RENFORCEMENT, RÉCOMPENSE, VALEUR ET ACTION

Séance 10

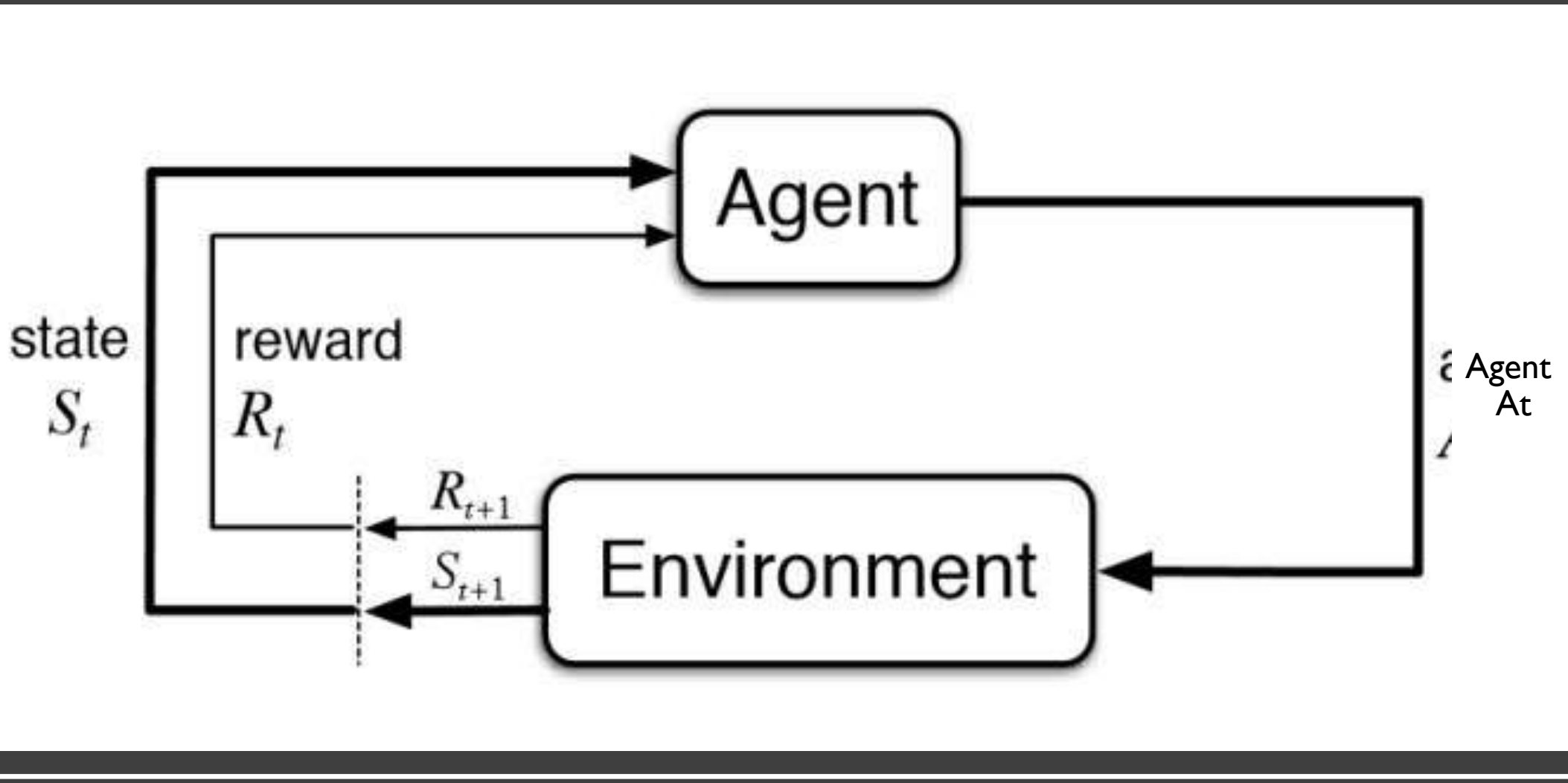
PHI 6385

Jonathan Simon

PROGRAMME

- 1) Qu’est-ce que l’apprentissage par renforcement et en quoi diffère-t-elle des autres formes d'IA ?
- 2) Qu'est-ce que l'apprentissage par renforcement a à voir avec l'agence réelle, celle qui nous intéresse dans la philosophie normative ?

QU'EST-CE QUE L'APPRENTISSAGE PAR RENFORCEMENT?


