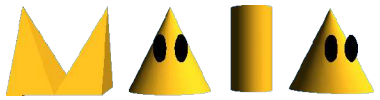


Introduction d'interactions directes dans les processus de décision markoviens

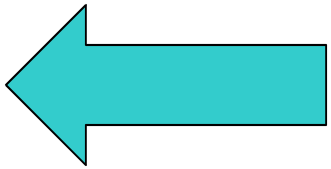
Vincent Thomas
Christine Bourjot
Vincent Chevrier

Présentation

- Travail en cours
- Systèmes multi-agents
  - Réactifs : règles stimulus-réponse
  - Sans mémoire
- Construction automatique de comportements
  - De manière décentralisée
  - Pour résoudre des problèmes collectifs
  - Dans un cadre coopératif

Plan

- Modèles markoviens
  - MDP
  - Extensions
- Notre proposition
  - Interac-DEC-MDP
  - Formalisme
- Exemples
- Résolution
- Conclusion



MDP

- MDP Markov Decision Process = <S,A,T,R>
  - S ensemble d'états
  - A ensemble d'actions
  - T matrice de transition : évolution du système stochastique
    - T: S x A ↙ P(S)
  - R récompense : fonction à optimiser
    - R: S x A ↙ P(Re)

- Un MDP = un problème de décision

Mono-agent

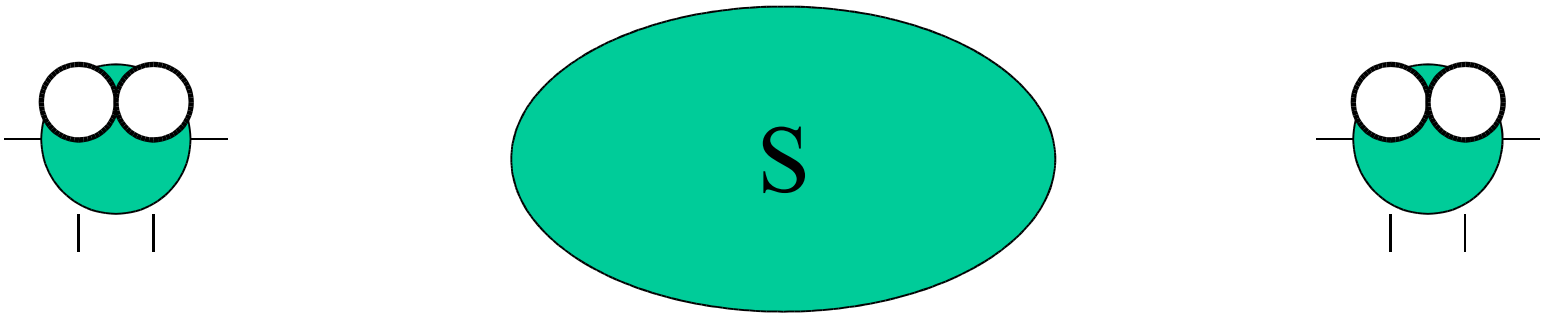
  - Trouver politique (comportement réactif) π: S ↙ P(A)
  - Qui maximise la somme des récompenses à long terme

- Algorithmes pour construire politique
  - Planification (value iteration, …)
  - Apprentissage (Q-learning, …)
  - Trouve politique **optimale**

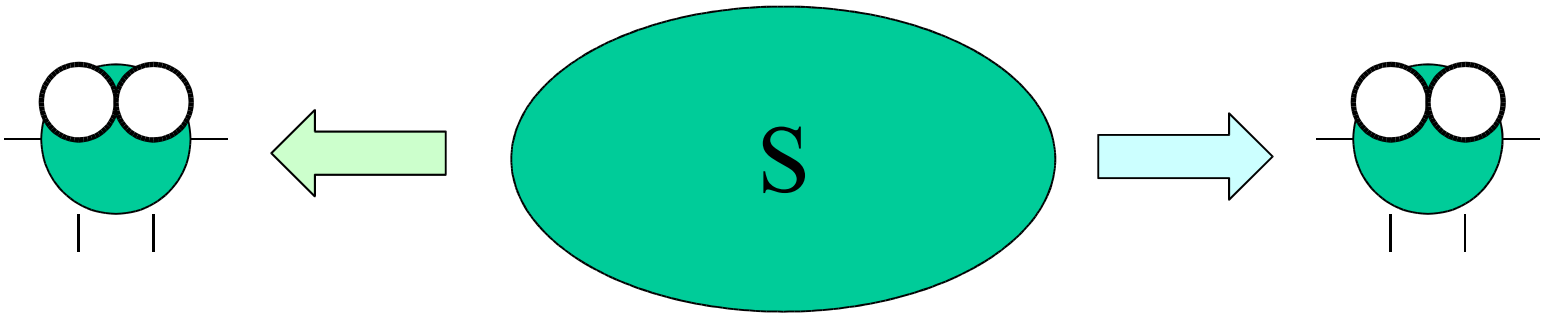
Extensions des MDPs

- DEC-MDP : Decentralized-MDP
- Formalisme pour problème de décision Multi-agent
  - Représenter agents réactifs
    - Exécution décentralisée et simultanée
    - Observabilité partielle
    - Fonction de Observations vers Actions : π_i: S_i ↙ $P(A_i)$
  - Représenter problème sous forme d'un processus
    - Matrice de transition
      - T : S x A_1 x A_2 x A_3 x … ↙ P(S)
    - Fonction de récompense
      - R : S x A_1 x A_2 x A_3 x … ↙ P(Re)
  - Actions des agents vues comme influences sur processus
  - Objectif: Maximiser la somme des récompenses

