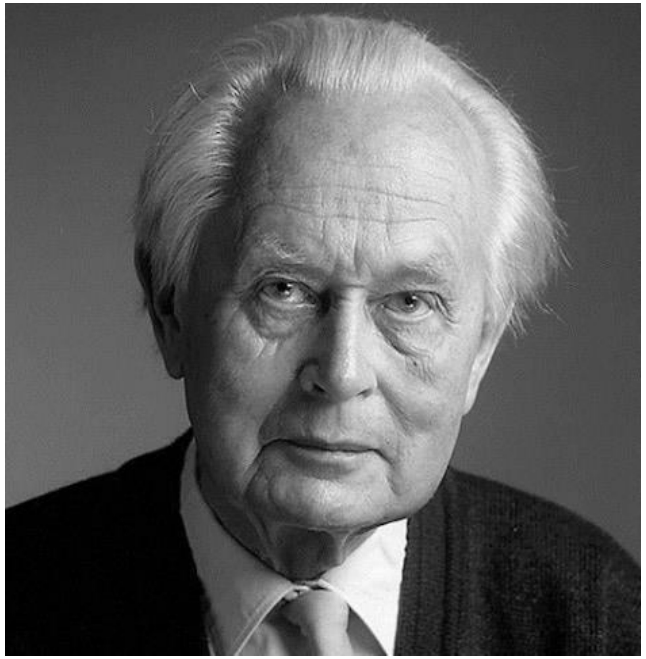
DES CASINOS À L’INTELLIGENCE ARTIFICIELLE