Agent
state
S_t
reward
R_t
Agent
At
R_{t+1}
S_{t+1}
Environment

LES BASES

- L'agent

observation à chaque pas de temps

action à chaque pas de temps

- L'environnement

états de l'environnement

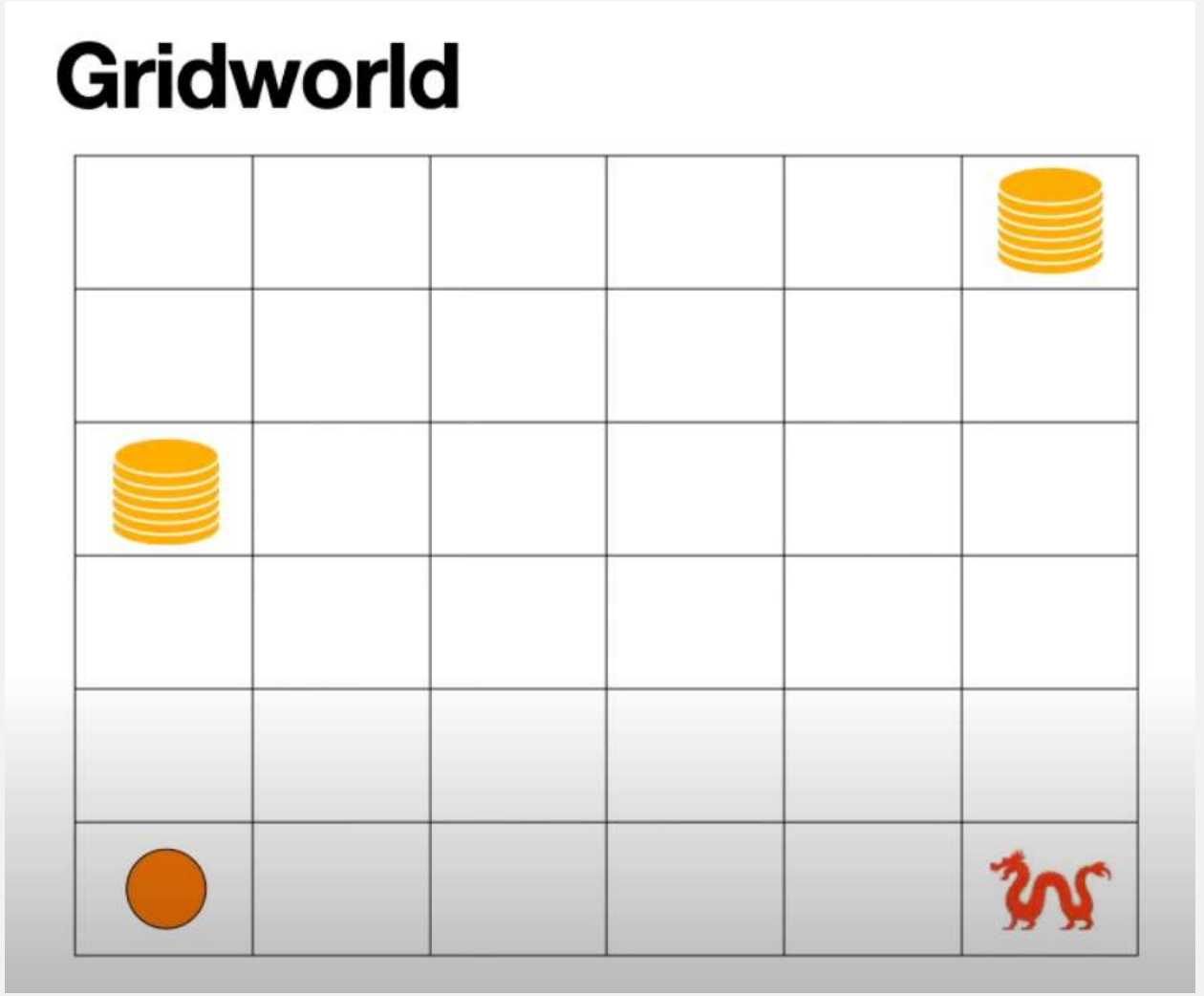
Récompenses (trouvées aux états)

LA TÂCHE D'APPRENTISSAGE

- La tâche d'apprentissage : nous voulons apprendre une fonction des états aux actions (un plan pour quelle action prendre étant donné son état, ou ce qu'il observe sur l'état)...
- L'objectif d'une politique est de maximiser la récompense globale (cumulative)
- Une « **politique» (policy function**)

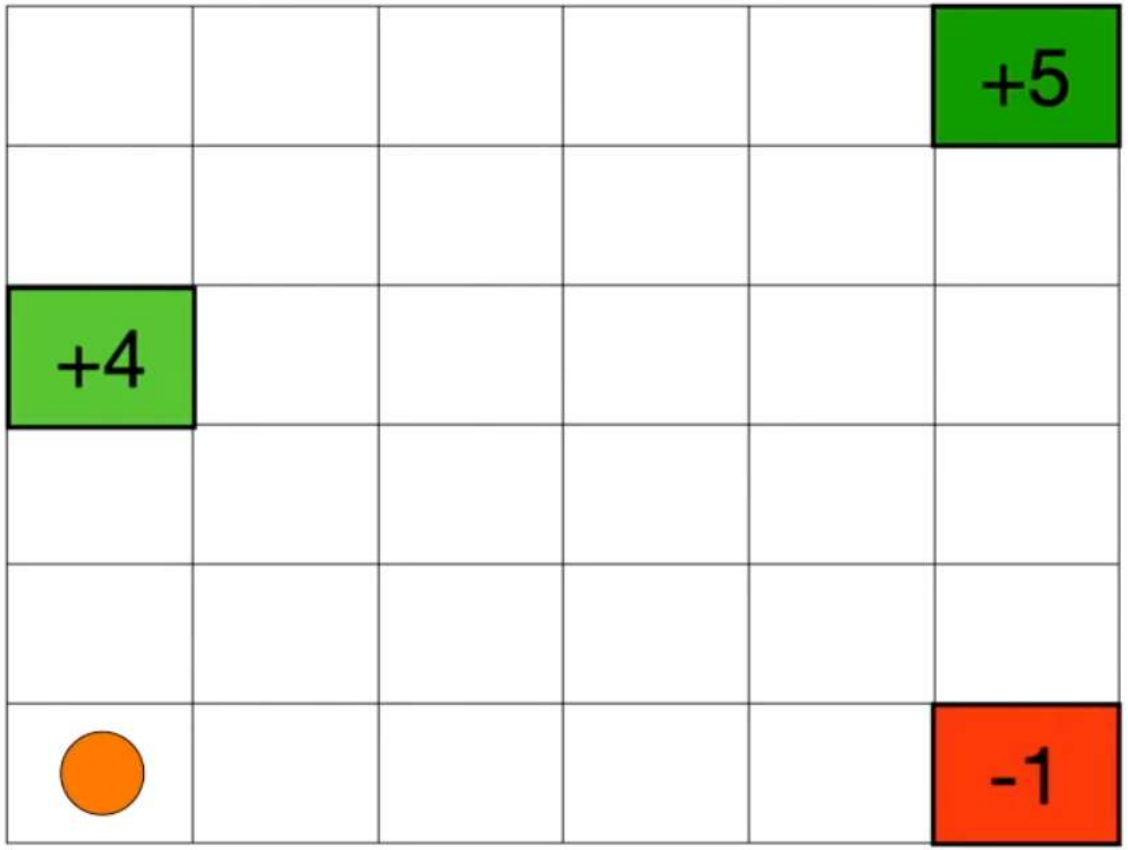
- Diapos suivant pris des videos de:
- Luis Serrano
- https://www.youtube.com/watch?v=SgC6AZss478