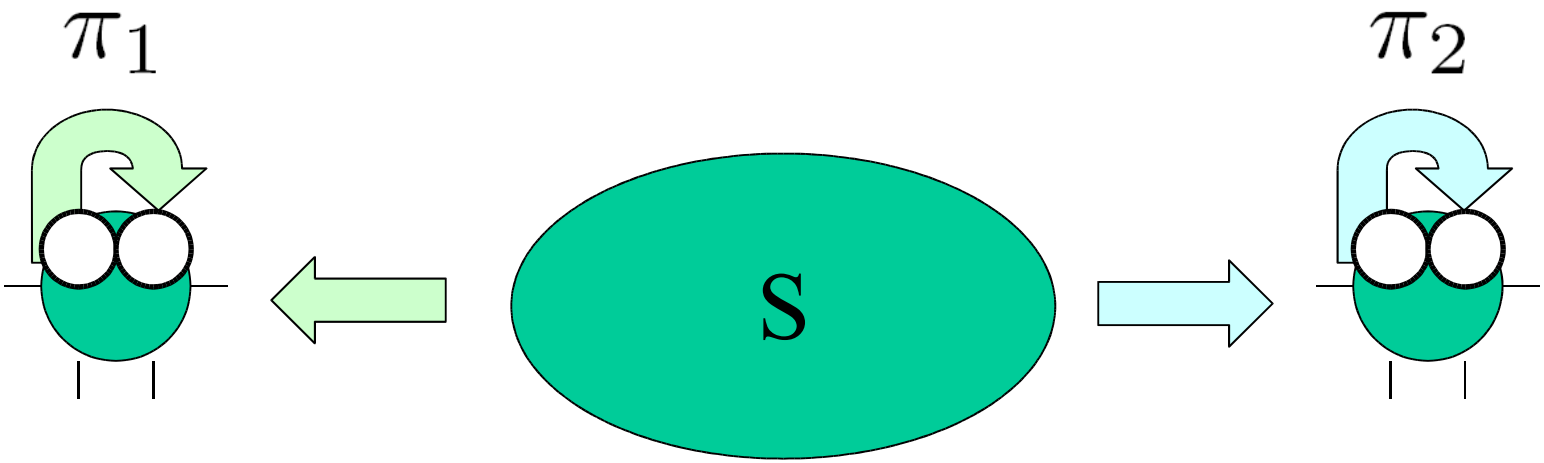
Fonctionnement (Initial)



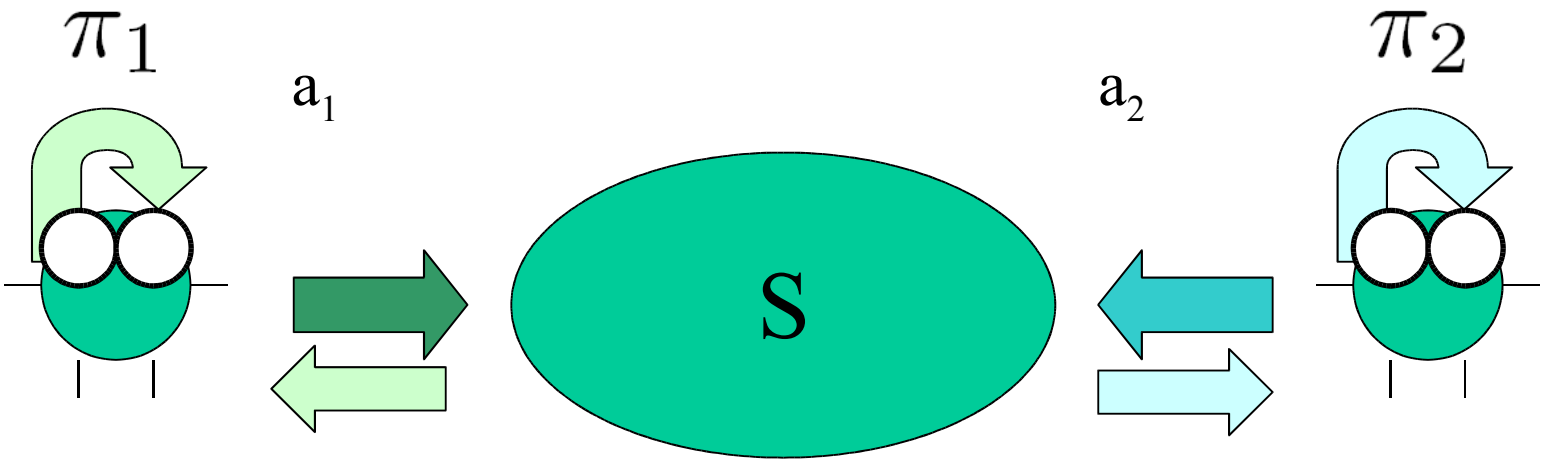
Fonctionnement (Observations)



