المدرسة الوطنية للذكاء الاصطناعي والرقمنة - بركان

ÉCOLE NATIONALE DE L'INTELLIGENCE ARTIFICIELLE ET DU DIGITAL - BERKANE

MACHINE LEARNING 2

Pr. BOUTAHIR Mohamed Khalifa

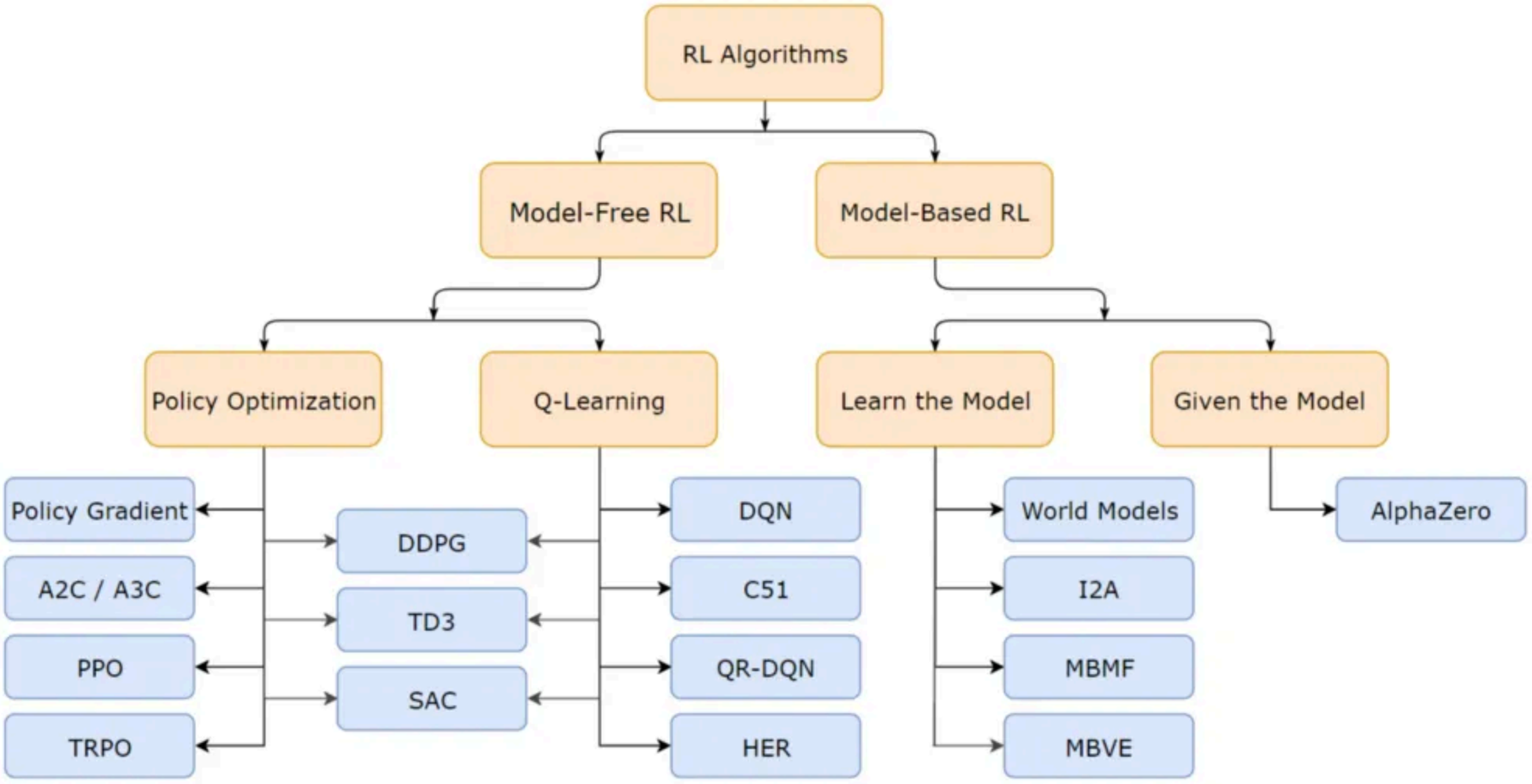
2024 – 2025

○ OBJECTIFS DE LA SESSION

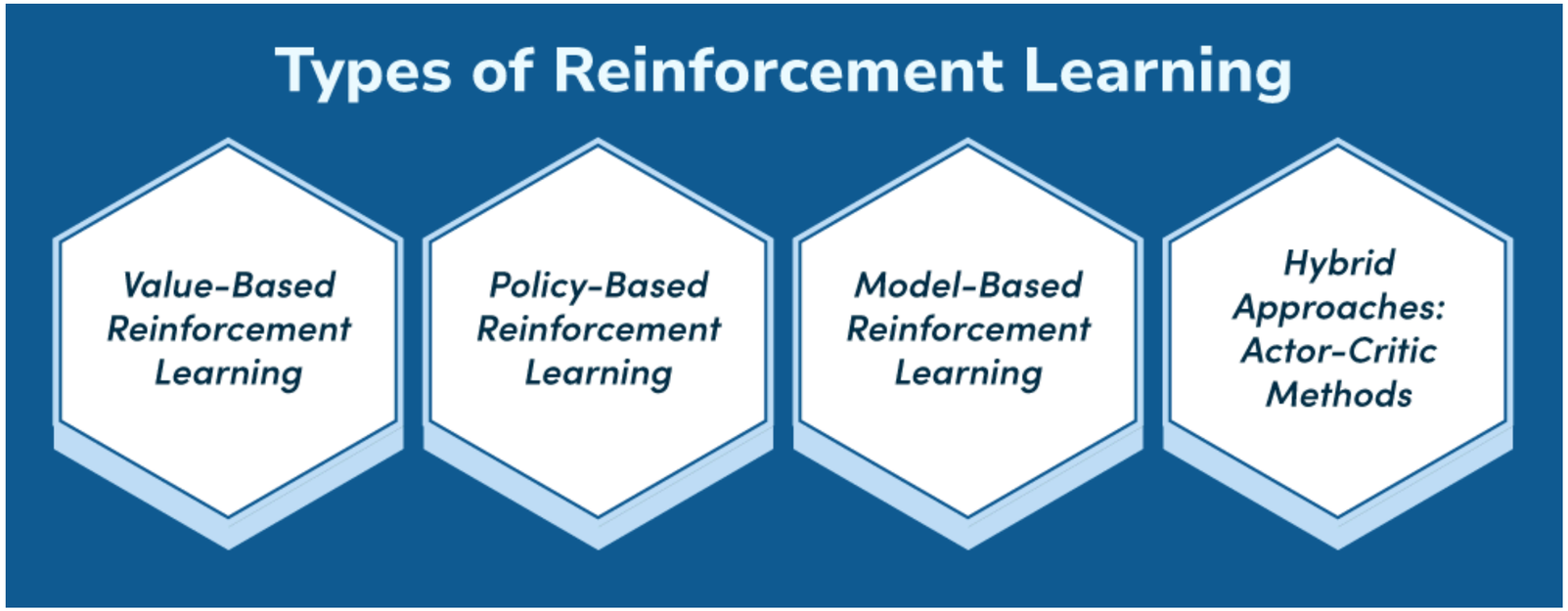
Dans les sessions précédentes, nous avons découvert les bases de l'apprentissage par renforcement (RL) et le cadre mathématique qui le structure (MDP). Aujourd'hui, nous allons voir comment un agent apprend réellement à optimiser ses décisions grâce aux premiers algorithmes.

✓ Comprendre comment un agent apprend dans un environnement RL
