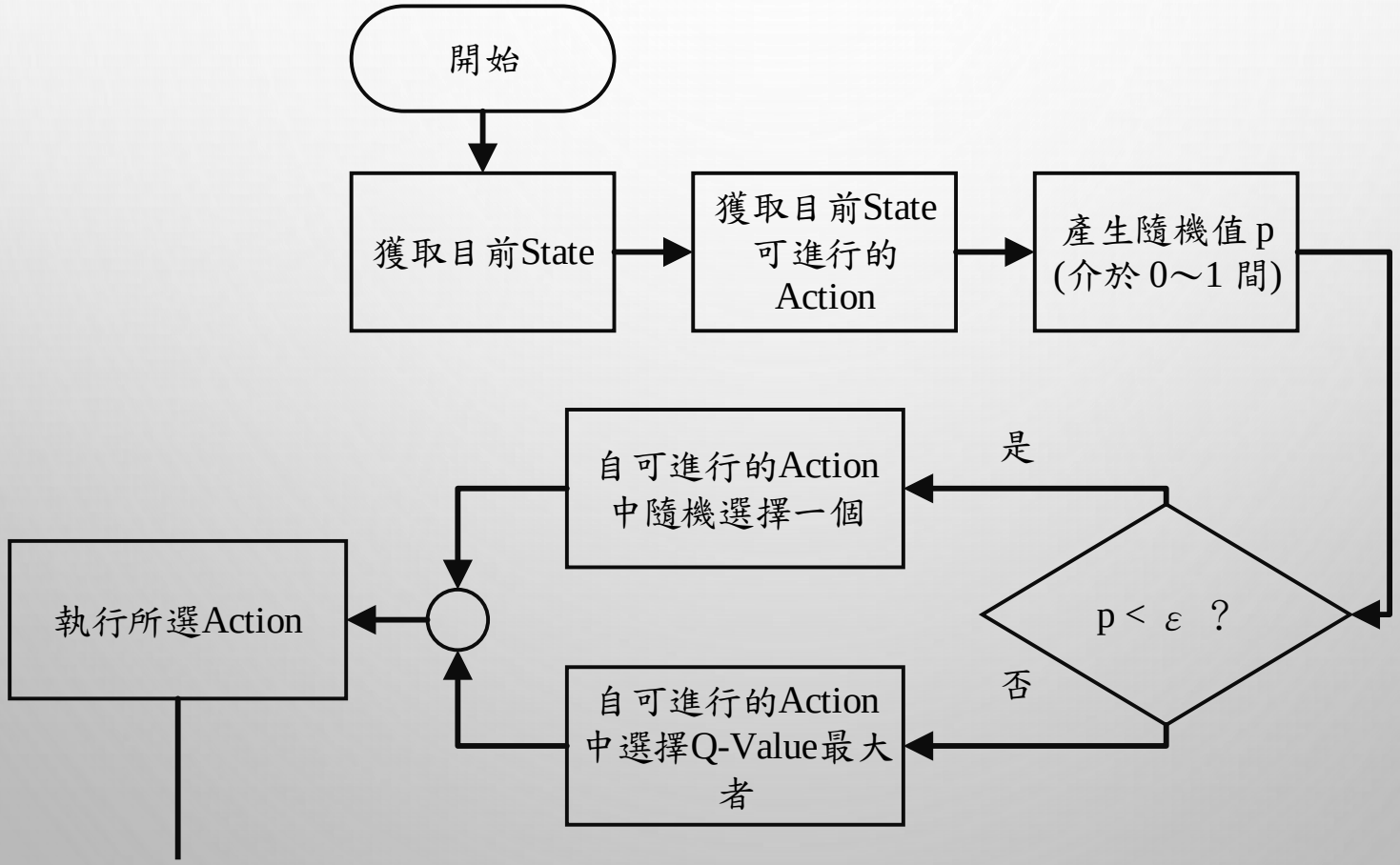
ALGORITHM FLOW CHART

- **吃豆人流程圖**



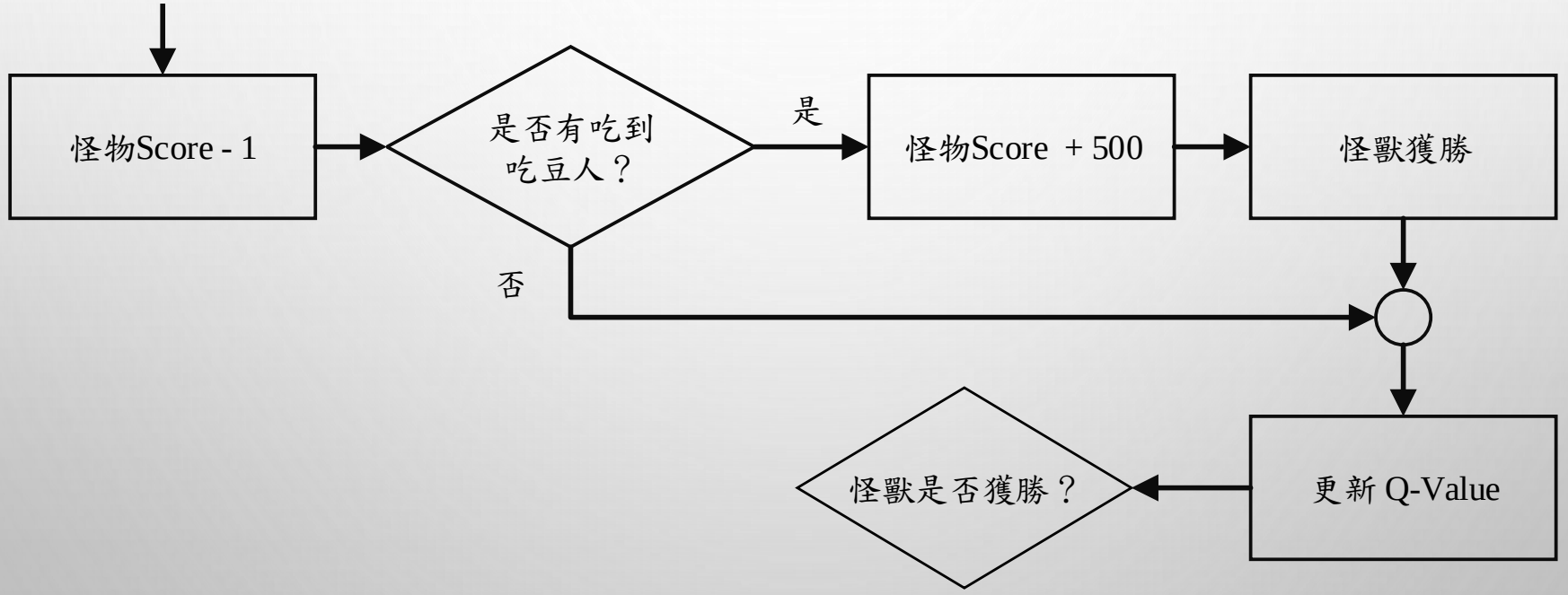
ALGORITHM FLOW CHART

- **怪物流程圖**



ALGORITHM FLOW CHART

- **怪物流程圖**



ALGORITHM

- **Q-LEARNING與APPROXIMATE Q-LEARNING差異**

1. 從Q-FUNTION原先以Q-TABLE的查表法(LOOK UP TABLE)的方式換成以權重(WEIGHT)*特徵(FEATURES)的方式來進行APPROXIMATION。

$$Q(s,a)=\sum_{k=1}^{n} f_k(s,a)\omega_k$$

2. 透過將Q-FUNCTION轉變成FUNCTION APPROXIMATION的方式且透過將更新的規則改成以更新權重的方式來進行計算，能夠有效地強化吃豆人在行為決策上的優化。

$$w_i \leftarrow w_i + \alpha \cdot \text{difference} \cdot f_i(s,a)$$
$$\text{difference} = \left(r + \gamma \max_{a'} Q(s',a')\right) - Q(s,a)$$

3. 而在本次期末專題報告中我們組設定了四個特徵（豆子會不會被吃掉、距離下一個豆子的距離遠近、是否即將撞上怪獸、是否與怪獸只差一步的距離）來進行訓練後，使得吃豆人在5X6MEDIUMGRID地圖的表現有了更顯著的提升。

ALGORITHM

- **Q-LEARNING與APPROXIMATE Q-LEARNING差異**

1. 從Q-FUNTION原先以Q-TABLE的查表法(LOOK UP TABLE)的方式換成以權重(WEIGHT)*特徵(FEATURES)的方式來進行APPROXIMATION。

