# Aplicació de tècniques d'aprenentatge computacional per la creació d'agents jugadors de Sushi Go

**Consultor: Joan M. Nuñez Do Rio**
**Professor Responsable: Carles Ventura Royo**

Juny 2020

**Jose Montufo Rosal**
Grau d'Enginyeria Informàtica
Àrea d'Intel·ligència Artificial

# CONTINGUT

Motivacions i objectius

Introducció a l'aprenentatge per reforç

Regles de Sushi Go

Metodologia

Comparativa i Anàlisi

Conclusions

# MOTIVACIONS I OBJECTIUS

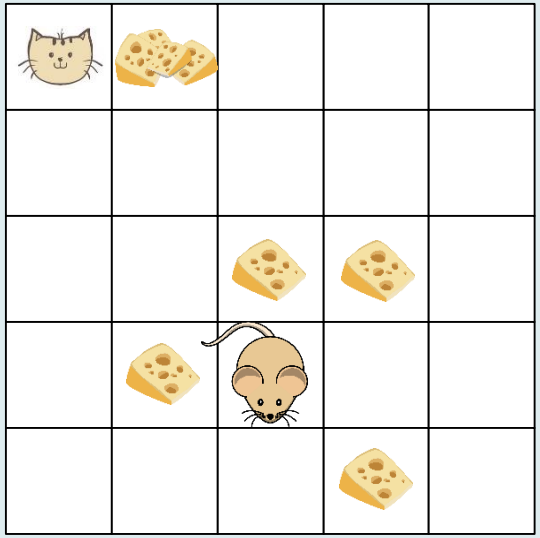
## Motivacions

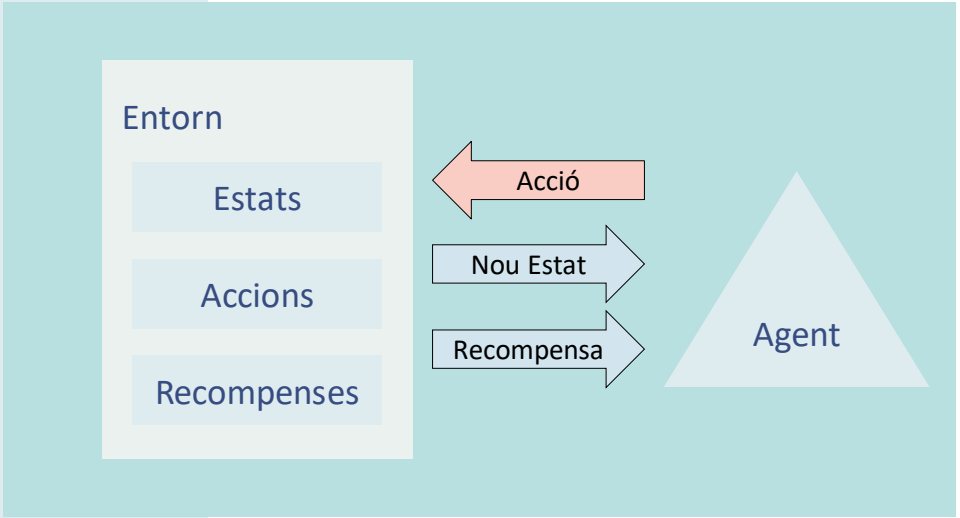
- Utilitzar un entorn de Sushi Go per generar agents creats mitjançant algoritmes d'aprenentatge per reforç
- Realitzar una comparativa dels agents creats mitjançant els diferents algoritmes
- Utilitzar els agents per determinar una estratègia òptima
- Crear un entorn de Sushi Go estandarditzat que permeti la implementació de noves versions dels agents

## Objectius específics

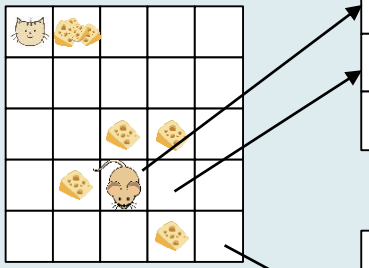
- Creació d'una implementació de Sushi Go. Adaptar-la per poder ser registrada com entorn OpenAI Gym
- Generació de diversos agents mitjançant els algoritmes Q-Learning, Deep Q-Learning i Monte Carlo
- Comparativa entre els agents generats, i anàlisi del comportament dels millors agents
- Creació d'una UI que permeti a un usuari jugar contra els agents generats