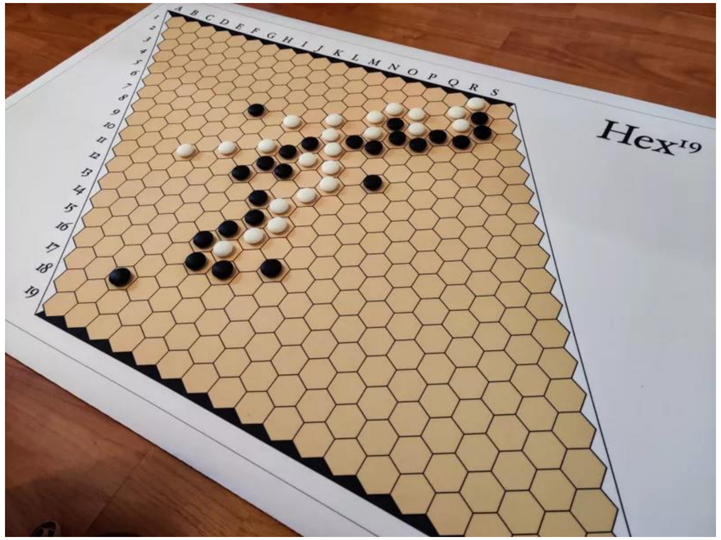
Travail Encadré de Recherche

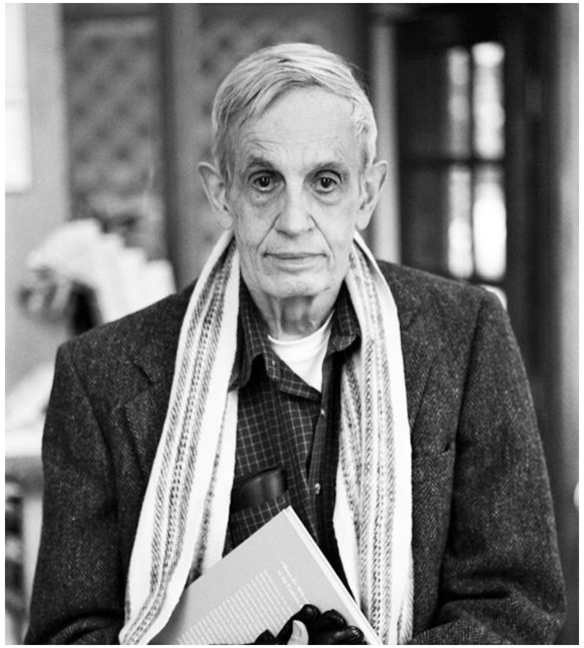
M1 DS

Le jeu de Hex



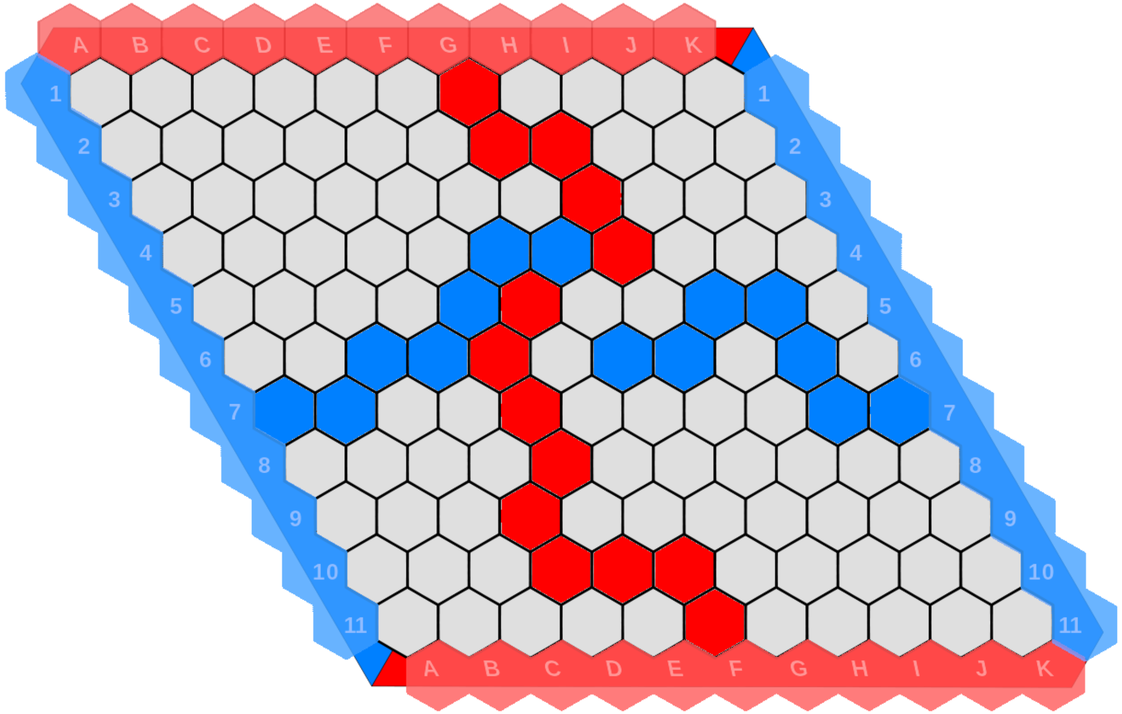
Piet Hein (1905 - 1996)