Gridworld



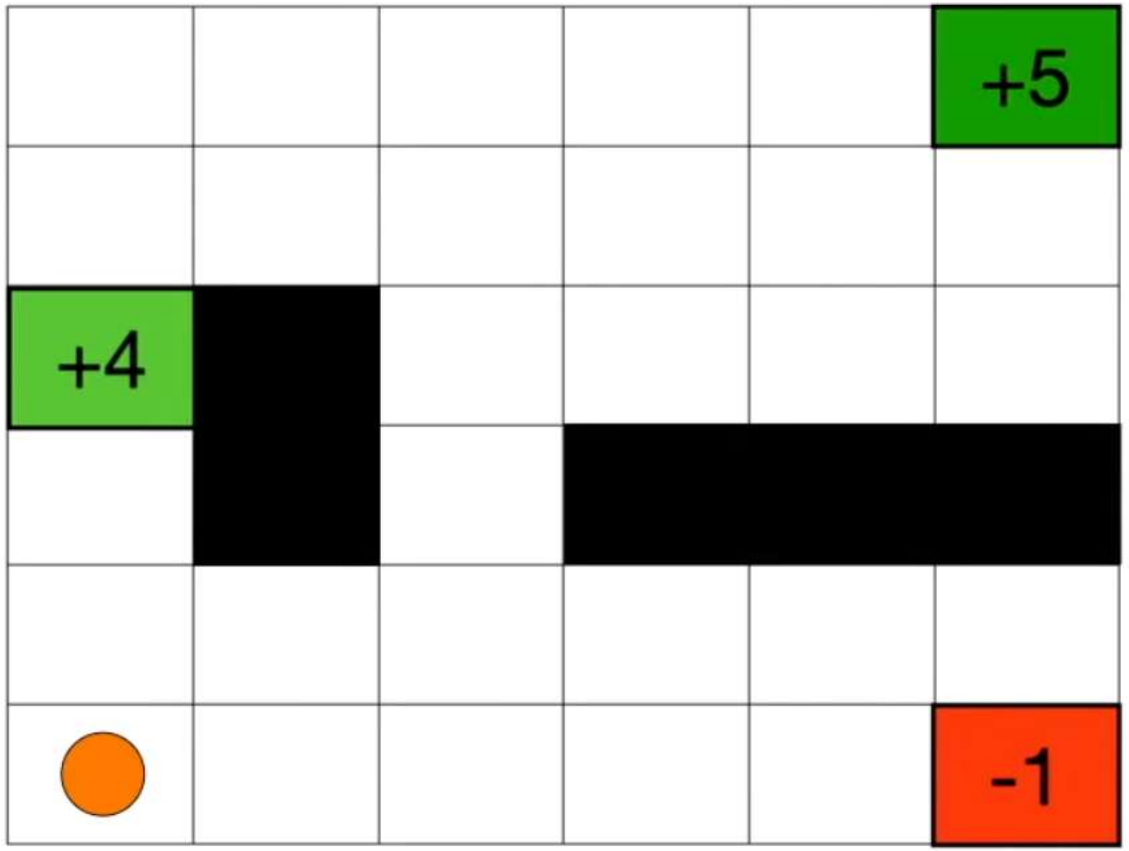
Luis Serrano

Gridworld

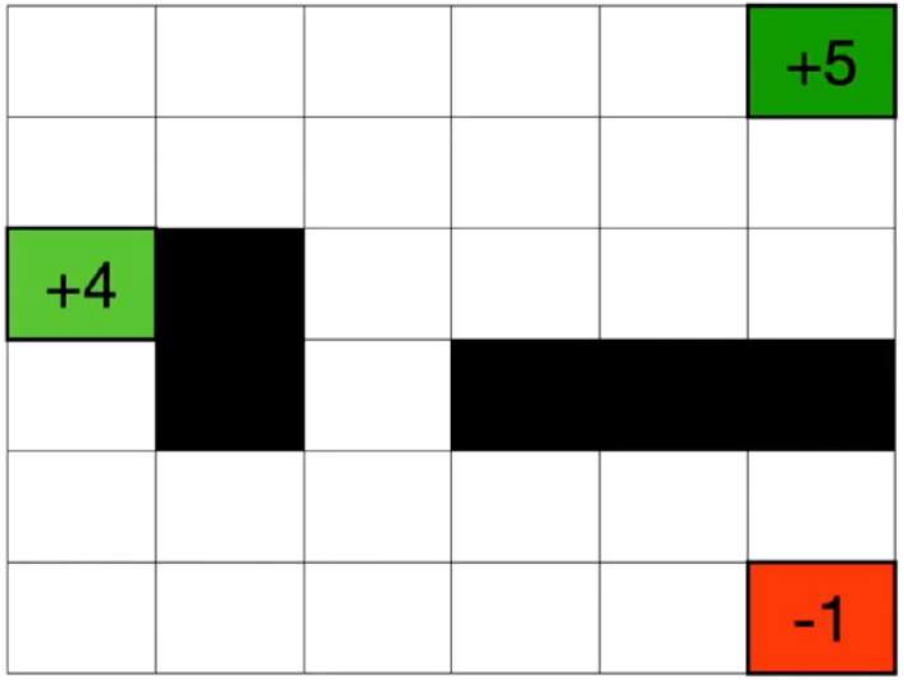


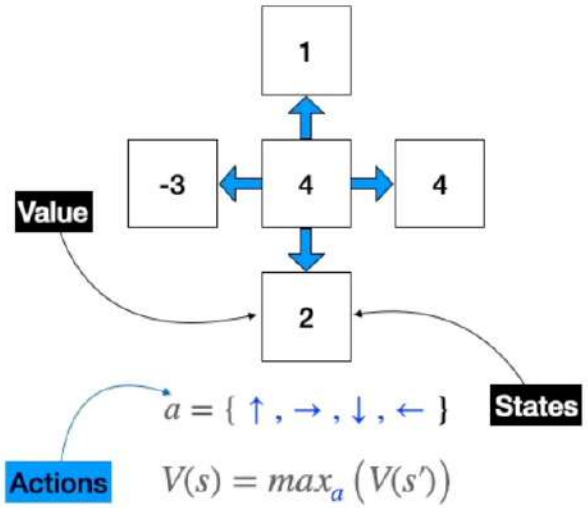
Luis Serrano

Gridworld



Value and policy





Luis Serrano

VALEUR

- $V(s) = Max_a V(s')$

- … La valeur d'un état est le plus que tu pourrais obtenir, à partir du meilleur choix que tu pourrais faire à partir de cet état, en passant au suivant. (une conception très optimiste)

VALEUR

- $V(s) = Max_a V(s')$

- Utile parce qu'il est récursif, tu peux travailler à partir d'un état final.