# INTRODUCCIÓ A L'APRENENTATGE PER REFORÇ





# ALGORITME Q-LEARNING



Q-TABLE

| → | ↑ | ← | ↓ |
|---|---|---|---|
| 5,8 | 5,2 | 4,9 | 5,9 |
| 3,5 | 2,9 | 3,8 | 3,1 |
| 4,8 | 4,2 | 3,5 | 2,1 |
| 0 | 1,2 | 1,5 | 1,9 |
| ... | | | |
| 4,9 | 5,2 | 5,7 | 5,2 |
| 0 | 5,9 | 6,2 | 0 |

## Mètode d'actualització per diferència temporal

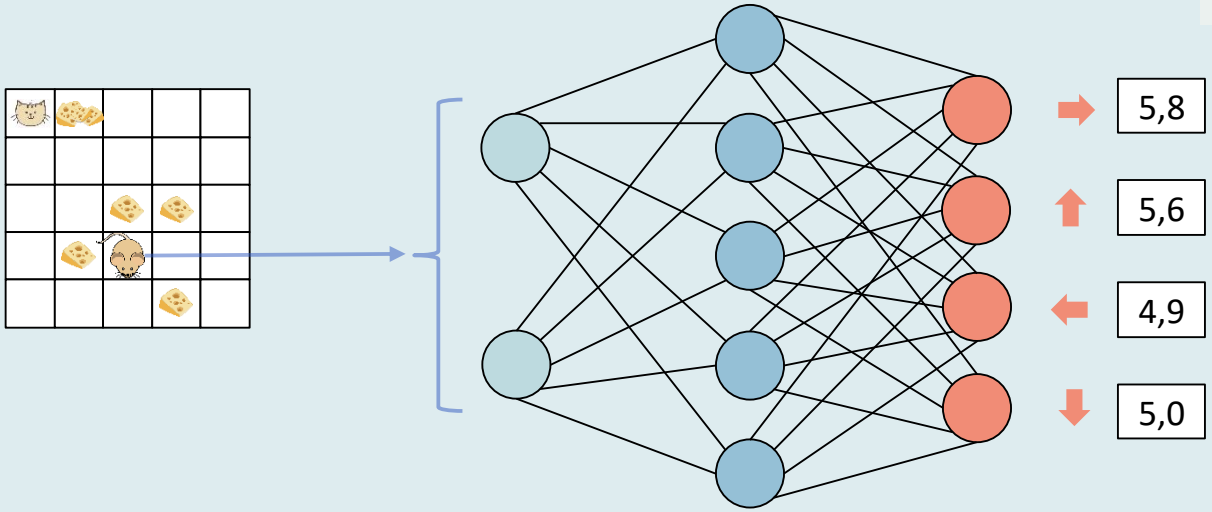
$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha * (r_t + \gamma * max_a\, Q(s_{t+1}, a) - Q(s_t, a_t))$$

- Valor de descompte
- Antic valor de l'acció en l'estat
- Valor del nou estat
- Diferència temporal entre el valor obtingut i l'esperat
- Ràtio d'aprenentatge

