

Cours 1 : Introduction

Reinforcement Learning

Pourquoi l'apprentissage par renforcement ?



Des résultats époustouflants

Atari (2013), AlphaGo (2016), StarCraft (2019)



Solides fondations mathématiques

Une théorie de plus de 50 ans



Besoins faibles d'annotations

Une récompense donnée par l'environnement.



Vers une intelligence artificielle forte

Capacité à éprouver une forme de conscience, une compréhension de ses propres raisonnements.

Objectifs du cours



Connection entre le RL et les autres champs

Entre apprentissage supervisé et non supervisé



Paysage des algorithmes

Model-based vs model-free, exploration vs exploitation



Formalisme du RL

Champ de Markov, fonction de valeurs, politique



Comprendre les limites

Algorithmes instables, malédiction de la dimension, manque de garanties.

Agenda

1

2

Histoire du RL

Dans cette section, vous verrez comment le RL s'est transformé d'un champ de la psychologie en une industrie puissante.

Les bandits manchots

Un premier exemple d'algorithme : comment gagner au casino ?

Histoire du RL

Réflexe conditionnel pavlovien

Pavlov mesure en 1889 la salivation de son chien, qui se déclenche lorsque le chien mange. La salivation facilite la digestion (**action > réaction**). Mais à la vue même de la salle à manger, le chien commence déjà à saliver (**réaction > action**).

Pavlov parvient à répéter son expérience avec un large panel de stimulus : sifflets, cloches, ...

Le chien **anticipe un gain probable lorsqu'il se situe dans un certain état**.

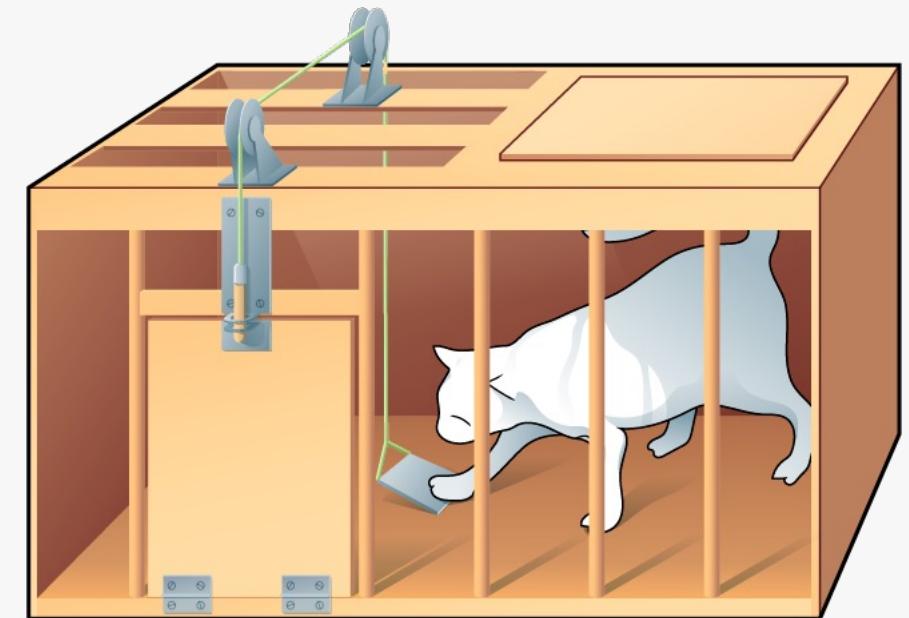


Apprendre par essais/erreurs

En 1911, Thorndike place un chat dans une boîte, qui s'ouvre en appuyant sur un levier. Le chat **explore erratiquement** une issue.

Lorsqu'il parvient à appuyer sur le levier, il acquiert une nouvelle connaissance. S'il est à nouveau enfermé, le chat va ré-appuyer immédiatement sur le levier.

Le chat a été capable quel a été l'étape clé à sa remise en liberté (**credit assignment**).



Réseaux de neurones

En 1992, Tesauro bat un nouveau record du monde sur backgammon en estimant la fonction de valeur à l'aide d'un réseau de neurones.



Atari (2013)

L'intégration des **réseaux de neurones profonds** change radicalement les algorithmes.



Go (2016)

DeepMind bat le champion du monde du jeu de Go, Lee Sedol. Un résultat inimaginable pour l'époque !