✓ Différencier les types d'apprentissage en RL
✓ Explorer les concepts clés : exploration vs exploitation, apprentissage des valeurs d'état et d'action

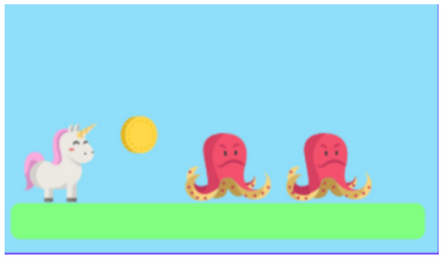
○ Les Types d’Apprentissage par Renforcement



- Les Types d’Apprentissage par Renforcement



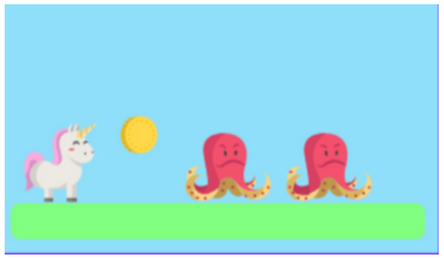
○ Les Types d’Apprentissage par Renforcement



Comment un agent apprend à prendre des décisions ?

Comment l’agent apprend à interagir avec l’environnement ?

○ Les Types d’Apprentissage par Renforcement



Comment un agent apprend à prendre des décisions ?