# ALGORITME DEEP Q-LEARNING

$$Q^{new}(s_t, a_t) \leftarrow r_t + \gamma * max_a Q(s_{t+1}, a)$$

XARXA NEURONAL



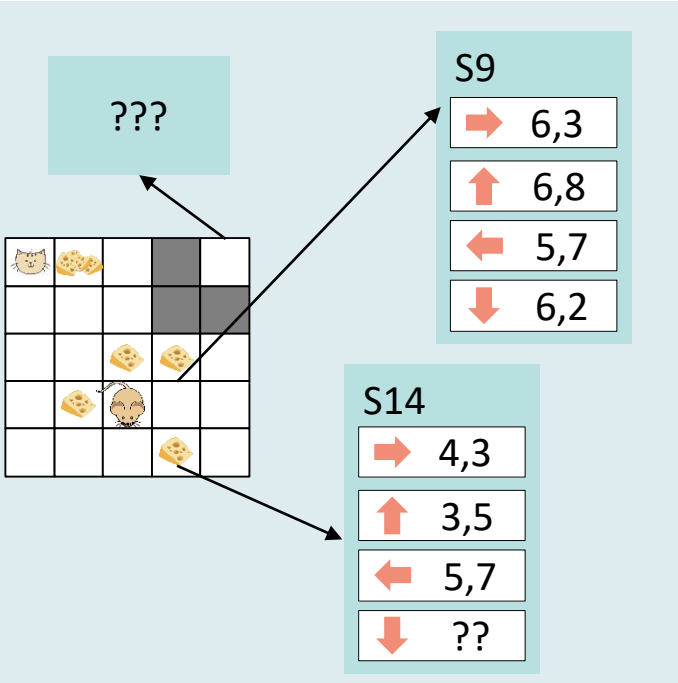

**MEMÒRIA D'EXPERIÈNCIES**

La xarxa no s'actualitza de forma ordenada.
Es guarden les experiències, i s'actualitza per paquets desordenats.

**VERSIÓ DOUBLE DEEP Q-LEARNING**

Utilitza dues xarxes virtuals: una principal que manté el valor de les accions als estats, i una auxiliar per obtenir el valor de les accions als estats següents.

# ALGORITME MONTE CARLO




**Expansió:** Es prioritzen els nodes no explorats

**Selecció:** Es tria l'acció a realitzar balancejant exploració i explotació

**Retropropagació:** S'actualitzen els valors de les accions amb el nou valor obtingut

| | | |
|---|---|---|
| $S_1$ | $a_1$ → | $r_1$ |
| $S_2$ | $a_2$ ↓ | $r_2$ |
| $S_3$ | $a_3$ → | $r_3$ |
| $S_4$ | $a_4$ ↑ | $r_4$ |
| $S_5$ | $a_5$ ↓ | $r_5$ |

$R_3$

$$Q^{new}(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{1}{n}[(R_t - Q(s_t, a_t)]$$

# OPEN AI GYM

**OpenAI Gym** és un toolkit per al desenvolupament i la comparació d'algoritmes d'aprenentatge per reforç

## Entorns predefinits



## Entorns personalitzats

**Env**

observation_space

action_space

reset() : observation

step(action) : new_observation, reward, done, info

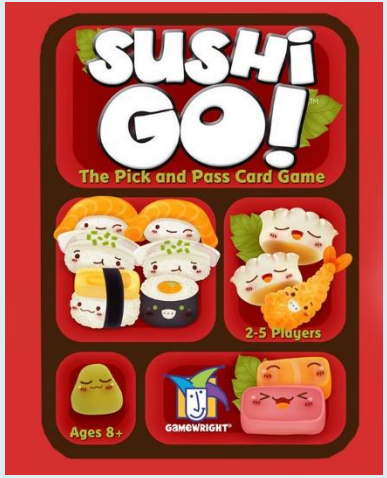
render()

# SUSHI GO





- 2 a 5 jugadors.
- A cada torn, cada jugador tria una carta de la mà, la mostren tots els jugadors a la vegada, i s'entreguen les cartes restants de la mà al jugador de la seva esquerra.
- Es juguen 3 rondes. Al final de cada ronda, es calculen els punts obtinguts per les cartes utilitzades, es retiren aquestes cartes, i es reparteix una nova mà.
- Objectiu: Ser el jugador amb més punts acumulats al llarg de les tres rondes.







# IMPLEMENTACIÓ / ESPAI D'ACCIONS

## Implementació

- A partir d'una de les implementacions preexistents
- Refactorització del codi
- Adaptació com a entorn d'OpenAI Gym
- Creació d'una UI per línia de comandes

## Espai d'accions

| Accions simples | |
|---|---|
| 0 | Palets |
| 1 | Nigiri |
| 2 | Wasabi |
| 3 | Maki |
| 4 | Sashimi |
| 5 | Tempura |
| 6 | Gyoza |
| 7 | Puding |

| Accions de palets | | |
|---|---|---|
| 8 | Nigiri | Nigiri |
| 9 | Nigiri | Wasabi |
| 10 | Nigiri | Maki |
| … | … | … |
| 33 | Gyoza | Gyoza |
| 34 | Gyoza | Puding |
| 35 | Puding | Puding |
| 36 | Wasabi | Nigiri |