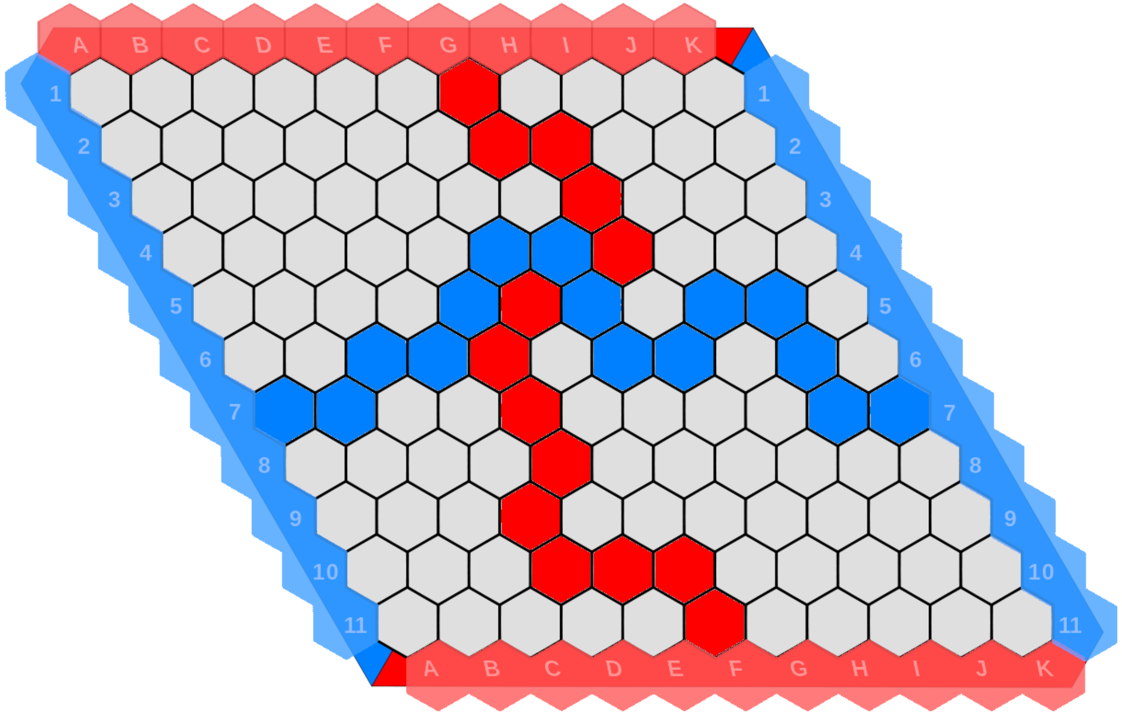




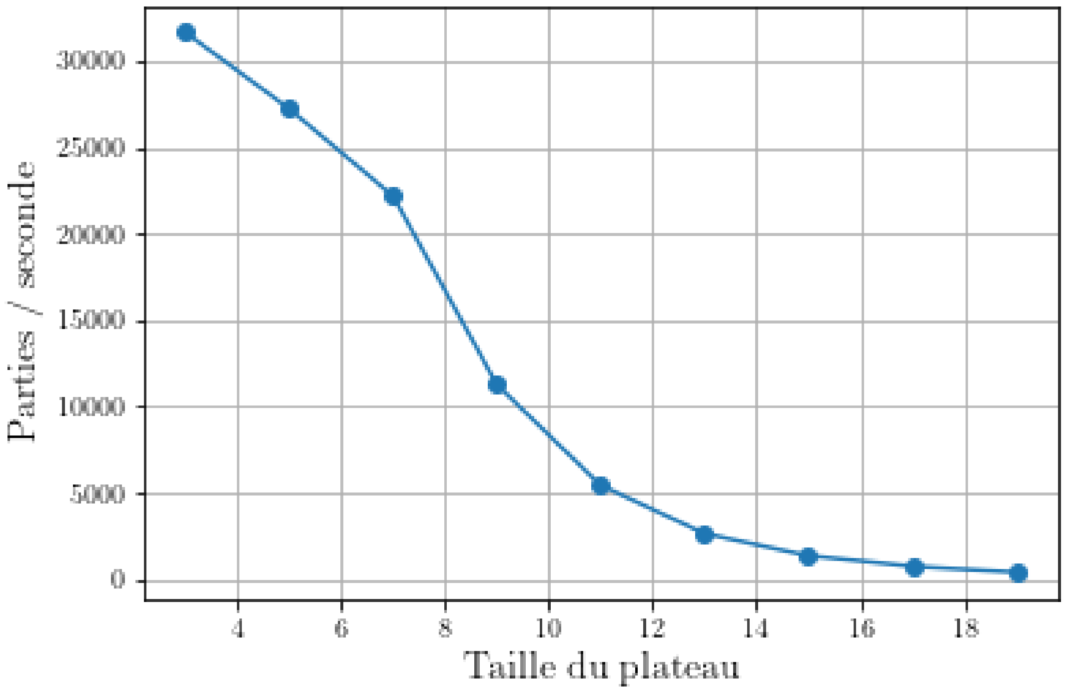
John Forbes Nash, Jr., (1928 - 2015)