VALEUR

- $V(s) = Max_a V(s')$

- En fin de compte, tu finis, pour chaque état, avec la valeur attendue de cet état, la plus grande récompense que tu peux espérer avoir à la fin, en entrant dans cet état

VALEUR

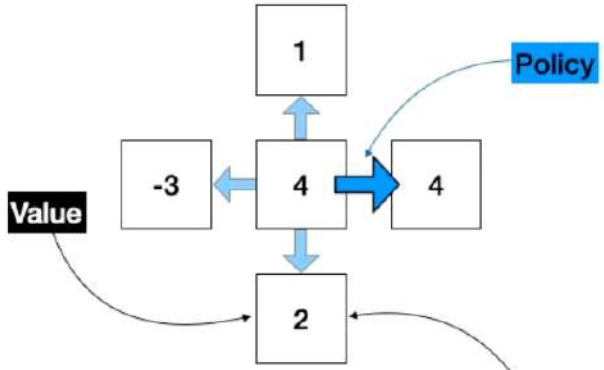
- $V(s) = Max_a V(s')$

- (Question : en additionnant le total de la récompense «collectée», additionnes-tu chaque carré ou seulement le dernier carré ?

- Réponse : Il s'agit plus d'une question d'interprétation informelle du cadre que d'une distinction mathématiquement importante.)

Value and policy

5	5	5	5	5	+5
5	5	5	5	5	5
+4		5	5	5	5
5		5			
5	5	5	5	5	5
5	5	5	5	5	-1



VALEUR

- $V(s) = Max_a V(s')$

- Nous devons modifier l'équation pour obtenir plus de structure et permettre au système d'en apprendre davantage sur le meilleur chemin à suivre.

VALEUR

- $V(s) = Max_a V(s')$

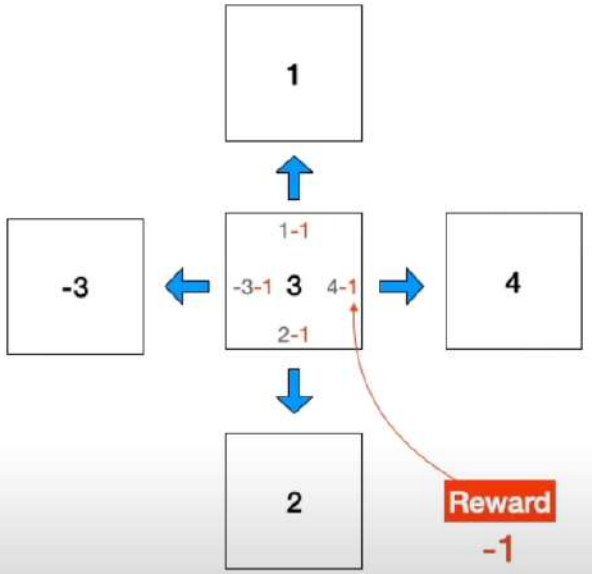
- Nous pouvons ajouter des récompenses supplémentaires (des miettes de biscuits, ou des pénalités... disons que le sol est en feu, donc tu veux minimiser le nombre de pas).

VALEUR

- $V(s) = Max_a V(s')$
- $V(s) = Max_a \ (R(s,a) + V(s'))$

Reward

2	1	2	3	4	+5
3	2	1	2	3	4
+4		0	1	2	3
3		-1			
2	1	0	-1	-2	-2
1	0	-1	-2	-2	-1



$$V(s) = max_a \left(R(s,a) + V(s')\right)$$

$$a = \{ \uparrow, \rightarrow, \downarrow, \leftarrow \}$$

VALEUR

- $V(s) = Max_a\ (R(s,a) + V(s'))$

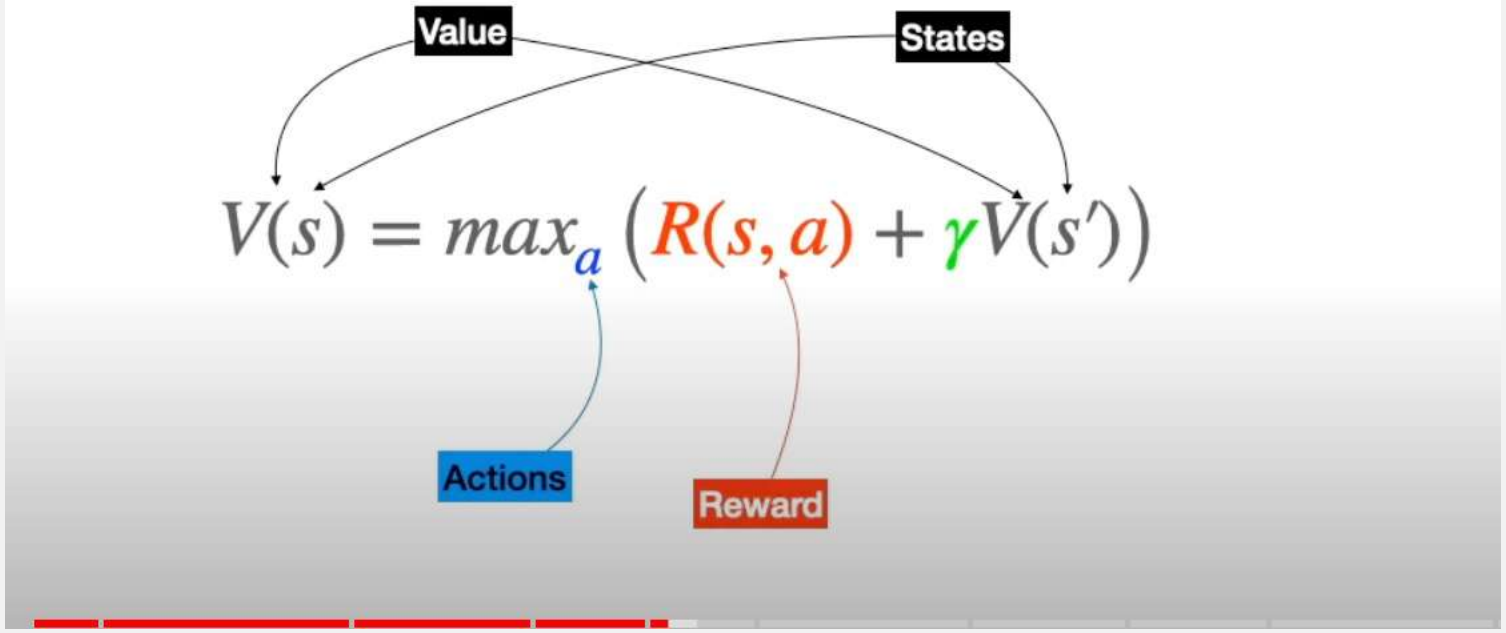
- Nous pouvons également escompter les récompenses futures : peut-être qu'un dollar demain ne vaut pas plus que 90 cents aujourd'hui