Eliminades accions equivalents i accions subòptimes

Transformació d'identificador a carta o parell de cartes, i a la inversa

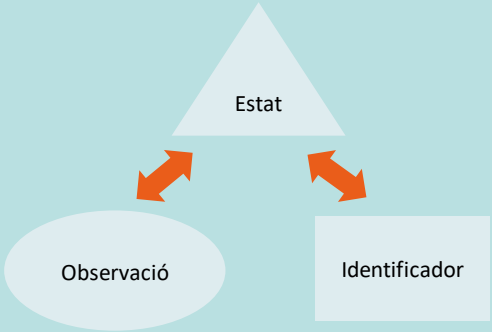
Construcció llista d'accions possibles a partir de conjunt de cartes

# ESPAI D'ESTATS / AGENTS

## Espais d'estats

Obté la informació de la implementació des de la perspectiva d'un jugador

Sistema modular:
Un espai pot contenir altres estats




**Espai de partida**

Torn
Ronda

**Espai de mà**

Maki Màxim
Nigiri Màxim

**Espai de taula**

Maki
Tempura
Gyoza
Sashimi
Wasabi
Pudding
Palets

**Espai de jugador**

Espai partida
Espai de mà
Espai de taula

**Espai complet**

Espai partida
Espai de mà
Espai de taula

Espais de taula de la resta de jugadors

## Agents

Tenen assignat un espai d'estats per obtenir l'estat de l'entorn

Tenen una funció que implementa la policy: per cada estat, determina l'acció òptima

**GENERATS MANUALMENT**

RandomAgent: Tria l'acció a l'atzar

Agents que prioritzen unes accions sobre la resta.

**GENERATS AUTOMÀTICAMENT**

Agents que determinen la policy mitjançant tècniques d'aprenentatge per reforç

# ESTRATÈGIA DE CREACIÓ DELS AGENTS

## Decisions prèvies

Partides de 2 jugadors

Es maximitza la puntuació obtinguda, no les victòries

Són necessaris agents rivals per realitzar l'entrenament dels agents automàtics

A cada fase, es crea un agent per cadascun dels algoritmes d'aprenentatge per reforç

Més quantitat d'agents rivals amb diferents policies implica menor biaix i major variància

## FASE 1

El rival al qual s'enfronten els agents és sempre el RandomAgent

## FASE 2

A cada experiment, el rival al qual s'enfronten els agents es tria aleatòriament entre tots els agents generats manualment

## FASE 3

A cada experiment, el rival al qual s'enfronten els agents es tria aleatòriament entre els agents de fase 1 i de fase 2

## FASE 4

A cada experiment, el rival al qual s'enfronten els agents es tria aleatòriament entre els agents de fase 1, 2 i 3 i tots els agents manuals

# METODOLOGIA D'IMPLEMENTACIÓ. Q-LEARNING

## Definició d'espais d'estats

**Espai senzill de jugador**

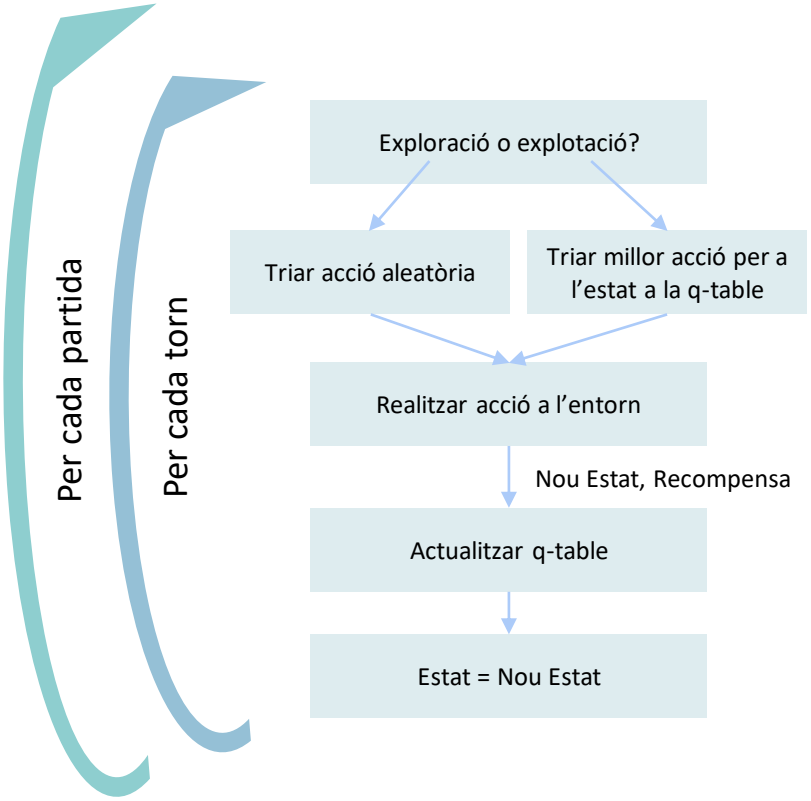
Espai partida
Espai simple de taula

**Espai de partida**

Torn (1-10)
Ronda (1-3)