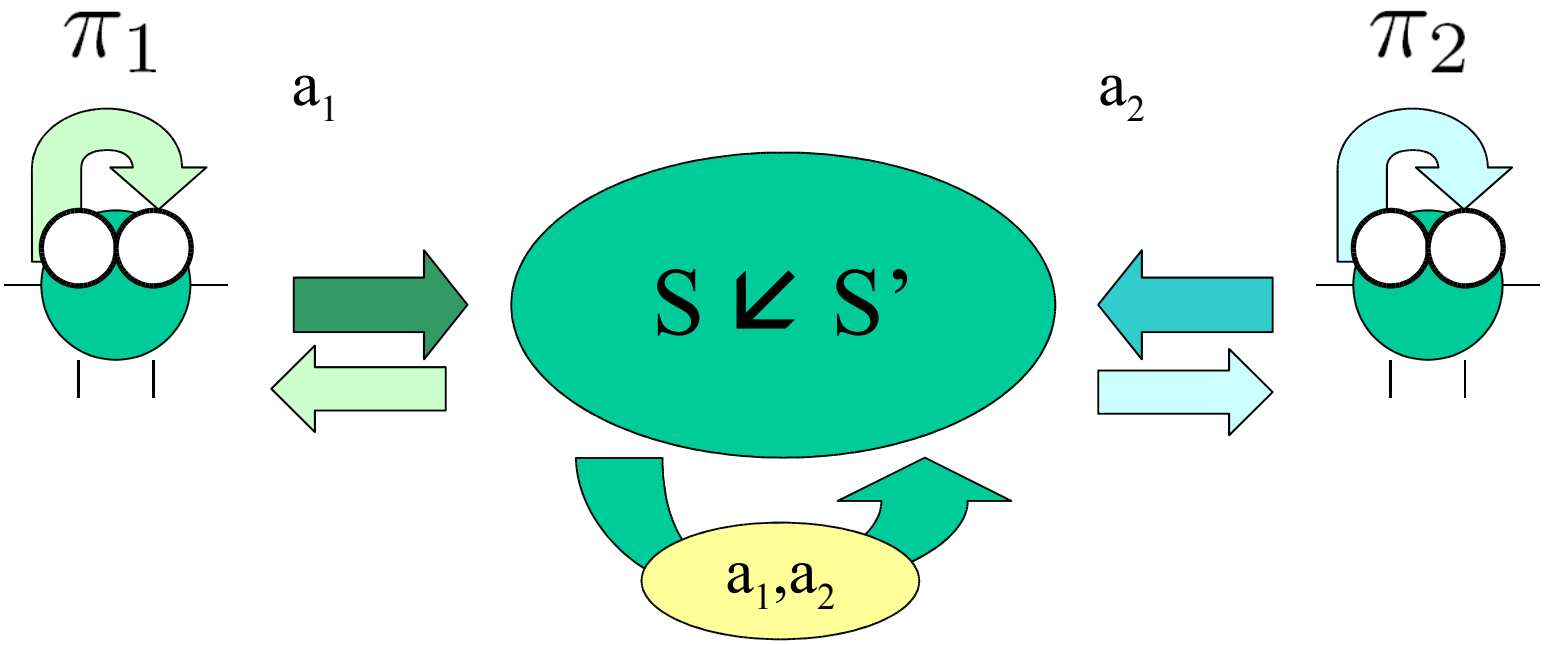
Fonctionnement (Décision)