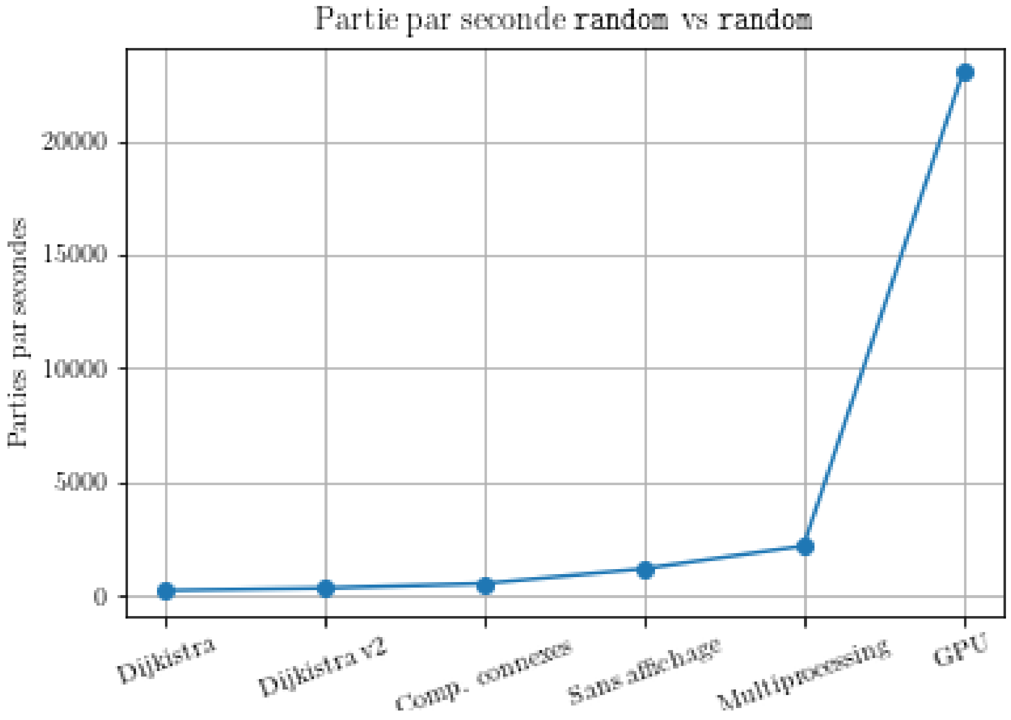
Condition de fin de partie



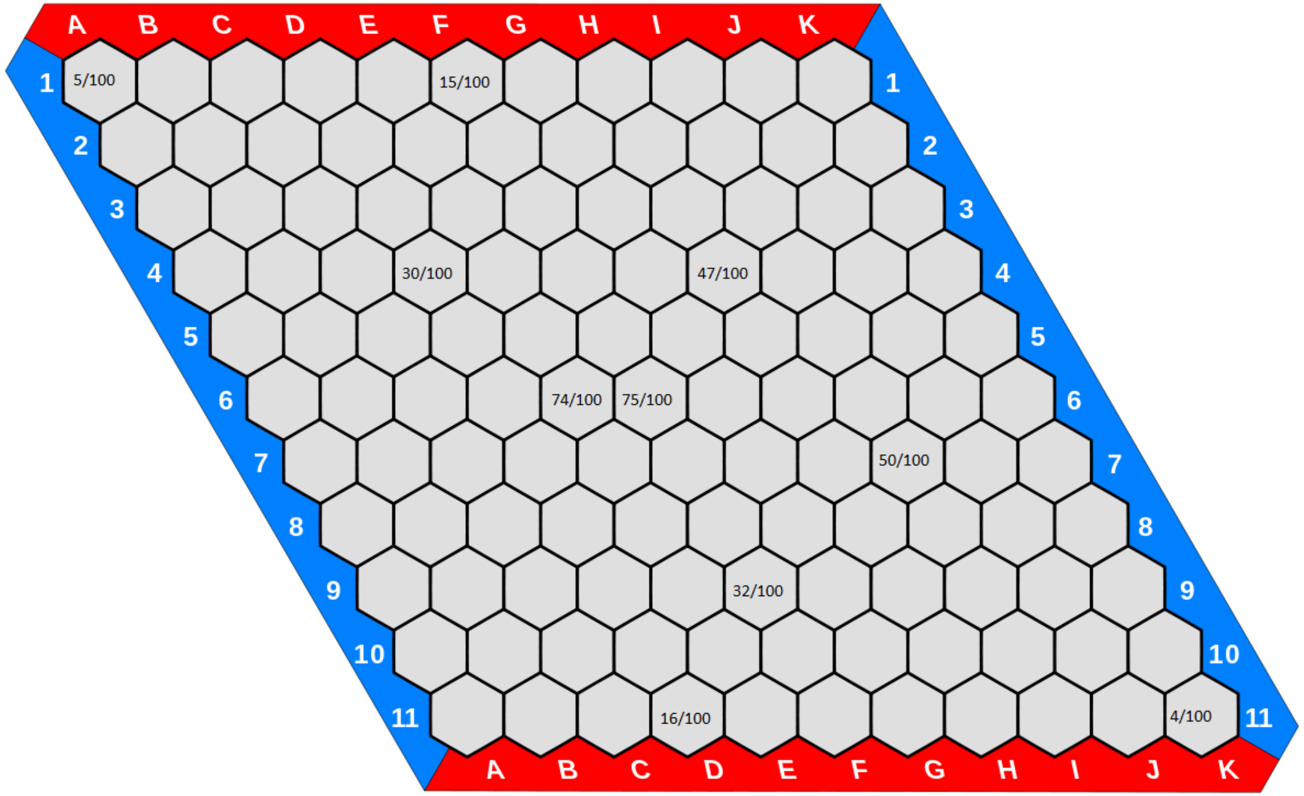


Condition de fin de partie





Première méthode : Monte-Carlo



UCB et problème du bandit manchot

Theorem 1. *For all $K > 1$, if policy* UCB1 *is run on K machines having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then its expected regret after any number n of plays is at most*

$$\left[8 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln n}{\Delta_i}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$$

where $\mu_1, \ldots, \mu_K$ are the expected values of $P_1, \ldots, P_K$.

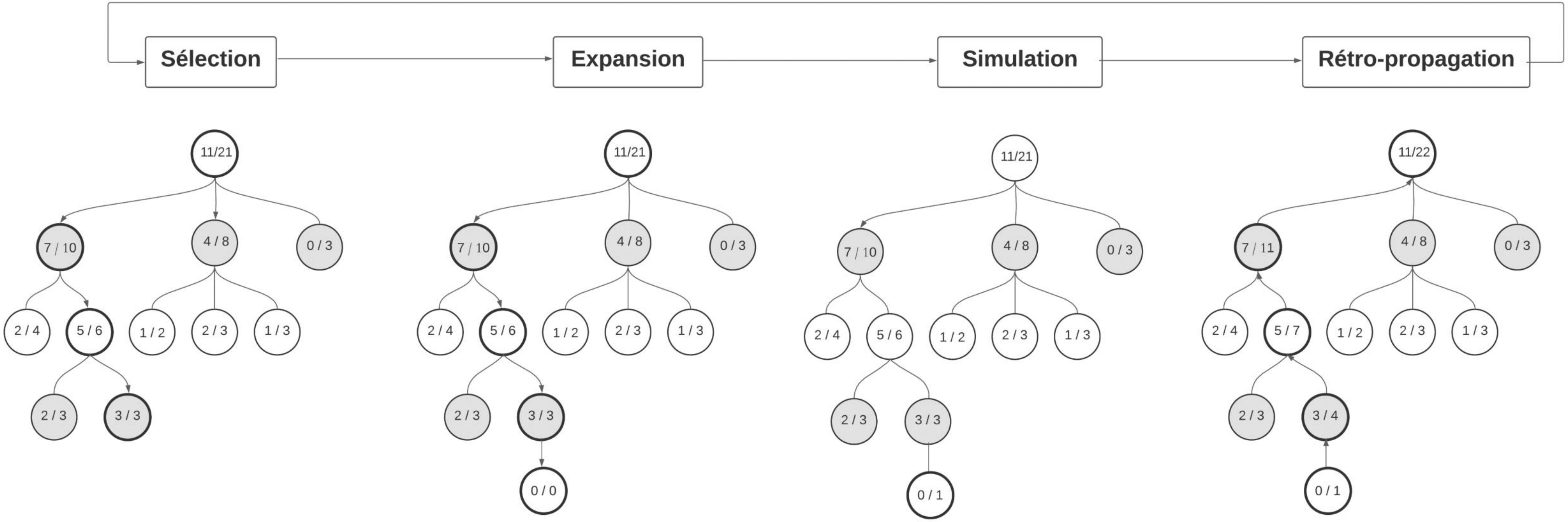
Deterministic policy: UCB1.
Initialization: Play each machine once.
Loop:

– Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

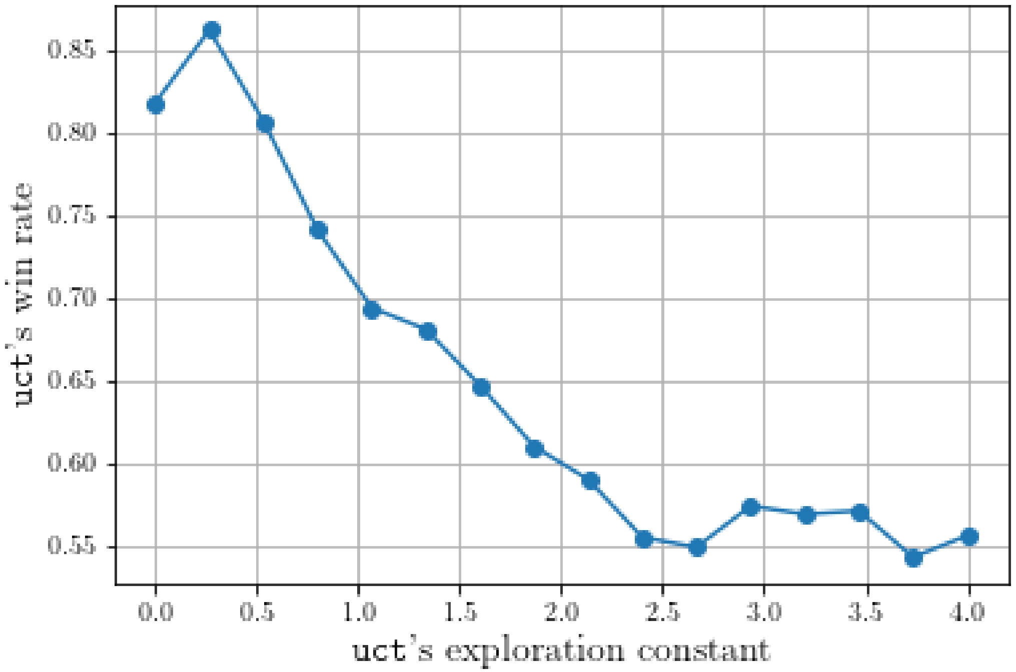
Figure 1. Sketch of the deterministic policy UCB1 (see Theorem 1).

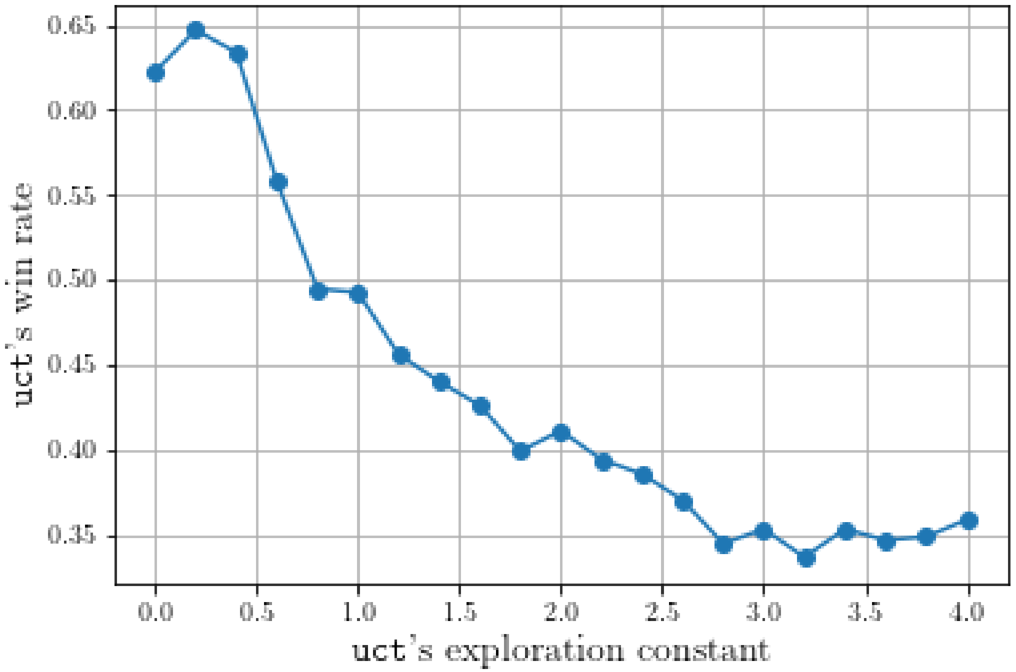
Finite-time Analysis of the Multiarmed Bandit Problem
Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer

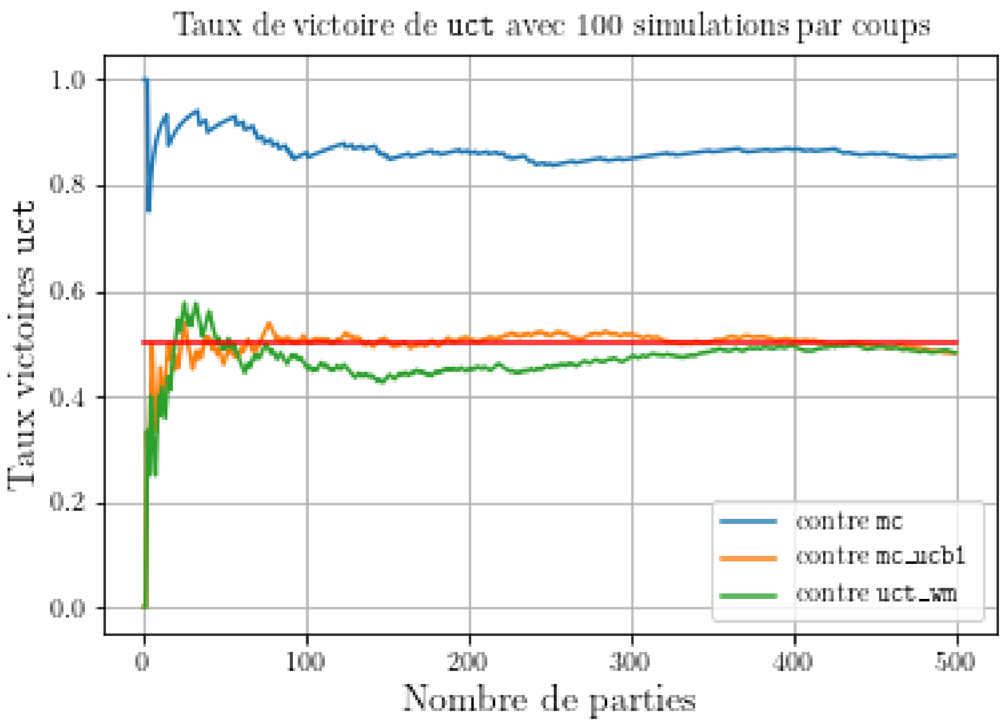
Upper Confidence Bound applied to Trees (UCT)