$$Q(s,a)=\sum_{k=1}^{n} f_k(s,a)\omega_k$$

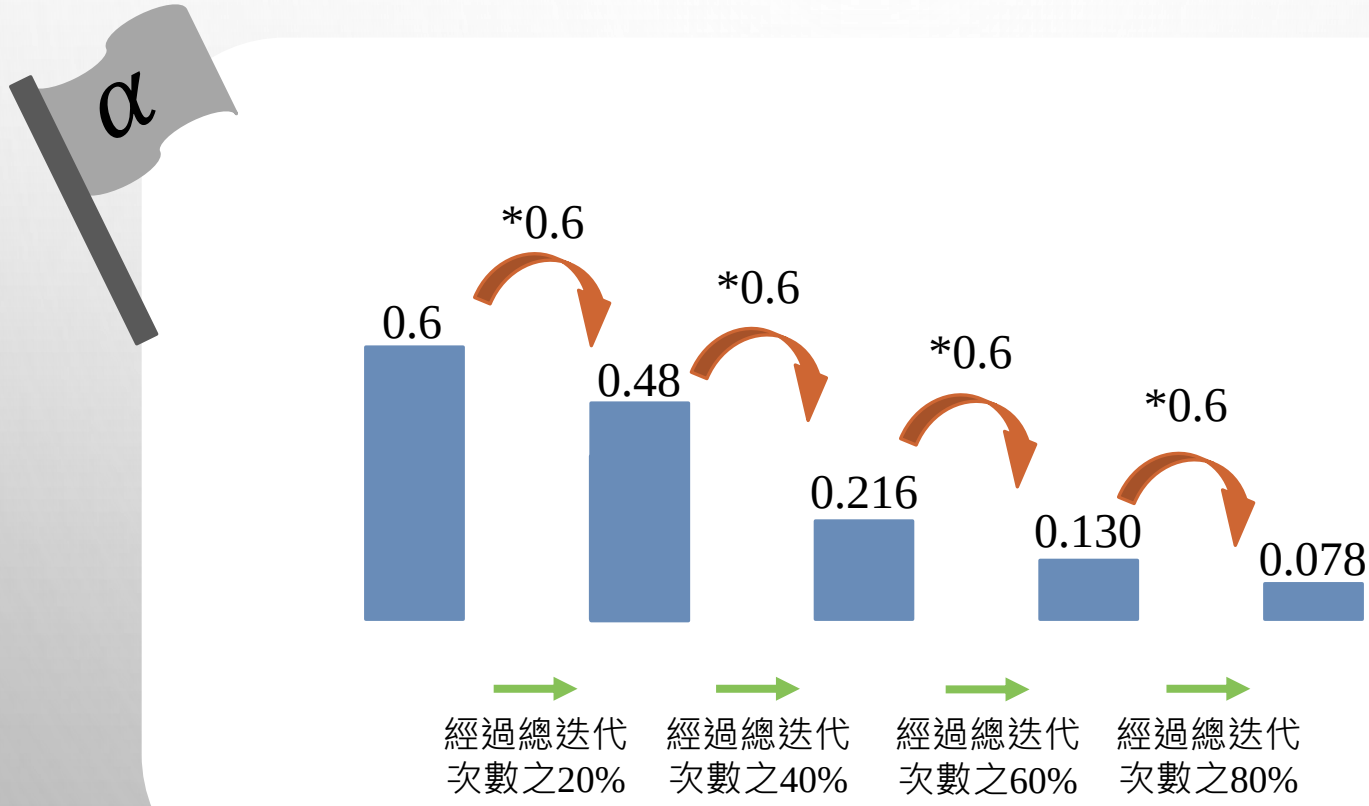
2. 透過將Q-FUNCTION轉變成FUNCTION APPROXIMATION的方式且透過將更新的規則改成以更新權重的方式來進行計算，能夠有效地強化吃豆人在行為決策上的優化。

$$w_i \leftarrow w_i + \alpha \cdot \text{difference} \cdot f_i(s,a)$$
$$\text{difference} = \left(r + \gamma Q(s',a')\right) - Q(s,a)$$