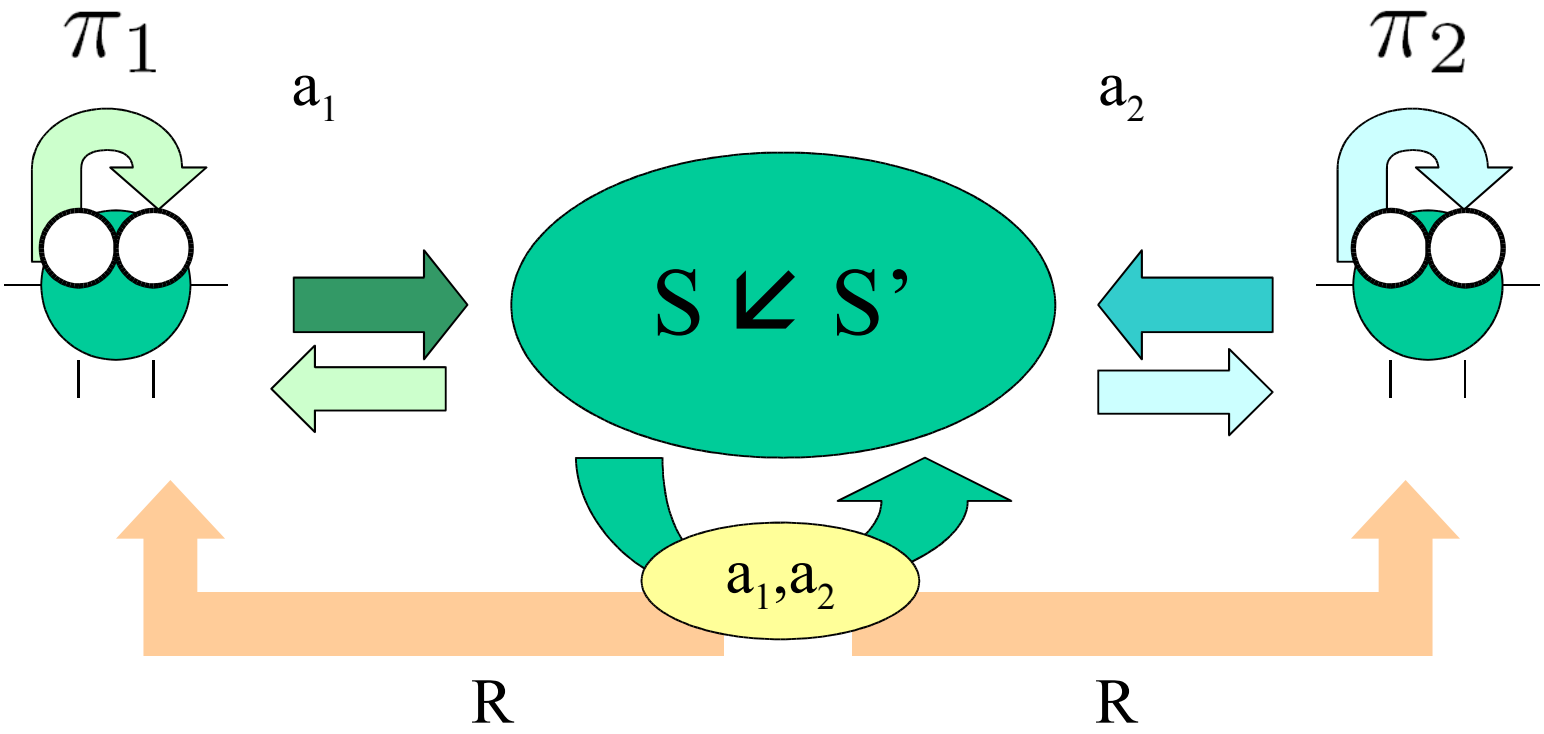
Fonctionnement (Action)



Fonctionnement (Évolution)

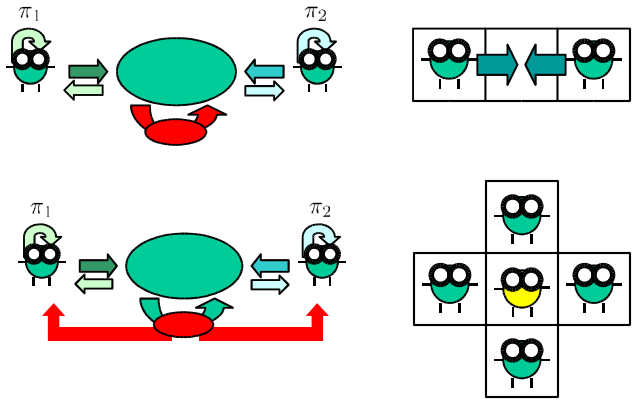


Fonctionnement (Récompenses)



Difficultés dans les DEC-MDP

- Difficultés
  - Couplages implicites
    - Dans transitions T
      - Résultat de action dépend des autres
    - Dans récompenses R
      - Récompense dépend des autres



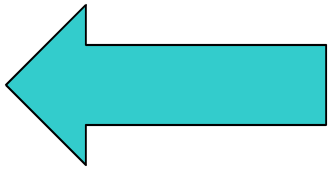
  - Évolution dépend des comportements des autres
- Résolution
  - Centralisée ↙ mono-agent
    - Explosion combinatoire
  - Décentralisée
    - Problème co-évolution
    - Tragédie des communs
    - Problème de « credit assignment »

Trouver un compromis

- Notre proposition

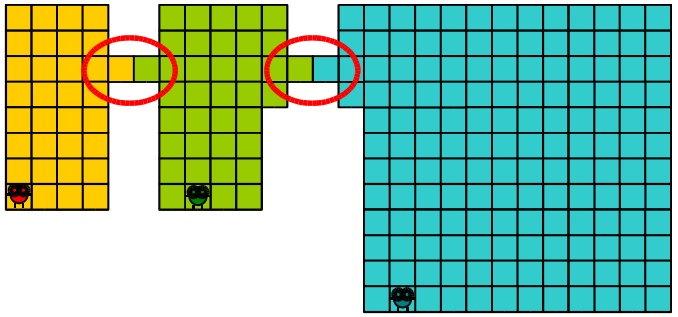
Plan

- Modèles markoviens
  - MDP
  - Extensions
- Notre proposition
  - Interac-DEC-MDP
  - Formalisme
- Exemples
- Résolution
- Conclusion