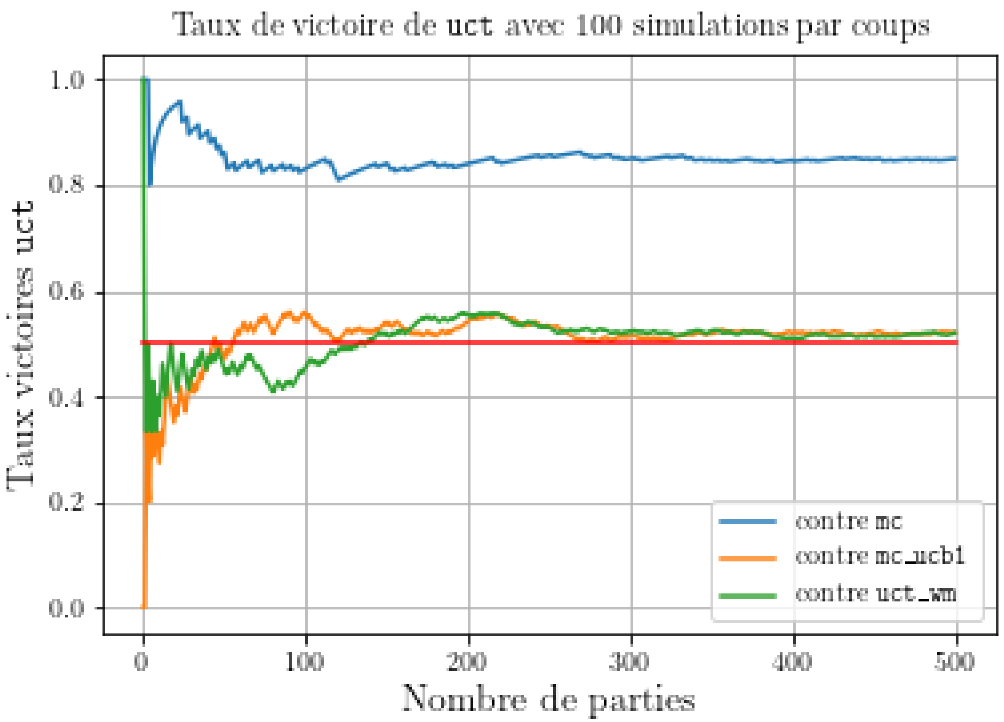
Constante d'exploration optimale



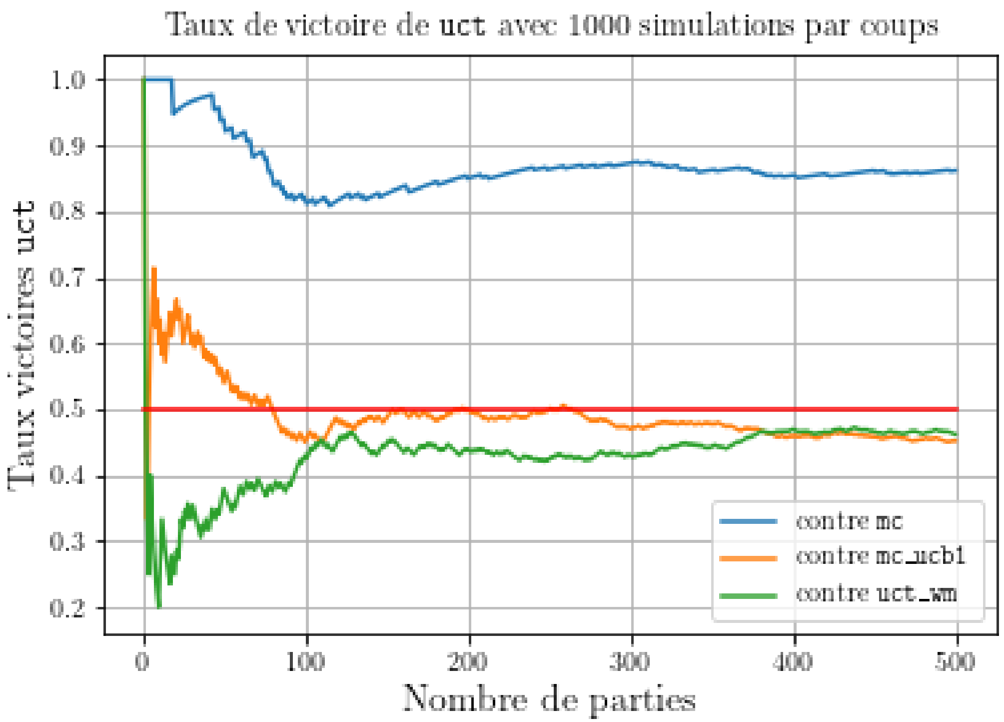




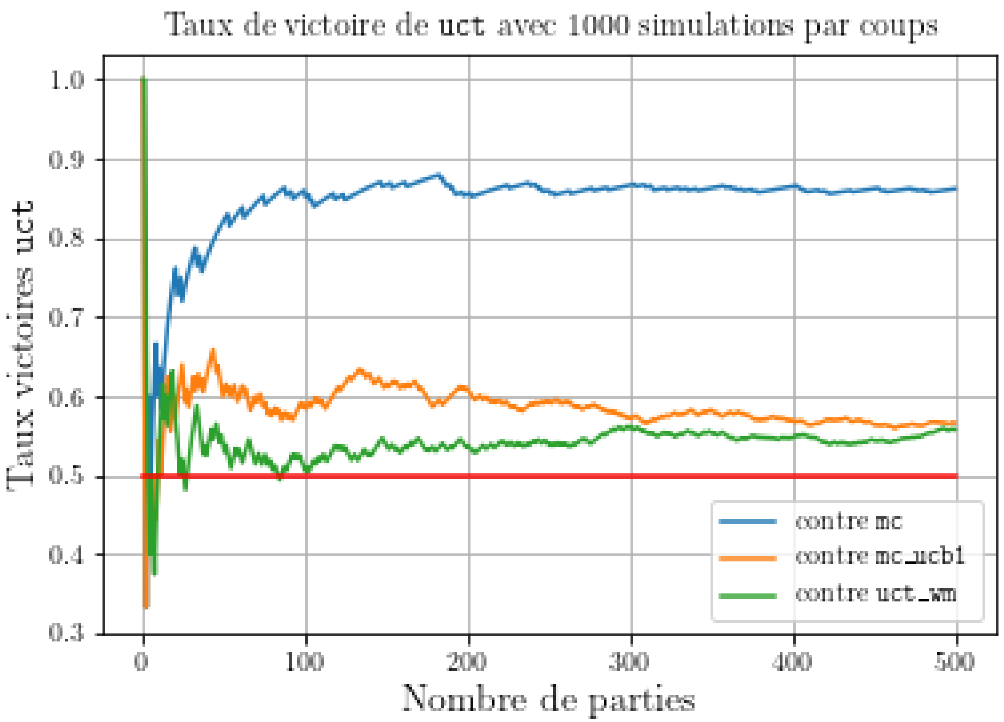
(a) uct joue en premier



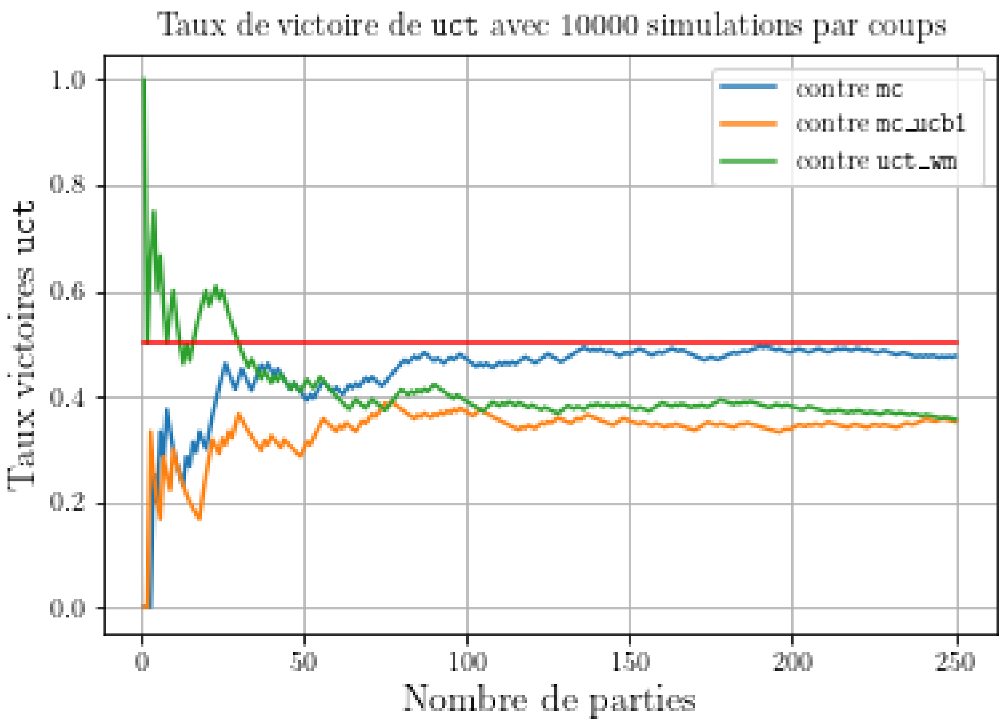
(b) uct joue en deuxième