4. 而在怪獸的部分，我們設定了兩個特徵，分別為：怪獸是否吃到吃豆人、怪獸與吃豆人的距離，藉此提升怪獸的效能。

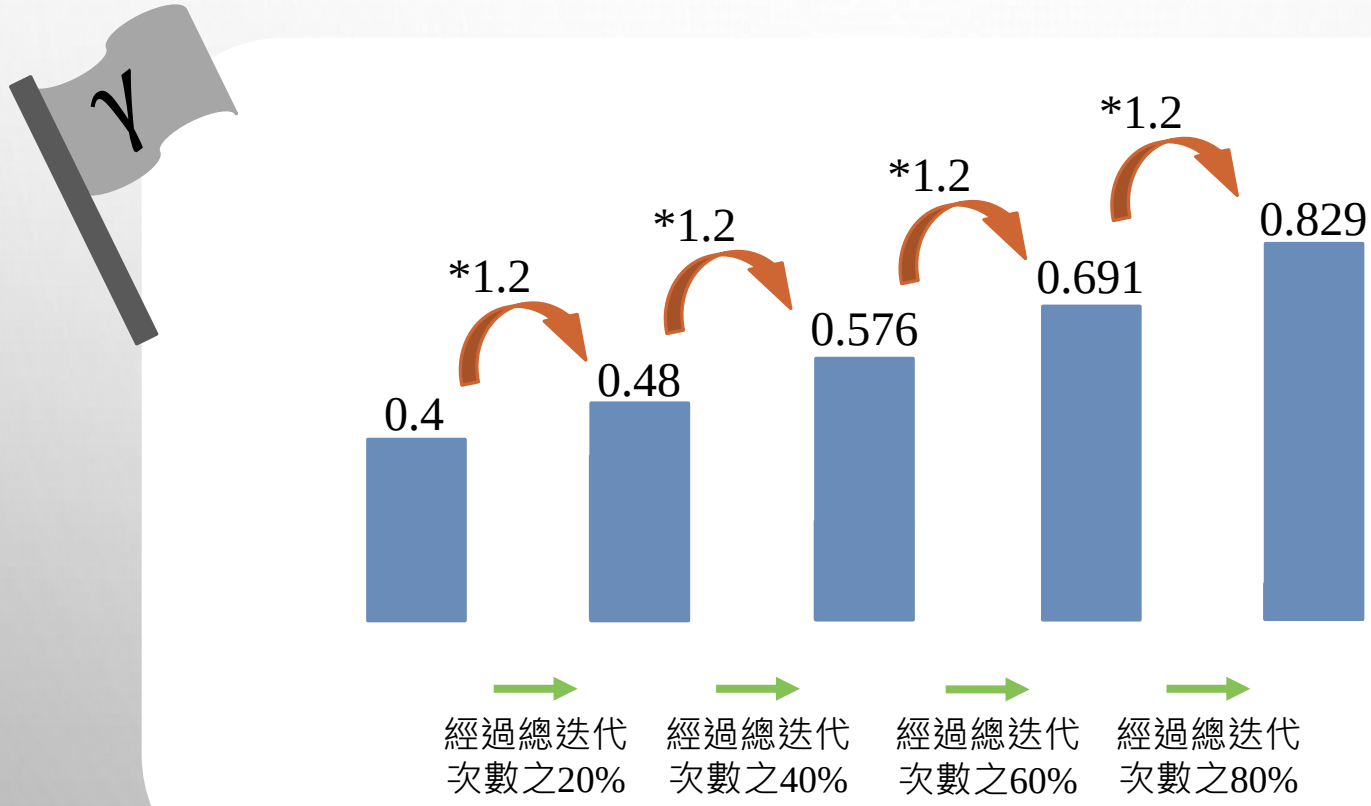
PARAMETER SETTING

- α =0.6、γ =0.4、ε =0.6 --- 動態調整



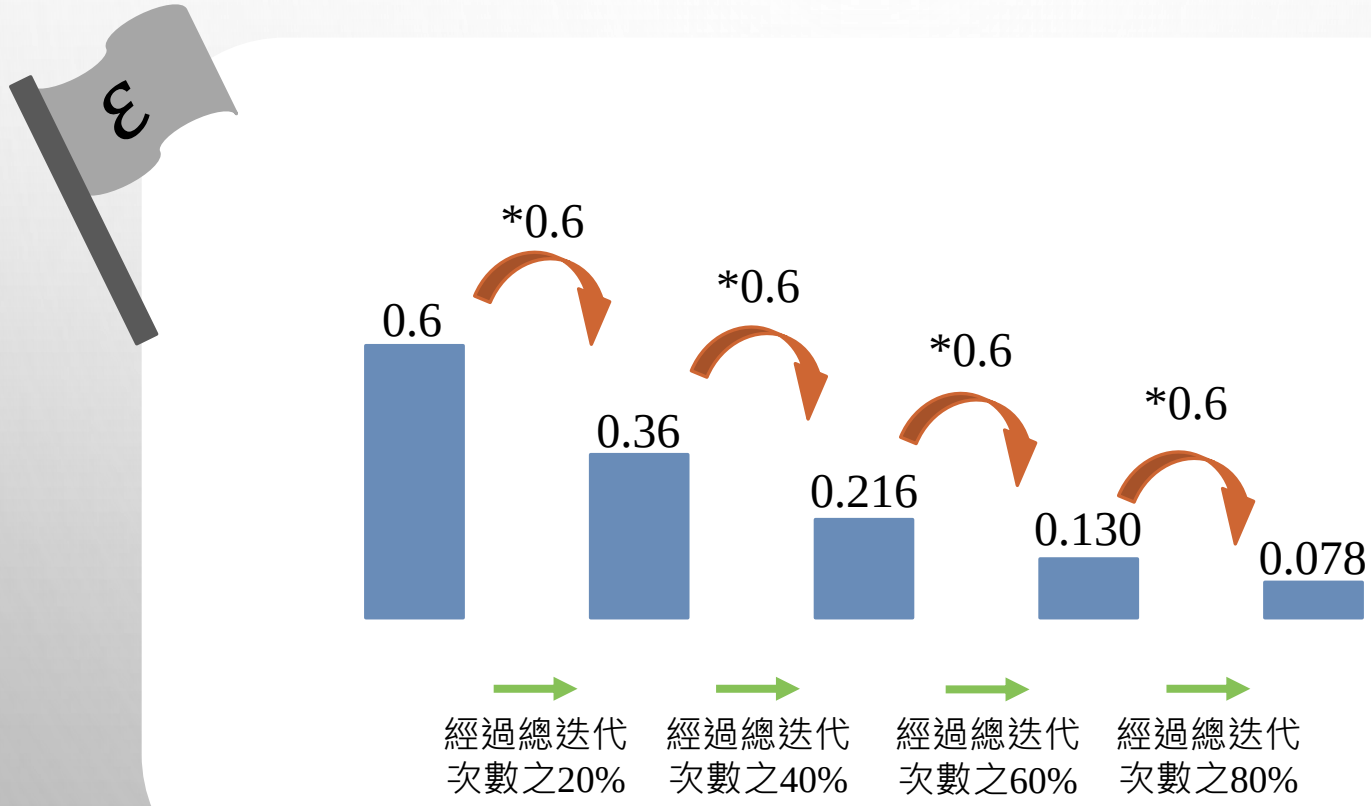
PARAMETER SETTING

- α =0.6、γ =0.4、ε =0.6 --- 動態調整



PARAMETER SETTING

- α =0.6、γ =0.4、ε =0.6 --- 動態調整



STATE&ACTION

- **吃豆人與怪物之STATE&ACTION**

吃豆人

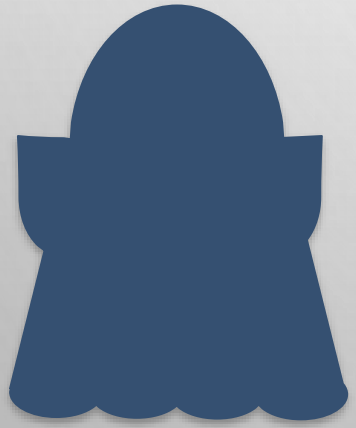
input	state	(x,y)				
	action	上	下	左	右	停止
output	state	(x,y+1)	(x,y-1)	(x-1,y)	(x+1,y)	(x,y)