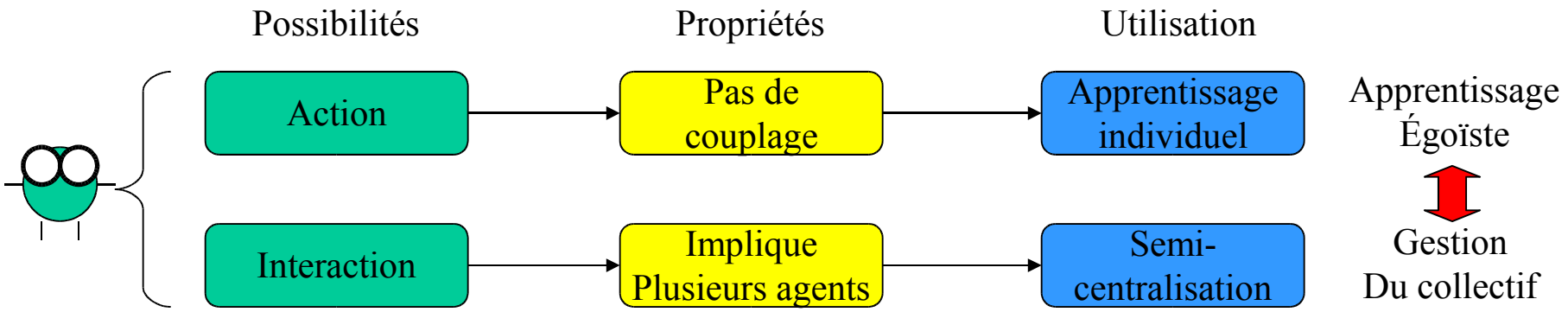


Proposition

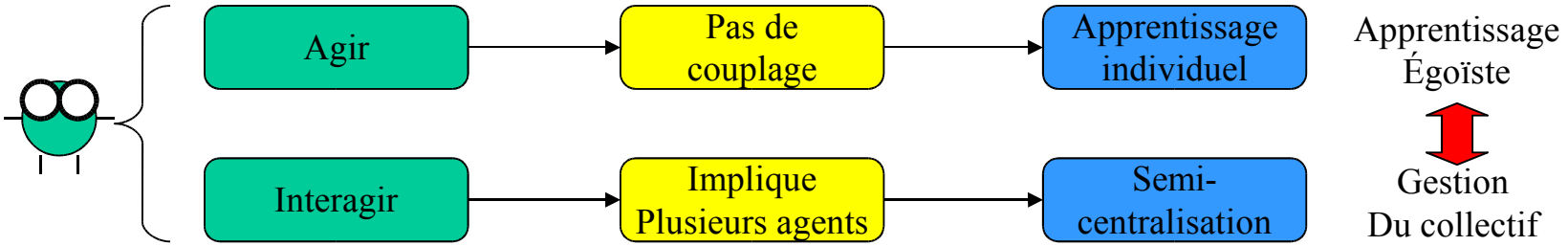
- Motivation :
  - Besoins de raisonner au niveau collectif sont limités
    - Échange, Partage de ressources, …
  - Raisonner individuel est moins coûteux
    - Gestion des ressources attribuées



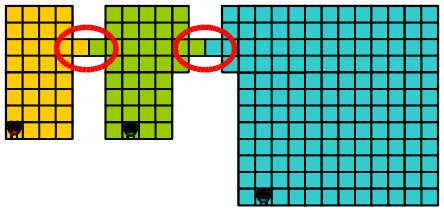
- Nouveau cadre formel
  - Interac-DEC-MDP
  - Restreindre les systèmes considérés
    - Séparer les décisions collectives des décisions individuelles
    - Moins expressif
- Restriction ↙ Système Factorisés