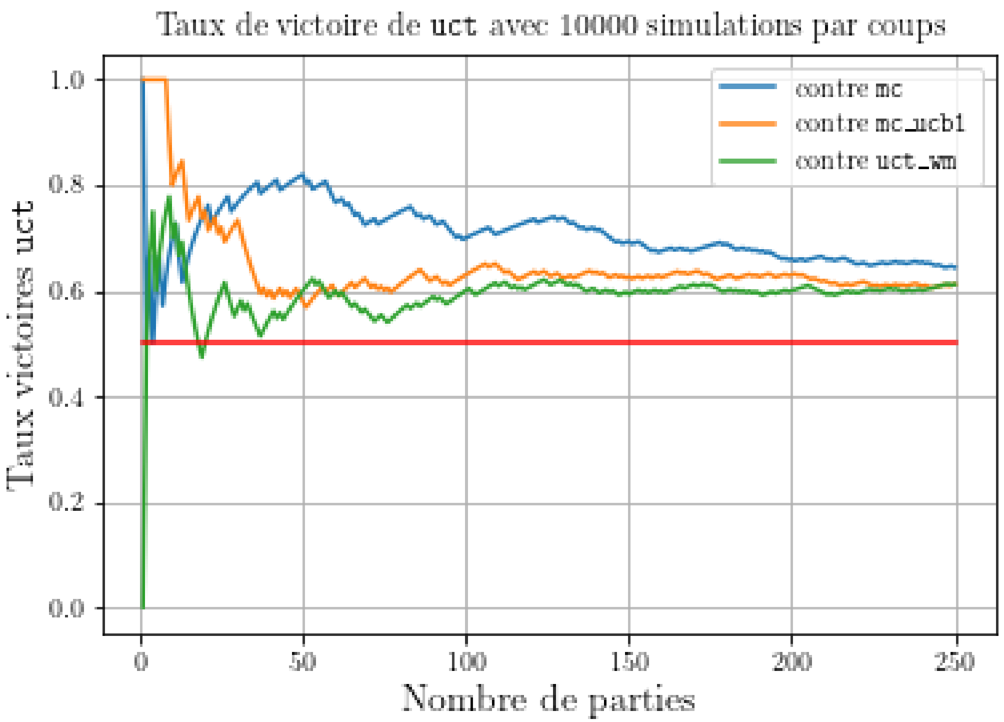
(a) uct joue en premier



(b) uct joue en deuxième



(a) uct joue en premier



(b) uct joue en deuxième

Axes d’amélioration

- Multiprocessing maître / esclaves pour UCT
- Début de partie
- Q-learning
- Connexion virtuelles, cases mortes, échelles