💡 Il doit choisir les bonnes actions

Comment l’agent apprend à interagir avec l’environnement ?

💡 Il doit choisir comment apprendre en fonction de sa connaissance du monde

○ Policy (Politique) π : Le cerveau de l'agent

Le processus de prise de décision de l'agent. Étant donné un **état**, une **politique** produira une **action** ou une distribution de probabilités sur **les actions**. En d'autres termes, étant donné une observation de **l'environnement**, une **politique** fournira une **action** (ou plusieurs probabilités pour chaque action) que **l'agent** devrait entreprendre.

State → **π (State)** → **Action**

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

L'objectif est de trouver u**ne politique optimale π** , c'est-à-dire **une politique** qui conduit à la **meilleure récompense cumulative** attendue.

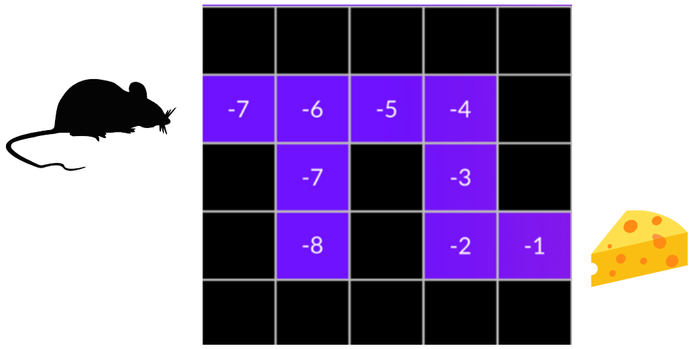
○ Valeur (Value) V : L'intuition de l'agent

La valeur mesure l'utilité d'un **état** pour **l'agent**. Elle représente **la récompense** attendue si **l'agent** commence dans cet **état** et suit une certaine **politique**.

En d'autres termes, plus **la valeur** d'un **état** est élevée, plus il est intéressant pour **l'agent** d'y être.

$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right]$$

- $v_\pi(s)$: la fonction de valeur
- $R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$: La somme des récompenses escomptées
- $S_t = s$: La condition initiale



L'objectif est que **L'agent** apprend à estimer les **valeurs** des **états** pour **prendre de meilleures décisions** et **maximiser sa récompense cumulée**.

○ Modèle (Model) M : La carte mentale de l’agent

Un modèle représente les règles de **l’environnement** : il permet à **l’agent** de prédire ce qui va se passer s'il effectue une **action** où d’une autre manière comment **l'environnement** réagit à les **actions** de **l’agent**.

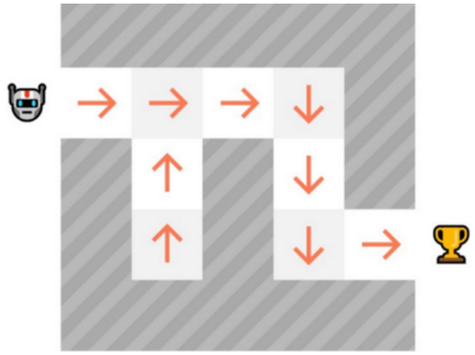
- Si **l’agent** a **un modèle**, il peut anticiper et planifier ses **actions** à l’avance.
- **Sans modèle**, il doit apprendre par **essai-erreur**, ce qui prend plus de temps.

○ Les Types d’Apprentissage par Renforcement

1️⃣ Comment un agent apprend à prendre des décisions ?

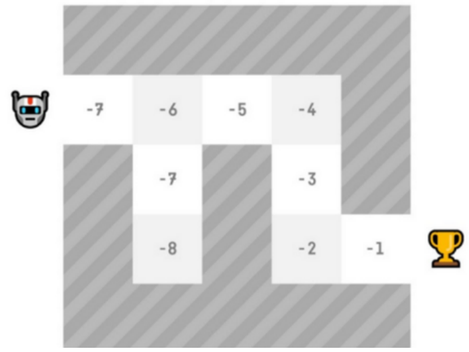
Policy-Based RL

L’agent apprend directement une politique pour choisir l’action optimale.



Value-Based RL

Entraîner une fonction de valeur pour apprendre quel état est le plus précieux et utiliser cette fonction de valeur pour prendre les mesures qui y mènent.



○ Les Types d’Apprentissage par Renforcement

1️⃣ Comment un agent apprend à prendre des décisions ?