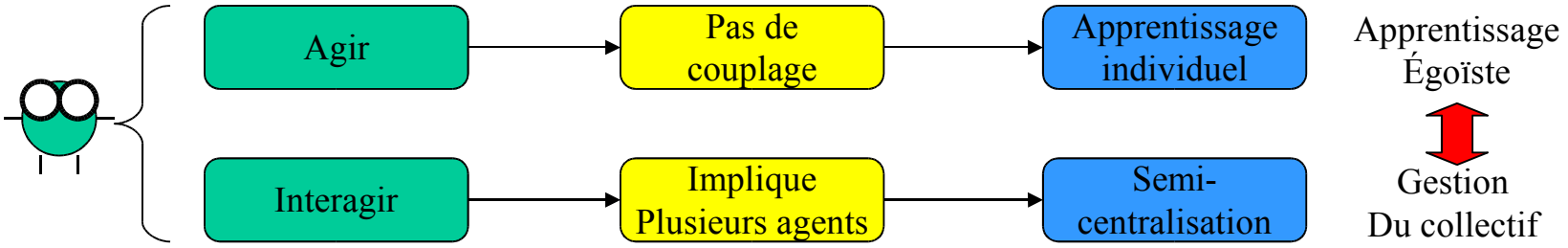
Cadre général



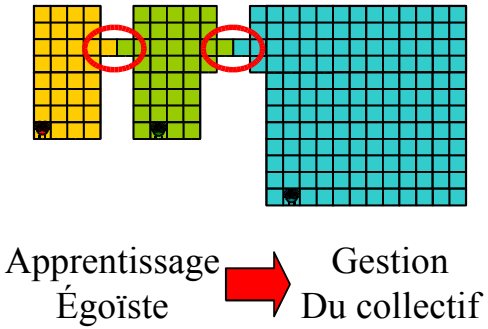
- Les agents peuvent agir individuellement
  - Pas influence des autres ↙ Transitions indépendantes
- Les actions des agents sont récompensées dans leur espace
  - Pas de couplage de R ↙ Récompenses indépendantes
- Chaque agent à des perceptions partielles
  - Etat, Récompenses, comportements des autres



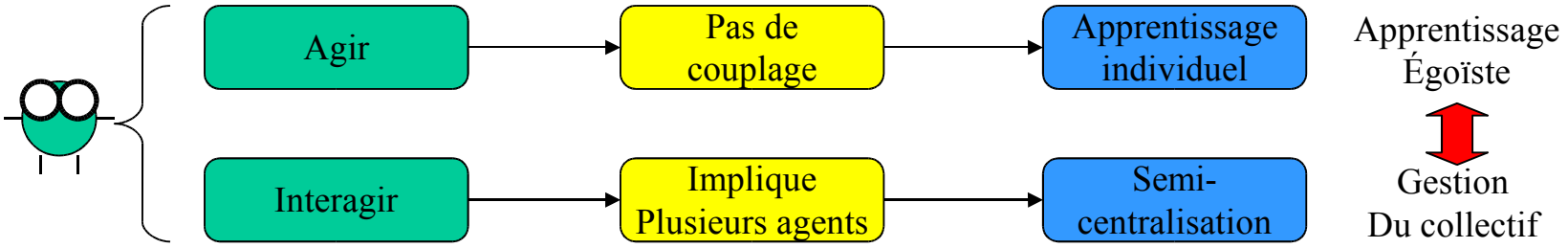
Cadre général



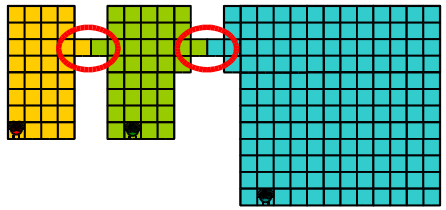
- Les agents peuvent agir individuellement
  - Pas influence des autres ↙ Transitions indépendantes
- Les actions des agents sont récompensées dans leur espace
  - Pas de couplage de R ↙ Récompenses indépendantes
- Chaque agent à des perceptions partielles
  - Etat, Récompenses, comportements des autres
- Interaction entre agents
  - Seuls couplages
  - Semi-centralisée entre agents impliqués



Cadre général



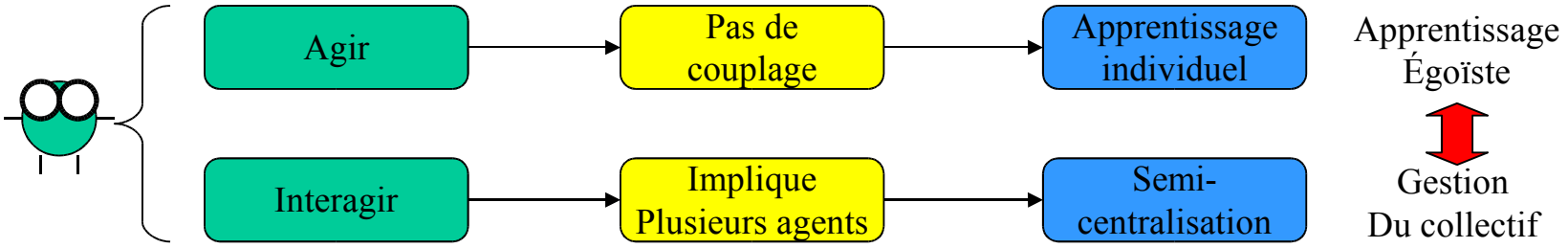
- Les agents peuvent agir individuellement
  - Pas influence des autres ↙ Transitions indépendantes
- Les actions des agents sont récompensées dans leur espace
  - Pas de couplage de R ↙ Récompenses indépendantes
- Chaque agent à des perceptions partielles
  - Etat, Récompenses, comportements des autres
- Interaction entre agents
  - Seuls couplages
  - Semi-centralisée entre agents impliqués
- Mais pas trivial
  - Remise en cause du comportement individuel