VALEUR

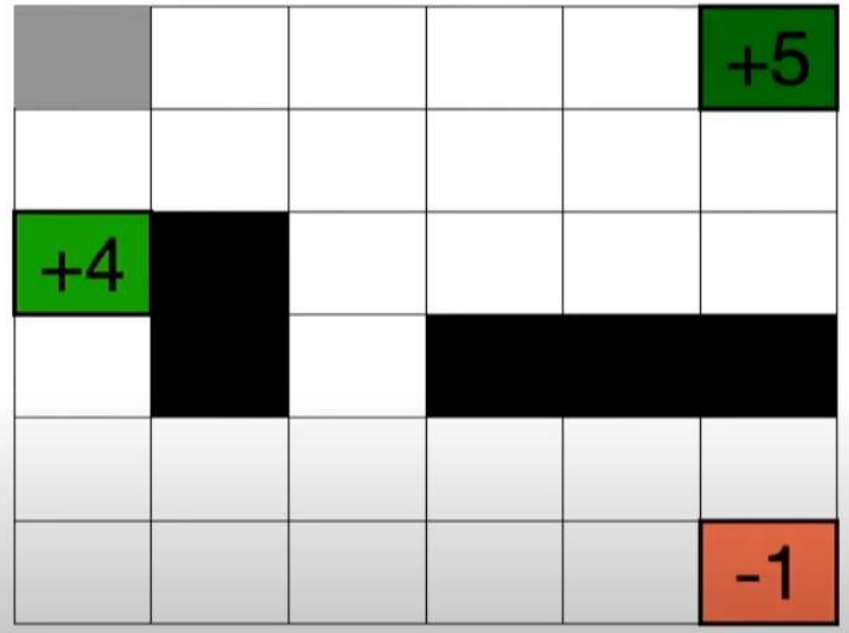
- $V(s) = Max_a\ (R(s,a) + \gamma V(s'))$

Bellman equation

$$V(s) = max_a \left(R(s,a) + V(s')\right) \qquad V(s) = max_a \left(\gamma V(s')\right)$$



How to solve the bellman equation?