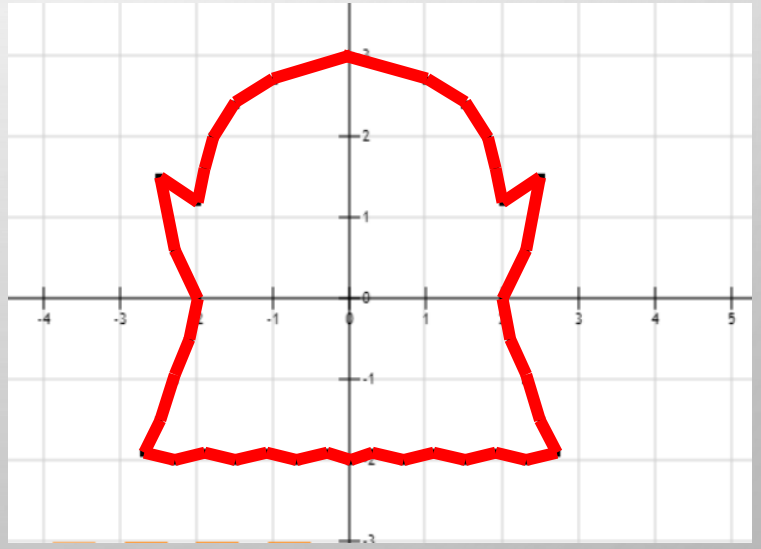
怪物

input	state	(x,y)				
	action	上	下	左	右	停止
output	state	(x,y+1)	(x,y-1)	(x-1,y)	(x+1,y)	(x,y)

INTERFACE AND ANIMATION DESIGN

- **繪圖技術**
  - 介面：使用PYTHON套件當中的TKINTER不斷更新遊戲畫面，其中畫面更新率為每0.05秒更新一次；選單畫面則由PYTHON套件當中的PYGAME進行更新。
  - 怪物：使用座標點建構出圖形，使用PYTHON內建POLYGON( )函式描繪；其中眼睛的部分再以額外一變數儲存目前X,Y位置。 POLYGON(已整理好且放入陣列之座標 , 顏色)





INTERFACE AND ANIMATION DESIGN

- **繪圖技術**
  - 吃豆人：使用PYTHON內建CIRCLE ( )函式描繪。
  - 豆子：使用PYTHON內建CIRCLE ( )函式描繪。
  - 初始畫面：利用PYGAME函式庫建立文字選單。
    - 使用LABEL( )建立文字---開始、結束、關於，各項選擇分別有各自的INDEX。
    - 利用PYGAME來不斷偵測”上鍵”、”下鍵”以及”ESC”之操作。

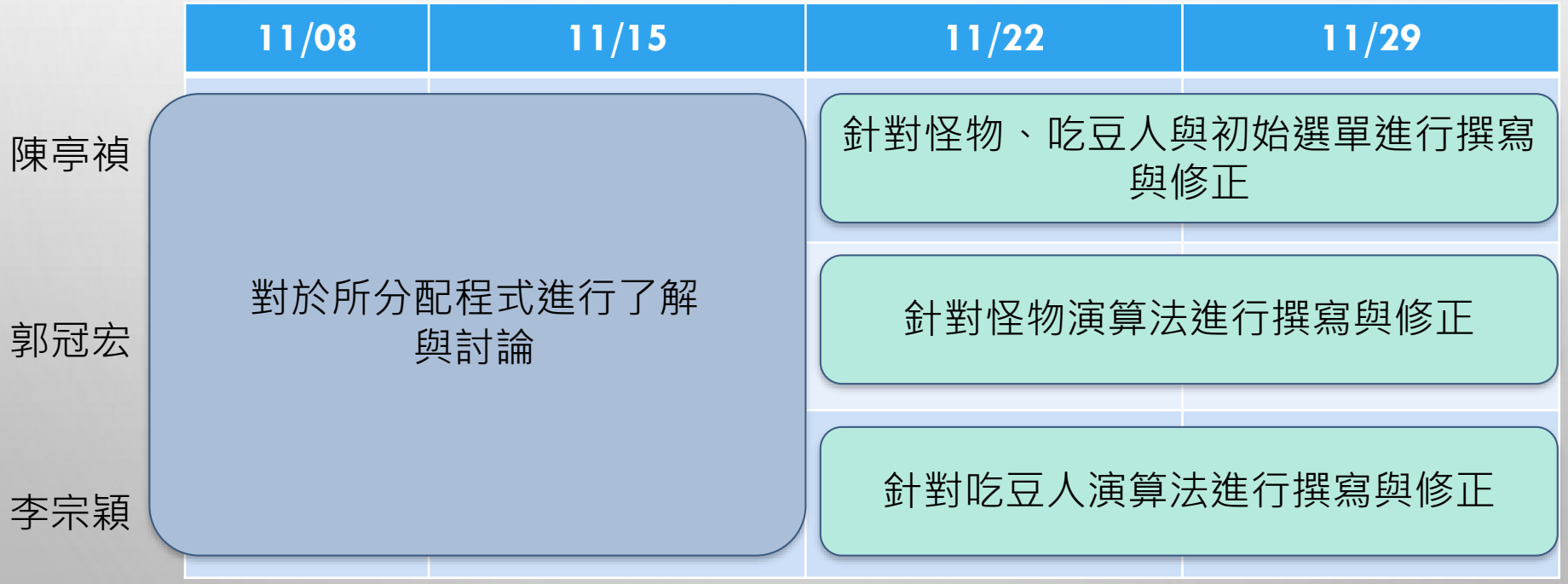
Index狀態	選單項目	對應結果
0	Start	呼叫演算法
1	About	呼叫顯示訊息頁面
2	Exit	停止所有遊戲