**Espai senzill de taula**

Maki (0-1)
Tempura (0-1)
Gyoza (0-2)
Sashimi (0-2)
Wasabi (0-1)
Pudding (0-1)
Palets (0-1)




## Implementació de l'agent

Seleccionar l'acció a triar de la q-table:

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| | 5,8 | 5,2 | 4,9 | 5,9 |
| Estat actual | 3,5 | 2,9 | 3,8 | 3,1 |
| | 4,8 | 4,2 | 3,5 | 2,1 |
| | 0 | 1,2 | 1,5 | 1,9 |

# METODOLOGIA D'IMPLEMENTACIÓ. MONTE CARLO

## Definició d'espais d'estats

**Espai simple de jugador v2**

Espai partida
Espai simple de taula
Espai de mà

... ...

**Espai de mà**
Maki màxim (0-3)
Nigiri màxim (0-3)

**Espai mig de jugador**

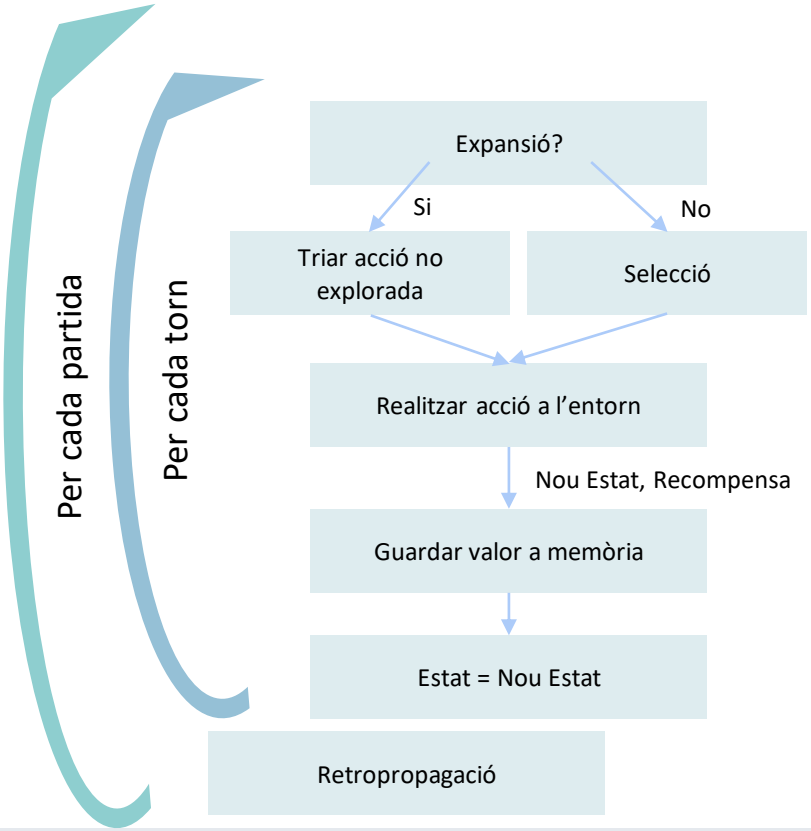
Espai partida
Espai simple de taula

**Espai de partida**

Torn (1-10)
Ronda (1-3)

**Espai mig de taula**

Maki (0-9)
Tempura (0-1)
Gyoza (0-5)
Sashimi (0-2)
Wasabi (0-1)
Pudding (0-4)
Palets (0-1)




## Implementació de l'agent

Trobar el node pertanyent a l'estat

Triar acció amb valor màxim

Si estat no explorat → A l'atzar

# METODOLOGIA D'IMPLEMENTACIÓ. DEEP Q-LEARNING

## Definició d'espais d'estats

### Espai complet de jugador

Espai partida
Espai complet de taula
Espai de mà

... ...