VALEUR

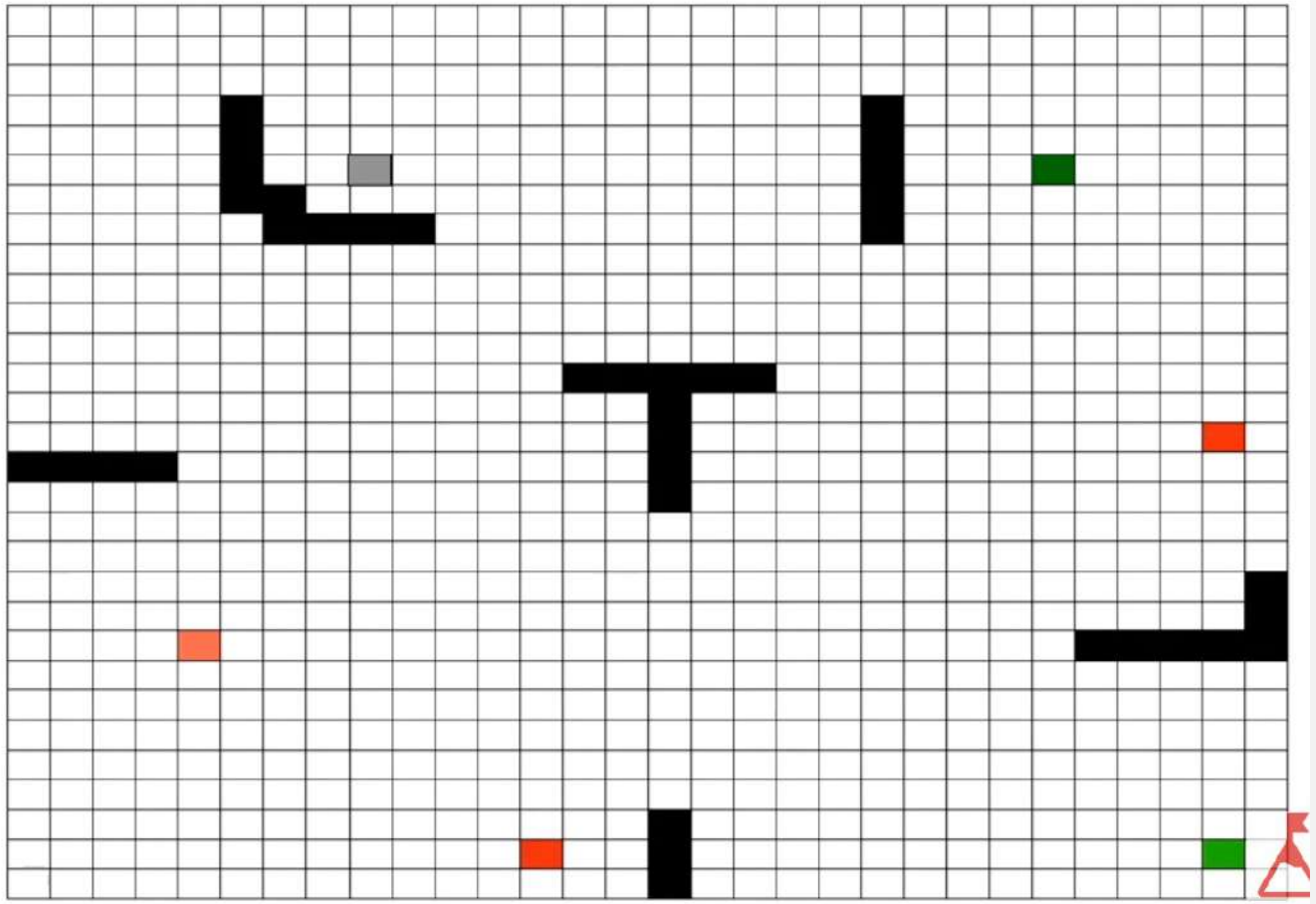
$V(s) = Max_a\ (R(s,a) + \gamma V(s'))$

- Explore, en commençant n'importe où au hasard, puis ajuste.

RESEAUX NEURONAUX?

Quel est le rapport avec les réseaux neuronaux ?

Problem



Luis Serrano

RESEAUX NEURONAUX?

Lorsque les choses deviennent compliquées (trop d'états du plateau, trop d'actions possibles)... il devient impossible de les résoudre à la main, c'est pourquoi l'approximation est importante.

RESEAUX NEURONAUX?

Deux approches principales : tu peux demander au réseau de trouver une fonction de valeur optimale, que tu peux ensuite convertir en une politique (apprentissage par la Valeur / value learning).

ou tu peux faire en sorte qu'il apprenne directement une politique (apprentissage de la politique / policy learning).

RESEAUX NEURONAUX?

L'idée générale est que les données glanées lors d'un tour de jeu servent d'exemple d'entraînement

RESEAUX NEURONAUX?

Remarque que cela signifie que l'on ne dispose pas de données d'apprentissage sous forme d'étiquettes : soit on implante les récompenses intrinsèques dans les environnements, soit on équipe le système d'un moyen de les trouver et on l'envoie apprendre.

Y. LeCun

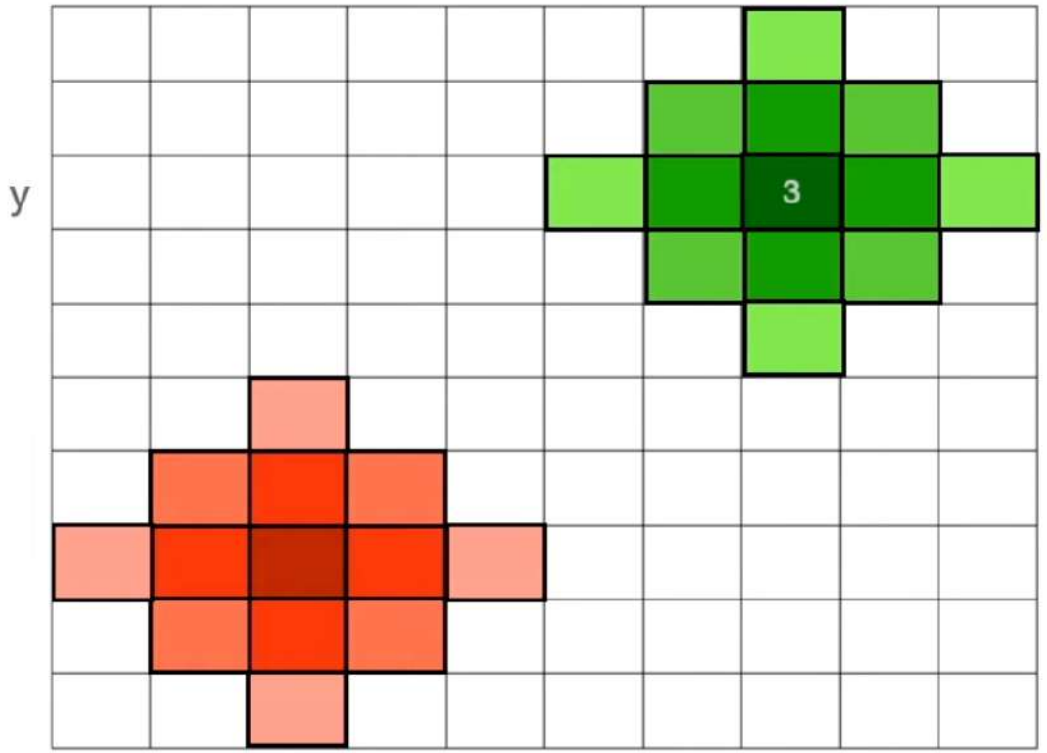
How Much Information is the Machine Given during Learning?

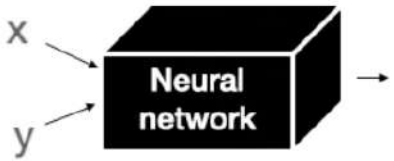
- **"Pure" Reinforcement Learning (cherry)**
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning (icing)**
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Self-Supervised Learning (cake génoise)**
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