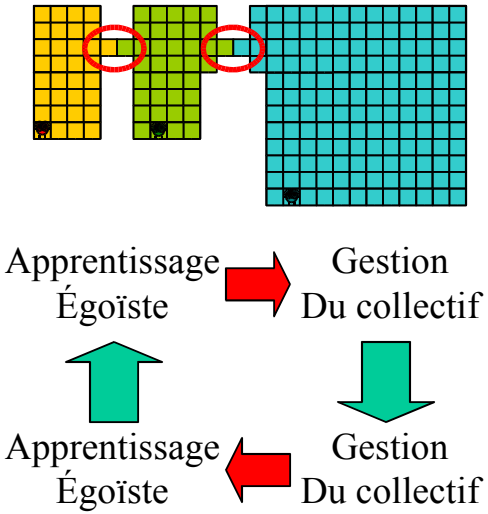
Apprentissage Égoïste → Gestion Du collectif

Apprentissage Égoïste ← Gestion Du collectif

Cadre général



- Les agents peuvent agir individuellement
  - Pas influence des autres ↙ Transitions indépendantes
- Les actions des agents sont récompensées dans leur espace
  - Pas de couplage de R ↙ Récompenses indépendantes
- Chaque agent à des perceptions partielles
  - Etat, Récompenses, comportements des autres
- Interaction entre agents
  - Seuls couplages
  - Semi-centralisée entre agents impliqués
- Mais pas trivial
  - Remise en cause du comportement individuel



Formalisme: Agents

- Chaque agent *i* est décrit par un MDP $<S_i,A_i,T_i,R_i>$
  - S_i espace état individuel
  - A_i espace action individuel
  - T_i transition individuelle
  - R_i récompense individuelle

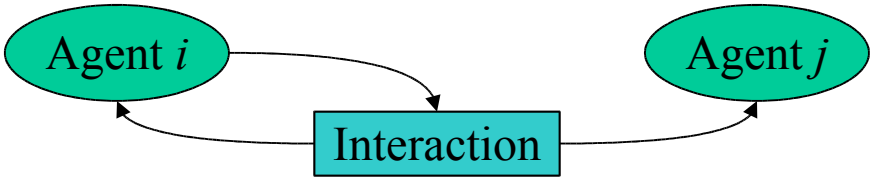
- Les agents agissent simultanément
  - Politique individuelle π_i

- L'objectif ↙maximiser la somme des récompenses individuelles
  - Pour le moment, sans interaction

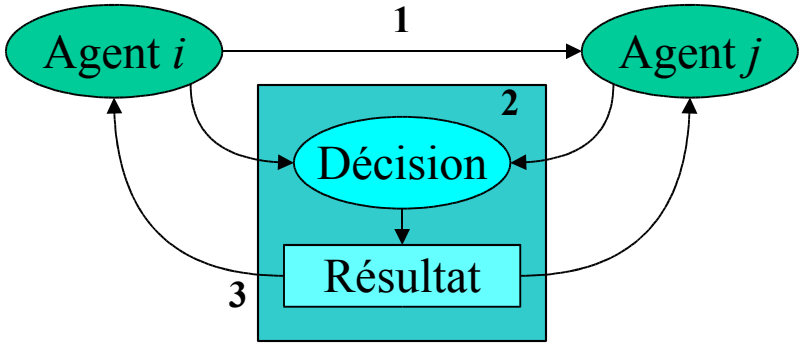
$$\max\left(\sum_i V(i)\right) = \sum_i \max\left(V(i)\right)$$

Interactions directes

- Définition
  - Influences mutuelles réciproques ponctuelles
- Il s'agit des seuls couplages du système
  - Agent i peut influencer état de j

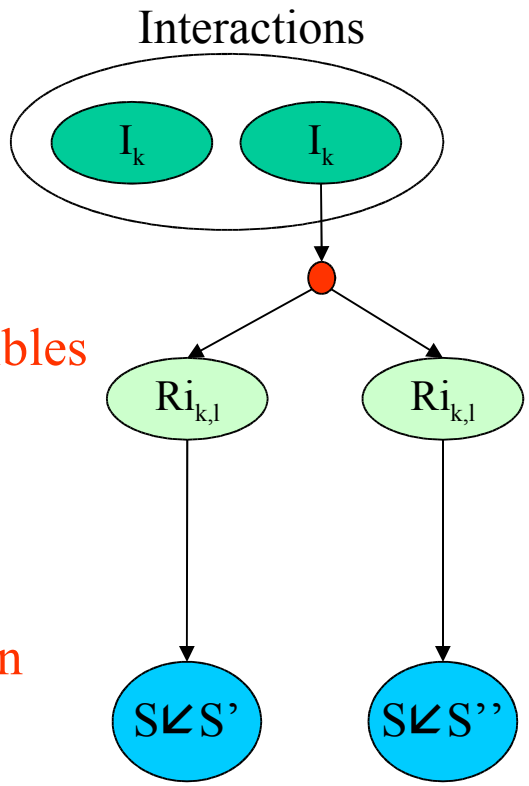


- Les agents impliqués peuvent raisonner
  - Politique dépend des agents impliqués
  - Processus de négociation