### Espai complet de taula

Maki (0-29)
Tempura (0-1)
Gyoza (0-5)
Sashimi (0-2)
Wasabi (0-5)
Pudding (0-9)
Palets (0-5)

### Espai complet

Espai complet de jugador
Espai complet taula (jugador a la dreta)
Espai complet taula (següent jugador)
....

## Memòria d'experiències

Mida total: 1024 experiències
Mida dels paquets: 256 experiències

## Estandardització de les observacions

Atributs no es troben al mateix rang

Atributs no al voltant de 0

Atributs no tenen una distribució normal: Transformació Min-Max

## Estructura de la xarxa

Tres capes:
- Entrada, longitud de l'espai d'estats
- Oculta, 256 nodes
- Sortida, longitud de l'espai d'accions

## Xarxa de valors objectiu

Actualització cada 100 episodis

# COMPARATIVA DELS AGENTS GENERATS

## GENERACIÓ DE LA COMPARATIVA

QL : Q-Learning
MC : Monte Carlo
DQL : Deep Q-Learning
DDQL : Double Deep Q-Learning

1 : Fase 1
2 : Fase 2
3 : Fase 3
4 : Fase 4

2000 partides entre cada parell d'agents

| Classificació per percentatge de victòries | |
|---|---|
| Agent | Percentatge |
| DQL1 | 61,65 |
| DQL4 | 58,05 |
| DQL3 | 57,33 |
| DDQL1 | 57,09 |
| DDQL3 | 56,71 |
| DQL2 | 55,88 |
| DDQL2 | 55,19 |
| DDQL4 | 51,83 |
| MC3 | 48,33 |
| MC2 | 48,21 |
| QL3 | 46,06 |
| QL2 | 44,46 |
| MC4 | 43,15 |
| MC1 | 42,68 |
| QL1 | 40,71 |
| QL4 | 38,11 |

| Classificació per puntuació mitjana | |
|---|---|
| Agent | Puntuació mitjana |
| DQL4 | 55,43 |
| DQL3 | 55,11 |
| DQL2 | 54,96 |
| DDQL1 | 54,90 |
| DQL1 | 54,87 |
| DDQL2 | 54,76 |
| DDQL3 | 54,74 |
| DDQL4 | 54,47 |
| MC3 | 52,89 |
| QL3 | 52,18 |
| MC2 | 52,02 |
| QL2 | 51,89 |
| MC4 | 51,66 |
| MC1 | 51,59 |
| QL1 | 51,24 |
| QL4 | 50,03 |

## CONCLUSIONS DE LA COMPARATIVA

Deep Q-Learning
↓
Double Deep Q-Learning
↓
Monte Carlo
↓
Q-Learning

| Deep Q-Learning | MC i QL |
|---|---|
| Fase 4 | Fase 3 |
| ↓ | ↓ |
| Fase 3 | Fase 2 |
| ↓ | ↓ |
| Fase 2 | Fase 1 |
| ↓ | ↓ |
| Fase 1 | Fase 4 |