RESOURCE REQUIRED

- **介紹專題需要的軟硬體設備和開發工具**
  - 軟體：PYTHON3.6
  - 硬體：INTEL(R) CORE (TM) I7-2600CPU 3.4GHZ
- **人員的工作分配**
  - 演算法設計
    - 郭冠宏、李宗穎
  - 動畫圖畫設計
    - 陳亭禎
  - 成果簡報製作
    - 陳亭禎、郭冠宏、李宗穎

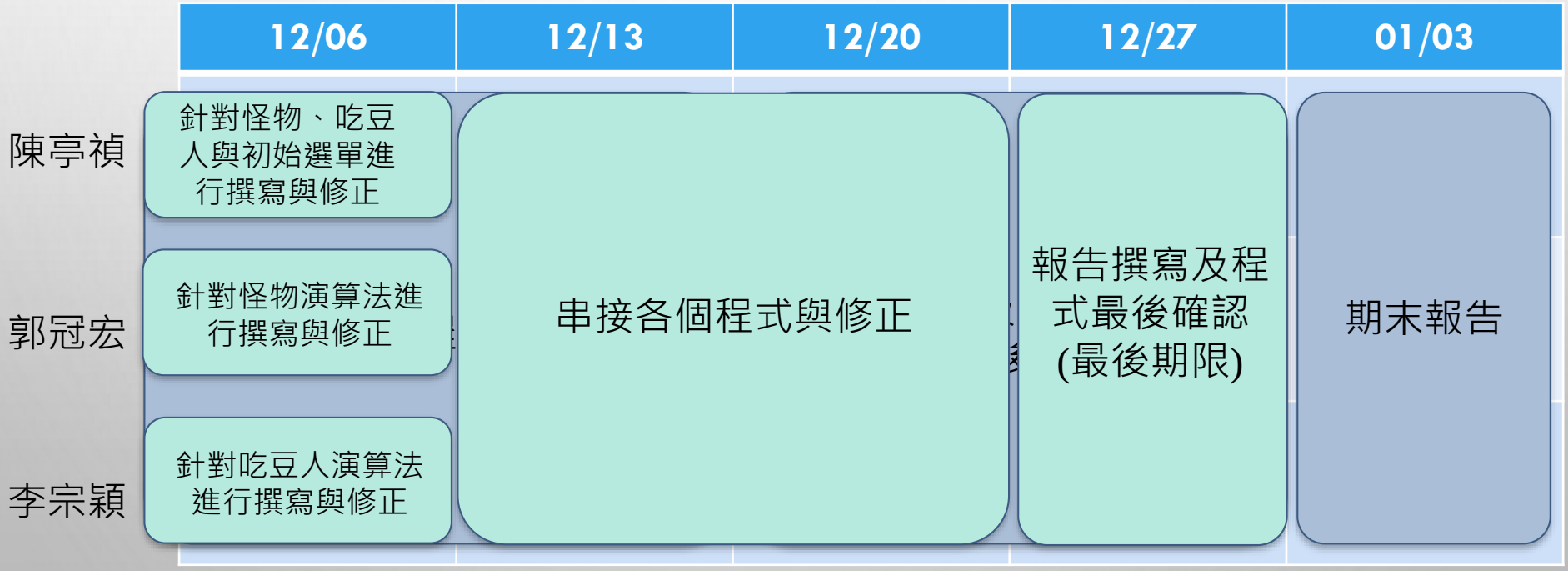
SCHEDULE

- **專題的規畫時程**

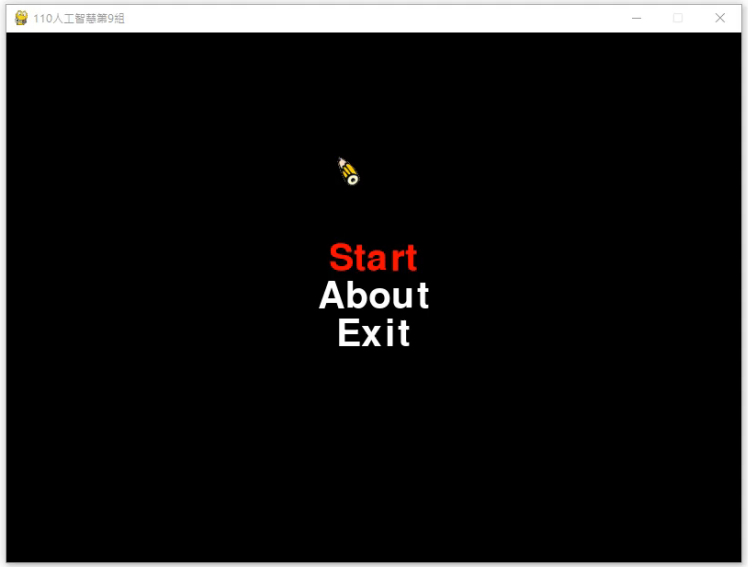


SCHEDULE

- **專題的規畫時程**



DEMO TIME

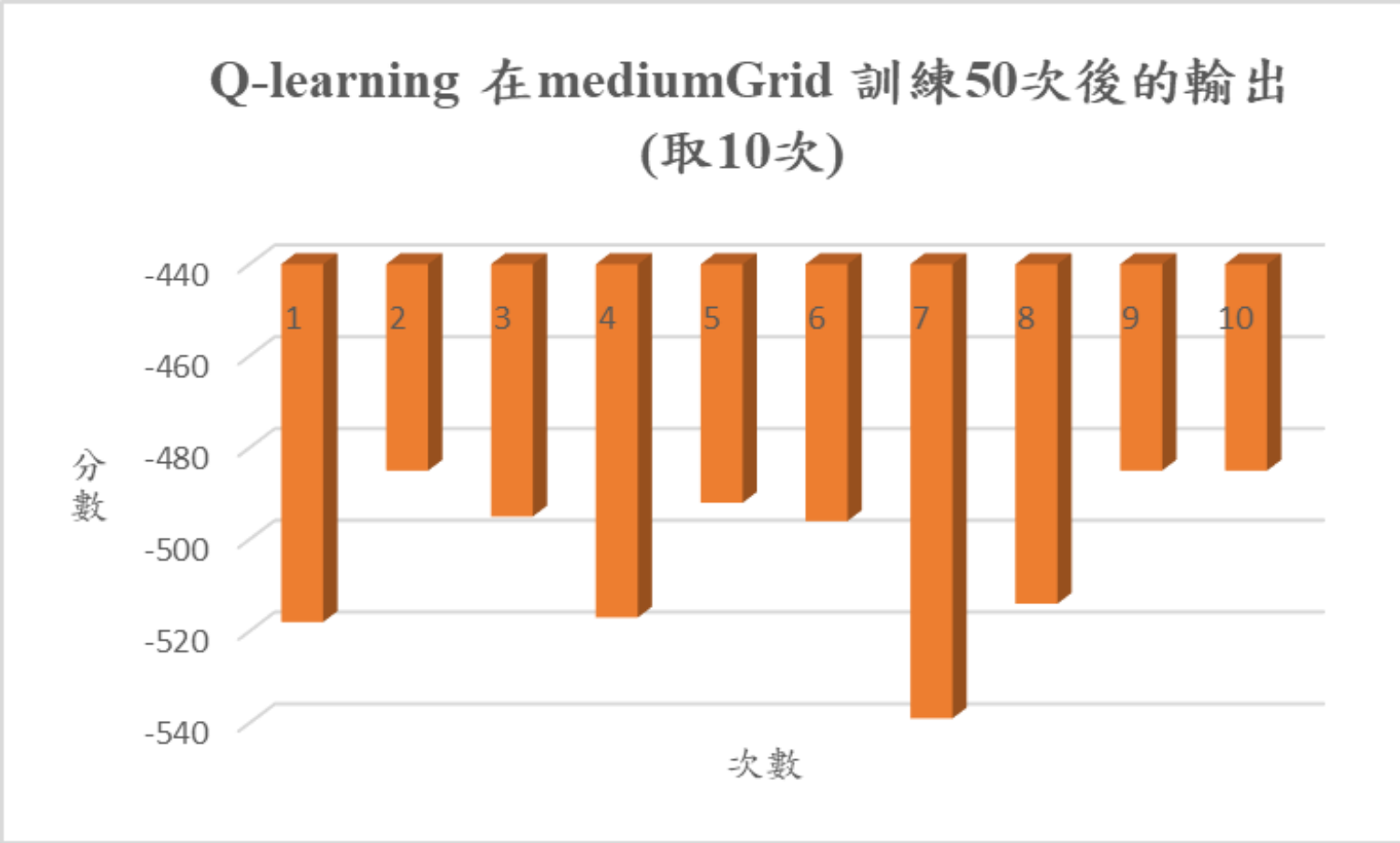


- **APPROXIMATE Q-LEARNING**

- **Q-LEARNING**

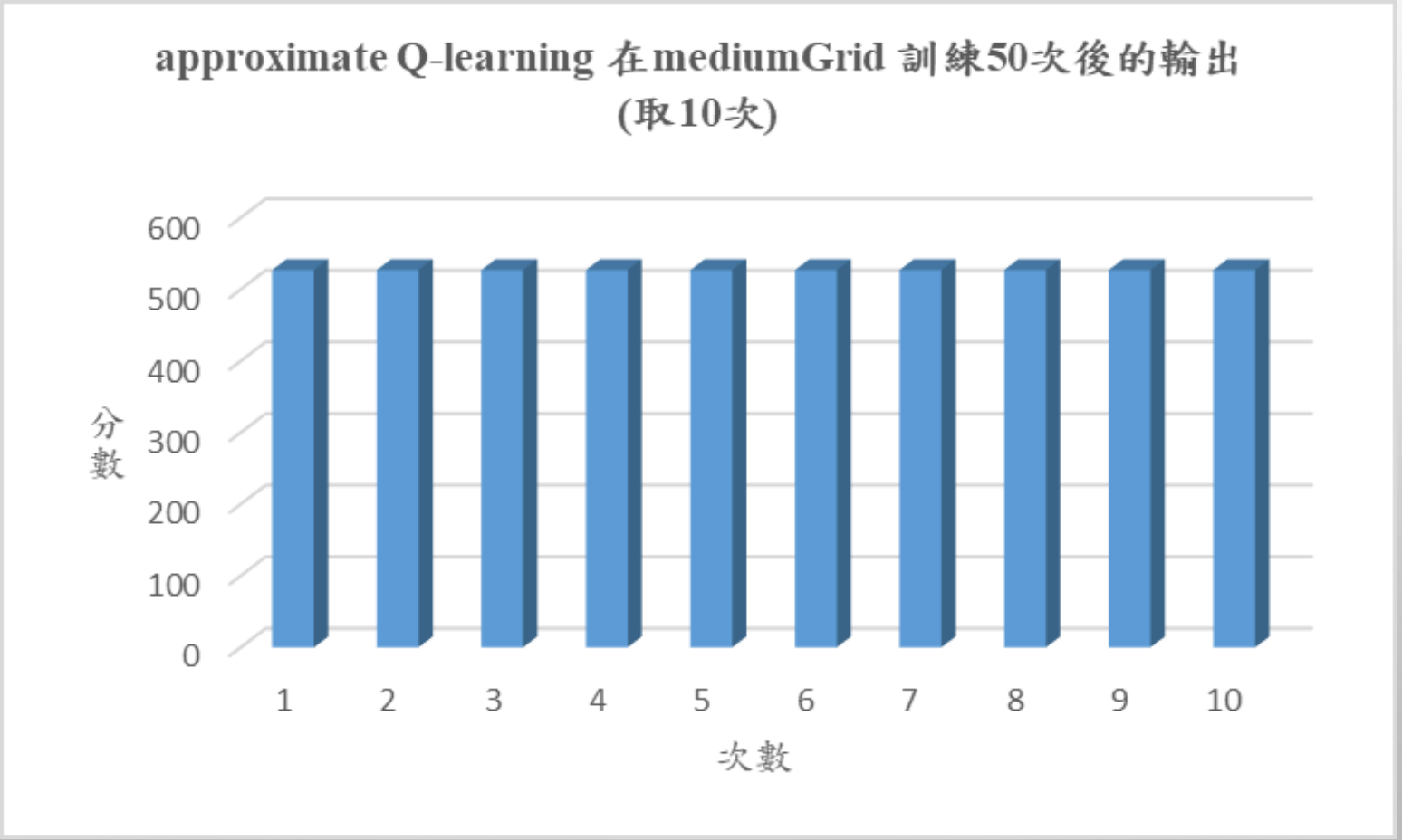
DEMO TIME

- **Q-LEARNING/APPROXIMATE Q-LEARNING訓練結果**



DEMO TIME

- **Q-LEARNING/APPROXIMATE Q-LEARNING訓練結果**



CONCLUSION