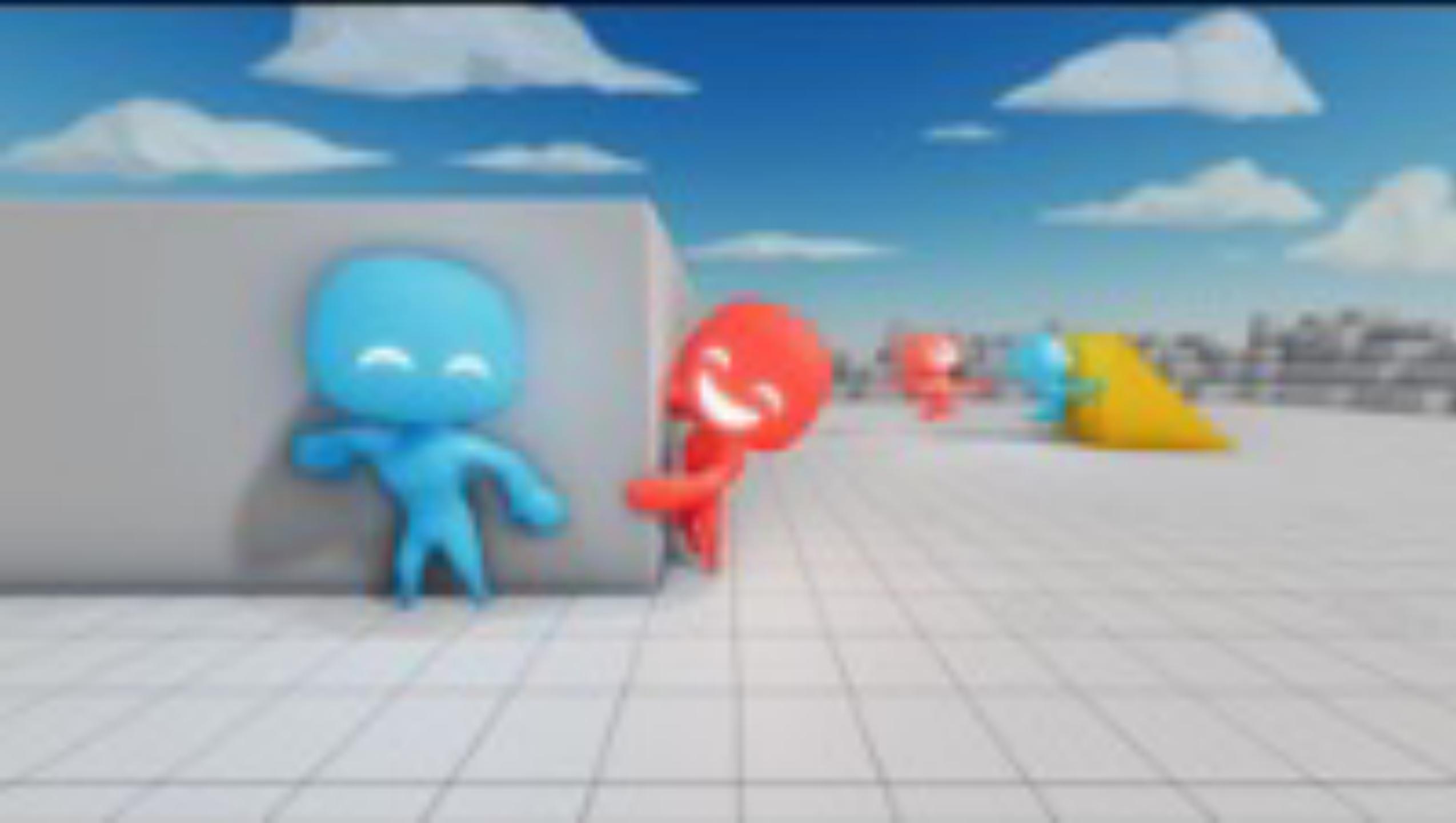


TI



DEEPMIND AI
LEARNED HOW TO WALK





La fonction de valeur

Elle évalue la **récompense** qu'un agent espère obtenir depuis son état courant.

Dans l'exemple d'un **grid world**, l'agent reçoit une récompense lorsqu'il atteint la croix en haut à gauche. A chaque fois qu'il l'atteint, il met à jour sa fonction de valeur, et progressivement, son estimation s'affine !



Bandit manchot

Comment gagner au casino ?

Présentation du modèle du **bandit manchot**. Mise en évidence du compromis entre **exploration et exploitation**.



Présentation

Un médecin a le choix entre 5 vaccins différents $X_{i,i<5}$. Il ne connaît pas l'efficacité e_i de chaque vaccin, mais il sait quel patient est devenu immunisé parmi les N premiers patients.
Un nouveau patient arrive. Quel vaccin lui donner ?

Exploitation : on estime l'efficacité de chaque vaccin ; le patient reçoit le vaccin le plus efficace.

Exploration : si N est trop petit, on donne le vaccin le moins testé.

Politique hybride ?