© 2019 IEEE International Solid-State Circuits Conference

Yann Lecun

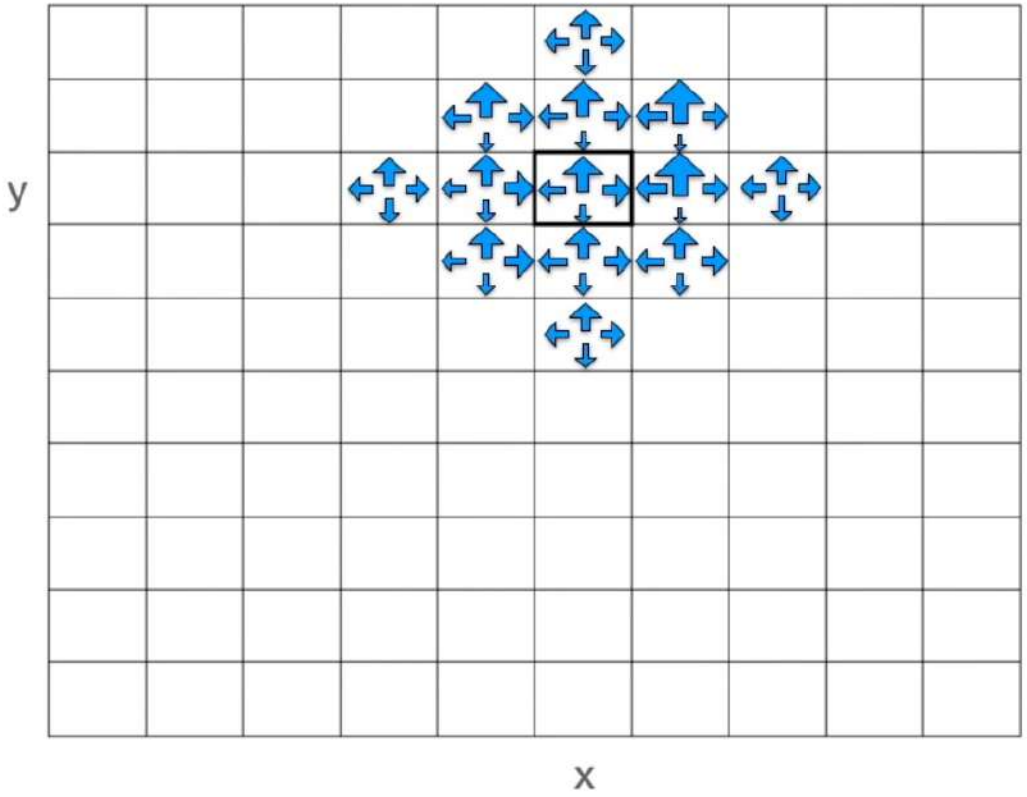
Value network





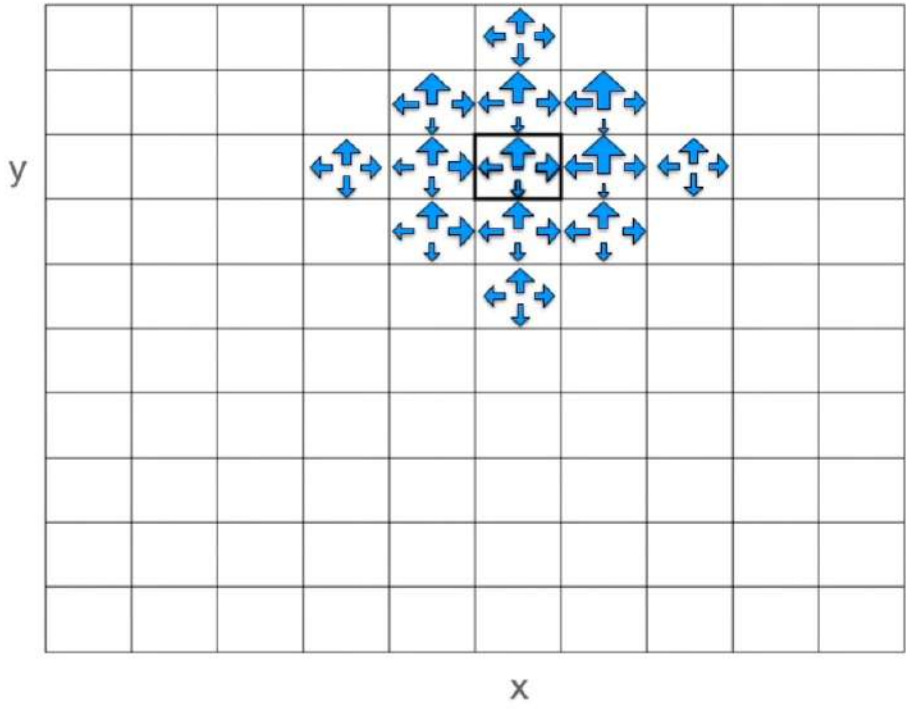
Luis Serrano

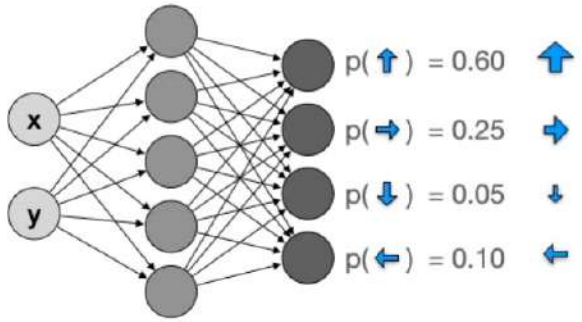
Policy network



Luis Serrano

Policy network

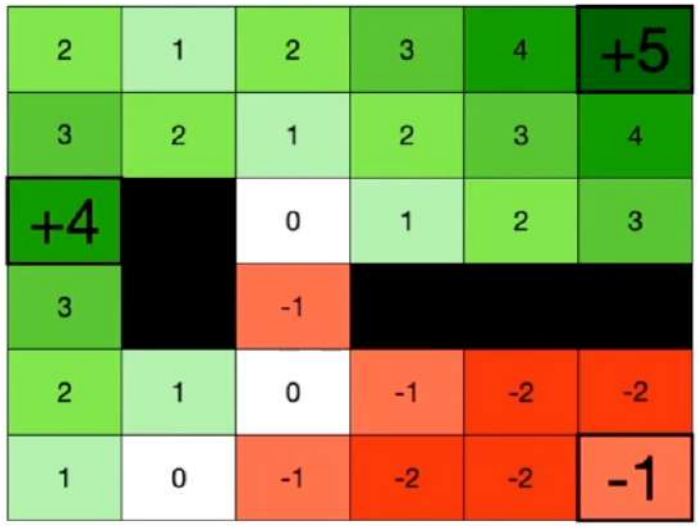




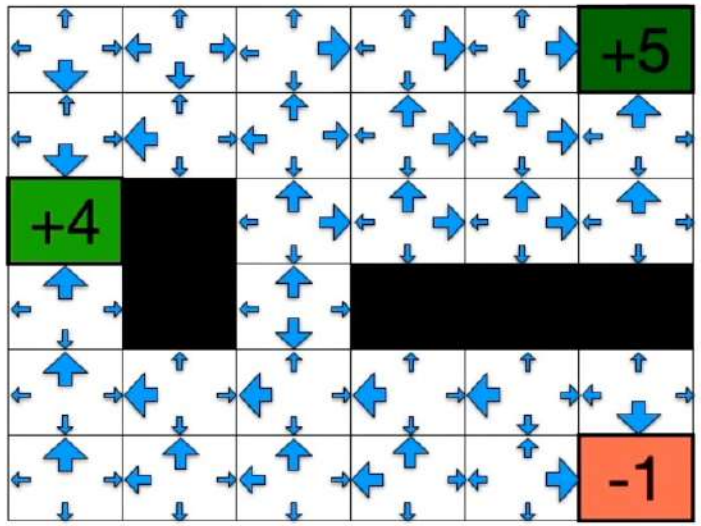
Luis Serrano

Two neural networks

Value neural network



Policy neural network



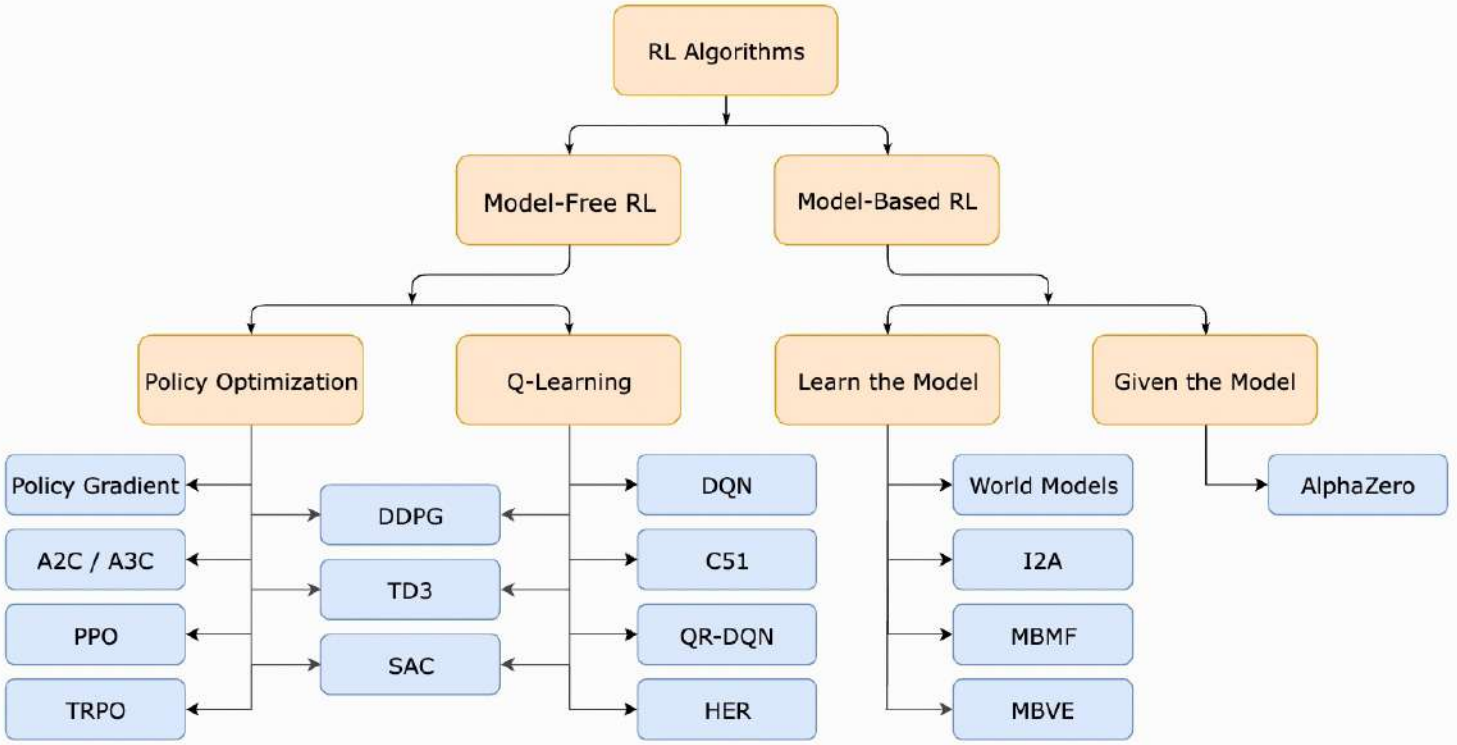
Luis Serrano

MODEL-FREE VS MODEL-BASED

Beaucoup d'algorithmes ici :

certains explorent simplement, d'autres utilisent des modèles de l'environnement, par exemple un arbre lui permettant de planifier plusieurs coups à l'avance (aux échecs ou au go).

A Taxonomy of RL Algorithms



A non-exhaustive, but useful taxonomy of algorithms in modern RL. Citations below.

RL ET AGENCE

RL ET AGENCE

- Comment cela informe-t-il exactement l'agence ?
- 1. l'agence n'est que la recherche d'une récompense
- 2. l'agence implique une action/un apprentissage intégré(e)
- 3. l'apprentissage politique représentatif
- 4. apprentissage politique basé sur un modèle représentatif

QUESTIONS OUVERT

L'origine de la récompense / de la valeur / de l'objectif?

Liberté et responsabilité?

Base biologique?

Trop minimale?

Conséquentialisme vs déontologie?

MAXIMISATION DES RÉCOMPENSES, ALIGNEMENT ET ÉTHIQUE DE L'IA

MAXIMISATION DES RÉCOMPENSES, ALIGNEMENT ET ÉTHIQUE DE L'IA

- Une fois qu'un agent RL est formé, il continuera à chercher à maximiser les valeurs en fonction de sa propre compréhension de ce qui est gratifiant.

- Et si cette compréhension est défectueuse ?

MAXIMISATION DES RÉCOMPENSES, ALIGNEMENT ET ÉTHIQUE DE L'IA

- Reward Hacking / Wireheading

MAXIMISATION DES RÉCOMPENSES, ALIGNEMENT ET ÉTHIQUE DE L'IA

- Perverse Instantiation