Critères	Basé sur la politique (Policy-Based)	Basé sur la valeur (Value-Based)
Approche	Apprend directement une politique π qui mappe les états aux actions.	Apprend une fonction de valeur qui estime la qualité des actions dans chaque état.
Efficacité des échantillons	Moins efficace car il apprend uniquement à partir des trajectoires générées.	Plus efficace car il réutilise les expériences passées pour améliorer l’apprentissage.
Pourquoi choisir ?	Préféré lorsque l’action optimale est complexe ou nécessite des décisions continues.	Préféré lorsque l’on peut facilement évaluer la qualité d’une action dans un état donné.
Exemple réel	PPO pour le contrôle de robots (ex: main robotique OpenAI).	DQN pour les jeux Atari.

○ Les Types d’Apprentissage par Renforcement

2️⃣ Comment l’agent apprend à interagir avec l’environnement ?

Model-Free RL

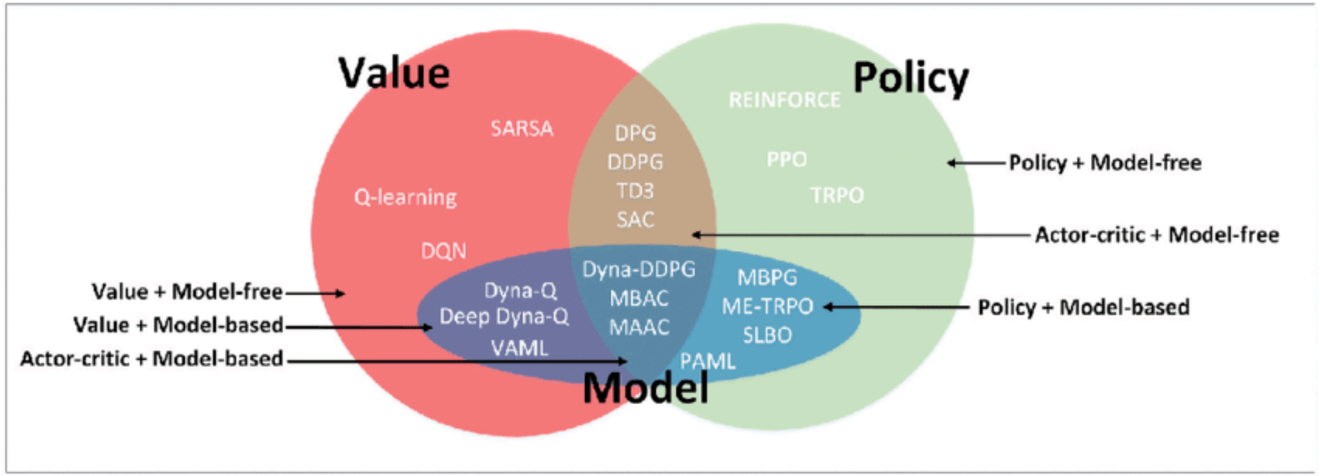
- L’agent apprend par essais-erreurs, sans connaître les règles de l’environnement.
- Quand un modèle du monde est disponible ou peut être appris efficacement.

Model-Based RL

- L’agent construit un modèle de l’environnement pour prédire les conséquences des actions.
- Quand il est difficile ou impossible d’apprendre un modèle précis.

○ Les Types d'Apprentissage par Renforcement

Type	Model-Based	Model-Free
Value-Based	Apprend un modèle et l'utilise pour estimer les valeurs des états/actions (ex : Dynamic Programming)	Apprend les valeurs directement sans modèle (ex : Q-Learning)
Policy-Based	Utilise un modèle pour optimiser directement la politique (ex : AlphaZero)	Apprend une politique sans modèle (ex : REINFORCE, PPO)

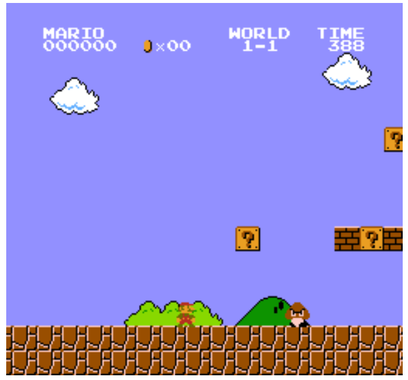


○ Exploration vs Exploitation

Types de tâches où Types problème d'apprentissage par renforcement

Tâche saisonnière où épisodique
Dans ce cas, nous avons un point de départ et un point d'arrivée (un état terminal). Cela crée un épisode : une liste d'états, d'actions, de récompenses et de nouveaux états.