Minimiser le regret

$$r_N = N\mu^* - \sum_{k=0}^{N-1} R_k$$

N: nombre de patients testés

μ^* : récompense moyenne du meilleur vaccin (variable caché pour le médecin)

R_k : récompense obtenue pour le patient k (1 s'il est immunisé, 0 sinon)

1. Approche gloutonne

On estime quel vaccin produit la meilleure immunisation sur les n premiers patients :

$$\bar{R}_i = \frac{1}{T_i} \sum_{k=0}^{N-1} \chi_{v_k, i} R_k$$

R_k : récompense obtenue pour le patient k (1 s'il est immunisé, 0 sinon)

v_k : le vaccin utilisé pour le patient k

$\chi_{j,i}$: fonction indicatrice (vaut 1 si $i = j$, 0 sinon)

T_i : le nombre de fois que le vaccin i a été utilisé ($\sum_{k=0}^{N-1} \chi_{v_k, i}$)

On choisit le vaccin avec le meilleur \bar{R}_i pour tous les patients suivants.

2. Upper Confidence Bound

La moyenne empirique de récompense est estimée pour chaque vaccin :

$$\bar{R}_i = \frac{1}{T_i} \sum_{k=0}^{N-1} \chi_{v_k, i} R_k$$

On y ajoute un biais :

$$\hat{R}_i = \bar{R}_i + \sqrt{\frac{2 \log N}{T_i}}$$

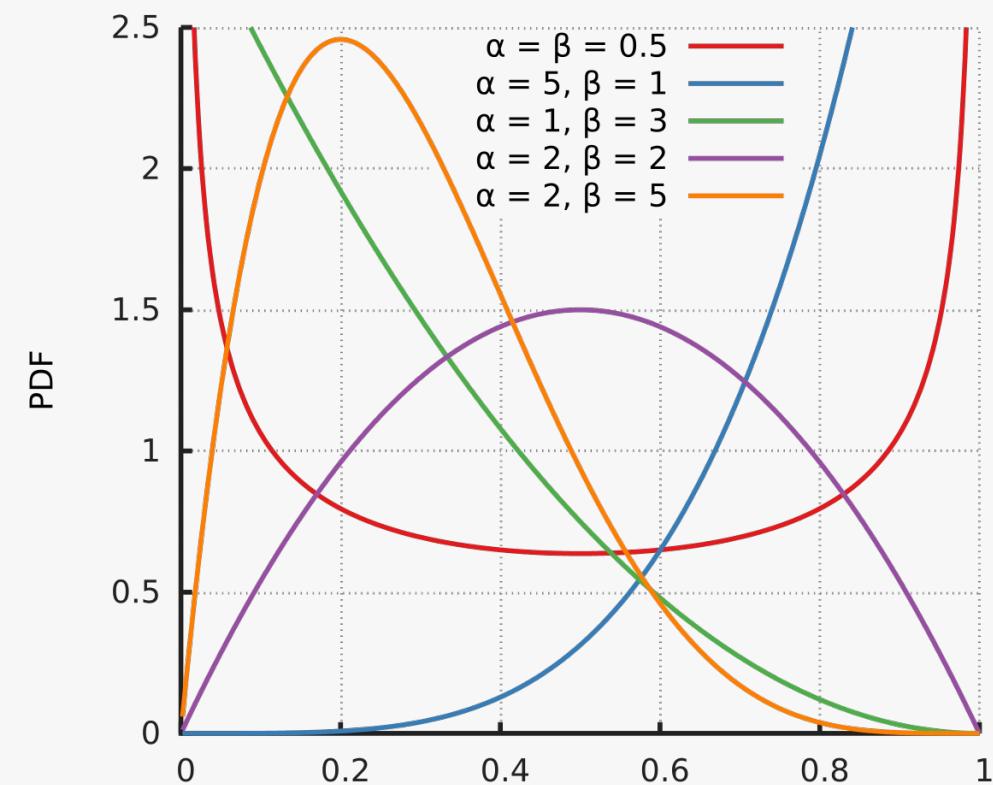
Que représente ce biais ?

3. Echantillonnage de Thompson

A chaque vaccin est associé un index, une variable aléatoire qui suit la loi bêta $\beta(a_j, b_j)$ dont les paramètres a_j et b_j sont initialisés à 1.

Pour chaque patient, on tire un index (on simule la loi aléatoire) pour chaque vaccin. Le vaccin utilisé est celui qui a le plus grand index.

Les paramètres de la loi bêta du vaccin choisi sont alors mis à jour : $a_j = a_j + 1$ s'il obtient la récompense et $b_j = b_j + 1$ sinon.





C'est l'heure
de pratiquer

A close-up photograph of a pile of walnuts on a light-colored wooden surface. In the foreground, one walnut shell is cracked open, revealing the white, oily nut肉. The background is filled with more whole and partially cracked walnuts.

En résumé...