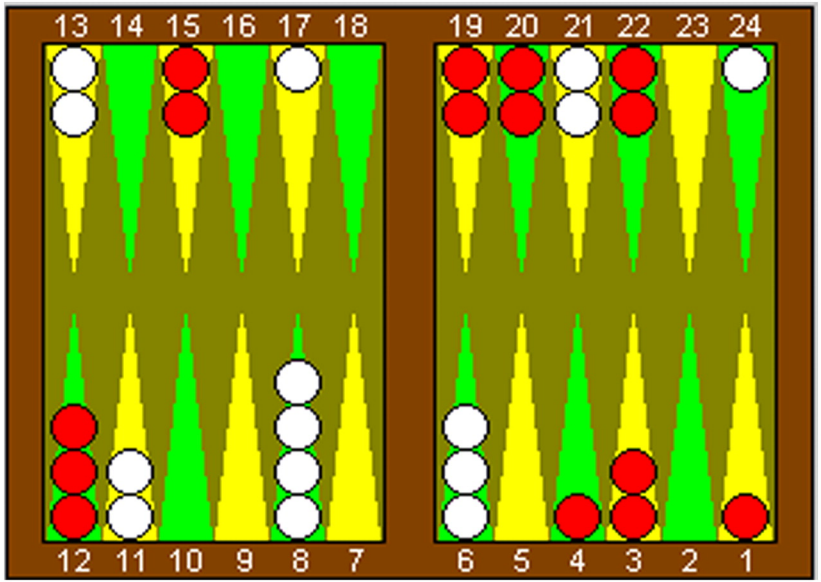
Objectifs

- Apprentissage par renforcement passif
  - méthode par estimation directe
  - méthode par programmation dynamique adaptative (PDA)
  - méthode par différence temporelle (TD)
- Apprentissage par renforcement actif
  - méthode PDA active
  - méthode TD active
  - méthode *Q-learning*
  - méthode par recherche de plan/politique (*policy-gradient*)
- Dilemme exploration vs. exploitation
- Généralisation en apprentissage par renforcement

Mise en situation

- Comment développer une intelligence qui apprend elle-même un jeu?



Mise en situation

- Comment apprendre un contrôleur d'hélicoptère?

Pourquoi l'apprentissage par renforcement?

- On a vu que l'**apprentissage automatique supervisé** permet de modéliser une expertise **à partir de données étiquetées**

- Pour obtenir un agent intelligent qui joue bien aux échecs, il faudrait amasser des paires (état du jeu, mouvement à jouer) d'un joueur expert
  - amasser de telles données peut être fastidieux ou trop coûteux

- On préférerait que l'agent apprenne seulement à partir du résultat de parties qu'il joue
  - si l'agent a gagné, c'est que son plan (sa politique) de jeu était bon
  - si l'agent perd, c'est qu'il y a une faiblesse derrière sa façon de jouer

Pourquoi l'apprentissage par renforcement?

- L'**apprentissage par renforcement** s'intéresse au cas où l'agent doit apprendre seulement **à partir de telles récompenses** ou **renforcements**

- L'apprentissage se fait à l'image d'un animal qui perçoit des récompenses négatives (douleur, faim) et positives (plaisir, manger)
  - l'animal veut maximiser les récompenses positives et éviter les négatives

Pourquoi l'apprentissage par renforcement?

- On a vu des algorithmes pour les processus de décisions markovien (MDP) qui trouvent le plan optimal qui maximise la récompense espérée
  - *value iteration*, *policy iteration*

- Ces algorithmes nécessitent une connaissance totale du modèle de transition $P(s'|s, a)$ et de la fonction de renforcement $R(s)$

Pourquoi l'apprentissage par renforcement?

- Dans un environnement réel, on ne connaît pas $P(s'|s, a)$
  - ex.: robot aspirateur placé dans une nouvelle pièce
  - ex.: agent qui contrôle un hélicoptère
  - ex.: agent qui joue à un jeu pour lequel $P(s'|s, a)$ est très complexe, avec un très grand espace d'état (Super Mario)

- Donc, l'apprentissage par renforcement vise aussi à **trouver un plan optimal, mais sans connaître le modèle de transition de l'environnement**

- Certains diront que c'est la forme la plus pure d'apprentissage en IA
  - c'est aussi une des plus difficiles à réussir...

Rappel: *Utility-based agents*