Par exemple, pensez à Super Mario Bros : un épisode commence au lancement d'un nouveau niveau de Mario et se termine lorsque vous êtes tué ou que vous avez atteint la fin du niveau.



Tâches continues
Il s'agit de tâches qui se poursuivent indéfiniment (pas d'état final). Dans ce cas, l'agent doit apprendre à choisir les meilleures actions et à interagir simultanément avec l'environnement.

Par exemple, un agent qui effectue des opérations boursières automatisées. Pour cette tâche, il n'y a pas de point de départ ni d'état final. L'agent continue à fonctionner jusqu'à ce que nous décidions de l'arrêter.

○ Exploration vs Exploitation

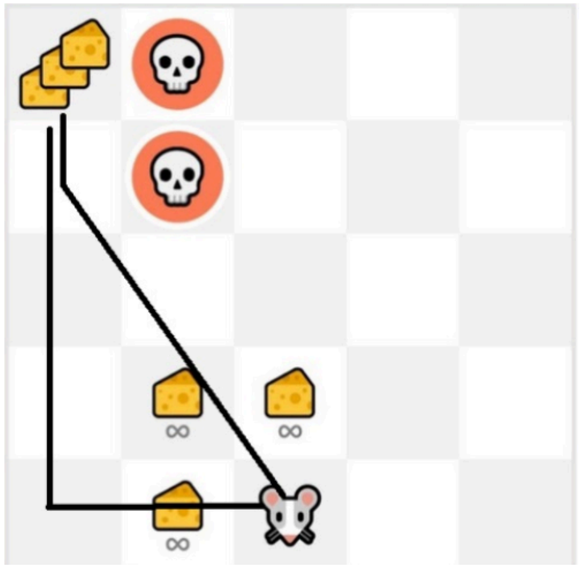
Concept	Exploration	Exploitation
Objectif	Découvrir de nouvelles stratégies, améliorer la compréhension de l'environnement.	Maximiser immédiatement la récompense en choisissant l'option connue comme la meilleure.
Avantage	Permet d'éviter de rester bloqué dans une stratégie sous-optimale.	Augmente les gains immédiats en utilisant les connaissances acquises.
Inconvénient	Peut ralentir l'apprentissage si trop fréquent.	Peut empêcher de trouver une meilleure solution sur le long terme.
Quand l'utiliser ?	Quand l'environnement est inconnu ou en début d'apprentissage.	Quand l'agent a déjà une bonne estimation des récompenses possibles.

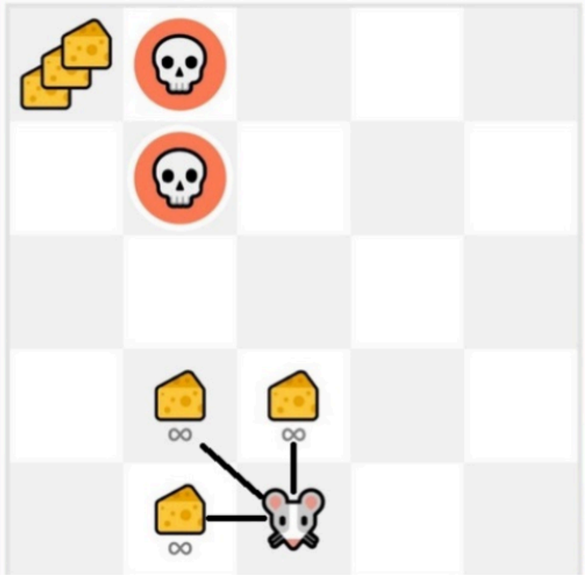
○ Exploration vs Exploitation

Concept	Exploration	Exploitation
Objectif	Découvrir de nouvelles stratégies, améliorer la compréhension de l'environnement.	Maximiser immédiatement la récompense en choisissant l'option connue comme la meilleure.
Avantage	Permet d'éviter de rester bloqué dans une stratégie sous-optimale.	Augmente les gains immédiats en utilisant les connaissances acquises.
Inconvénient	Peut ralentir l'apprentissage si trop fréquent.	Peut empêcher de trouver une meilleure solution sur le long terme.
Quand l'utiliser ?	Quand l'environnement est inconnu ou en début d'apprentissage.	Quand l'agent a déjà une bonne estimation des récompenses possibles.

Il y a un compromis exploration/exploitation.
Nous devons trouver un équilibre entre l'exploration de l'environnement et l'exploitation de ce que nous savons de l'environnement.

- Exploration vs Exploitation





- Exploration vs Exploitation