- 本專題核心目標在於吃豆人能自我學習避開怪物並將豆子全數吃光，而為了完成此目標，本專題利用強化式學習之APPROXIMATE Q-LEARNING演算法，使吃豆人能以當下所得環境進行目標得學習。從實作成果當中可發現，本專題除了成功使得吃豆人可避開怪物並吃掉豆子，還利用強化式學習當中的APPROXIMATE SARSA演算法讓怪物也可自我學習，以達成追逐吃豆人之目標。
- 我們冀希此專題後續能達成：
  1. 怪物數量增加；
  2. 吃豆人吃到大力丸時，能夠反追怪物，以獲得更高的REWARD；
  3. 地圖環境更為龐大。

  完成上述目標後，本專題發展可更為寬廣。

REFERENCES

- JO, T. (2021). MACHINE LEARNING FOUNDATIONS: SUPERVISED, UNSUPERVISED, AND ADVANCED LEARNING. SPRINGER NATURE.
- SUTTON, R. S., & BARTO, A. G. (2018). REINFORCEMENT LEARNING: AN INTRODUCTION. MIT PRESS.
- STUART, R., & NORVIG, P. (2010). ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 3RD EDITION. UPPER SADDLE RIVER, NEW JEYSEY.
- XU, Z. X., CAO, L., CHEN, X. L., LI, C. X., ZHANG, Y. L., & LAI, J. (2018). DEEP REINFORCEMENT LEARNING WITH SARSA AND Q-LEARNING: A HYBRID APPROACH. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS, 101(9), 2315-2322.
- ALFAKIH, T., HASSAN, M. M., GUMAEI, A., SAVAGLIO, C., & FORTINO, G. (2020). TASK OFFLOADING AND RESOURCE ALLOCATION FOR MOBILE EDGE COMPUTING BY DEEP REINFORCEMENT LEARNING BASED ON SARSA. IEEE ACCESS, 8, 54074-54084.
- MESUT YANG, & CARL QI. (2021). CS188|SUMMER 2021. RETRIEVED FROM HTTPS://INST.EECS.BERKELEY.EDU/~CS188/SU21/. (NOVEMBER 8,2021)

REINFORCEMENT LEARNING OF PACMAN FINAL PROJECT

陳亭禎、郭冠宏、李宗穎

授課教師：曾士桓 博士

TEAM09

電腦與通訊工程系

2022.01.10