# ANÀLISI DE L'ESTRATÈGIA DELS AGENTS

## Estratègia comuna

**Wasabi:** Primer torn sempre que sigui possible

**Palets:** Només tenen valor a l'inici de la ronda

**Nigiri:** Segon torn per aprofitar el Wasabi. Constant al llarg de la ronda.

**Tempura:** Molts punts per carta per poc risc a la primera meitat de la ronda.

**Gyoza:** Puntuació immediata sense risc i acumulable, per a la segona meitat.

**Maki :** Vàlids només al final, quan l'opció són cartes que no donaran cap punt.

**Puding :** Poc utilitzats, només al final pel mateix motiu que els maki.

## DQL4 : Millor puntuació

| Torn | Palets | Nigiri | Wasabi | Maki | Sashimi | Tempura | Gyoza | Puding |
|---|---|---|---|---|---|---|---|---|
| 1 | 52.64 | 19.23 | 98.37 | 0.00 | 12.04 | 4.26 | 11.28 | 0.86 |
| 2 | 13.72 | 45.52 | 30.83 | 0.04 | 26.82 | 19.40 | 27.41 | 1.75 |
| 3 | 19.95 | 36.97 | 41.52 | 0.30 | 23.47 | 42.02 | 33.98 | 3.30 |
| 4 | 16.37 | 38.75 | 36.83 | 0.99 | 23.16 | 54.43 | 43.74 | 10.19 |
| 5 | 13.05 | 51.30 | 23.73 | 3.71 | 23.81 | 56.84 | 55.44 | 22.12 |
| 6 | 19.13 | 60.76 | 29.75 | 11.45 | 26.67 | 56.09 | 68.99 | 39.95 |
| 7 | 23.04 | 66.56 | 28.70 | 30.40 | 36.01 | 54.76 | 77.58 | 49.41 |
| 8 | 23.83 | 65.94 | 32.15 | 58.81 | 42.46 | 45.92 | 82.73 | 53.56 |
| 9 | 33.86 | 66.78 | 33.18 | 79.49 | 60.19 | 43.56 | 87.35 | 60.58 |
| 10 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

## DQL1 : Més victòries

| Torn | Palets | Nigiri | Wasabi | Maki | Sashimi | Tempura | Gyoza | Puding |
|---|---|---|---|---|---|---|---|---|
| 1 | 36.01 | 15.96 | 97.88 | 0.03 | 26.53 | 11.80 | 0.89 | 0.00 |
| 2 | 29.23 | 45.06 | 34.68 | 2.01 | 37.54 | 24.35 | 4.82 | 1.61 |
| 3 | 23.34 | 42.82 | 39.17 | 1.50 | 34.55 | 44.19 | 14.20 | 4.39 |
| 4 | 23.47 | 39.64 | 31.99 | 3.46 | 31.72 | 59.13 | 29.11 | 8.33 |
| 5 | 23.83 | 45.05 | 18.06 | 8.77 | 23.09 | 66.02 | 48.08 | 15.39 |
| 6 | 17.93 | 50.83 | 17.39 | 22.73 | 17.89 | 66.77 | 63.93 | 27.22 |
| 7 | 14.33 | 53.52 | 14.96 | 44.30 | 14.71 | 58.81 | 72.61 | 35.60 |
| 8 | 12.25 | 61.15 | 20.54 | 66.38 | 18.26 | 49.49 | 76.86 | 42.90 |
| 9 | 23.19 | 76.62 | 41.00 | 80.89 | 40.22 | 51.05 | 83.49 | 61.21 |
| 10 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

## Diferència clau

**Sashimi :** DQL4 no l'utilitza. No perd punts per Sashimi sense tripleta, però el rival acumula tots els Sashimi. Per tant, puntuació més gran, però també el rival.

DQL1 tria Sashimi a l'inici, i perd punts per Sashimis sense tripleta, però el rival tampoc completa tripletes de Sashimi. Menys punts per l'agent, però encara menys pels rivals.

# CONCLUSIONS

## Lliçons apreses

- La gran quantitat de tècniques i mètodes que proporciona l'àrea de l'aprenentatge per reforç és completament inabastable en un espai limitat de temps.
- Abans d'iniciar processos d'entrenament d'agents que requereixin llargs temps d'execució, és primordial disposar d'una implementació de l'entorn totalment estable i fiable.
- La creació d'estructures flexibles per la gestió dels espais d'estats i d'accions, i per la creació de nous agents, han facilitat la tasca de creació i emmagatzematge dels agents generats.

## Objectius assolits

- Gran part dels objectius previstos inicialment s'han assolit.
- Objectius no assolits:
  - GUI
  - Algoritme Policy Gradients
  - Comparació amb agents preexistents

## Seguiment de la planificació

- La planificació ha sofert diverses modificacions al llarg de la durada del projecte, però cap de gran magnitud.
- Gran part dels reajustaments de planificació han estat provocats per una previsió massa optimista.
- La metodologia planificada a l'inici del projecte ha ajudat a dur a terme el projecte de manera satisfactòria.

# LÍNIES DE TREBALL FUTUR

Partides de més de 2 jugadors

Aplicar a diferents versions del joc

Millores als algoritmes d'aprenentatge utilitzats

Variacions dels espais d'estat utilitzats. Estats que mantinguin memòria de les mans

Creació d'una GUI

Utilitzar altres algoritmes al mateix entorn

Reutilitzar els algoritmes i les estructures per a nous entorns més enllà de Sushi Go

Universitat Oberta de Catalunya

# Moltes gràcies per la seva atenció