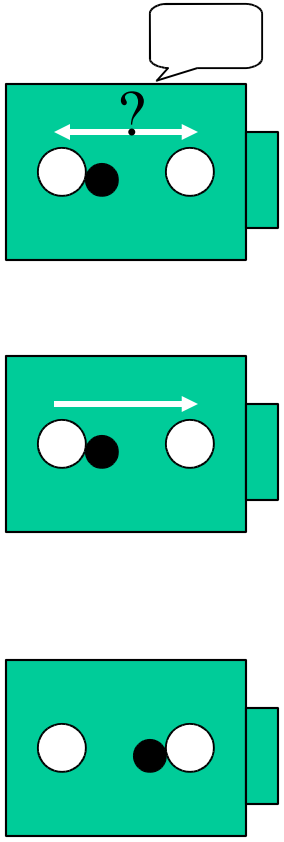


Représentation interactions

- Ajout d'instances d'interactions
  - I_k: interaction k
  - I=ensemble des interactions
- Interaction: différents résultats possibles
  - $Ri_{k,l}$: résultat l
  - Ri_k: ensemble des résultats de I_k
- Chaque résultat: matrice de transition
  - $TRi_{k,l}$



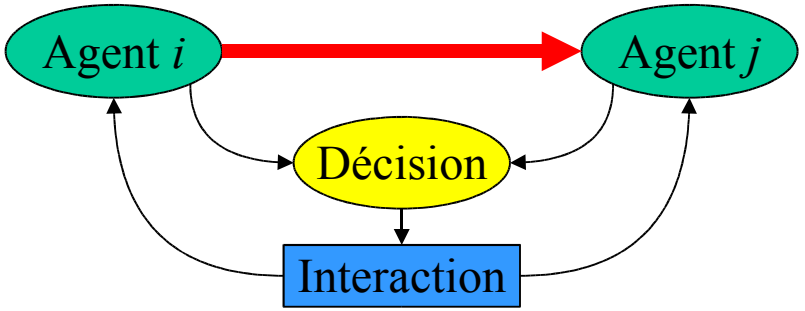
Sport collectif



Politiques d'interaction

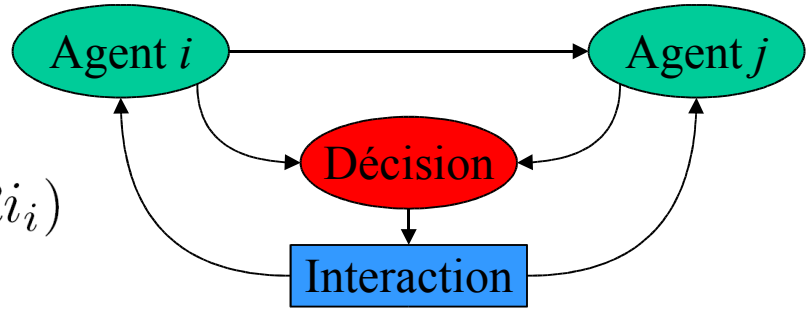
- Individuelle
  - Déclenchement

$$\pi_{i,trig} : S \rightarrow P(I, [0..n])$$



- Collective
  - Semi-centralisation
  - Résolution d'interaction
    - Pour chaque couple

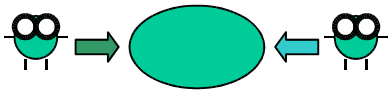
$$\Pi_{interac,i,n_i,i_i} : S \rightarrow P(Ri_i)$$



Formalisme: Modèle d'exécution

- Module d'action
  - Décision

$$\forall i, a_i \leftarrow \pi_i(s)$$



  - Exécution

$$\forall i, r_{i,earned} \leftarrow r_i(s, \{a_k\}_k)$$
$$s \leftarrow T(s, \{a_k\}_k)$$

- Module interaction
  - Pour tout agent i
    - Déclenchement

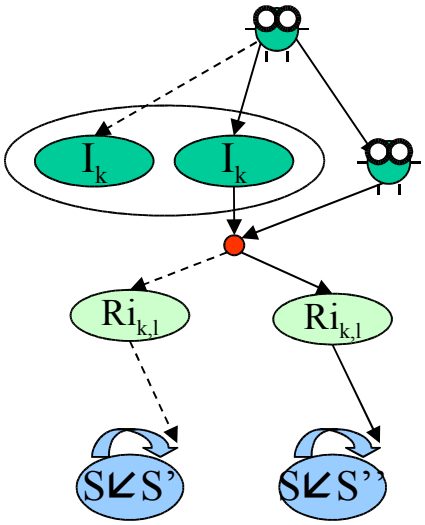
$$(i_i, n_i) \leftarrow \pi_{i,trig}(s)$$

    - Décision jointe

$$ri_i \leftarrow \Pi_{interac,i,n_i,i_i}(s)$$

    - Exécution de l'interaction

$$s \leftarrow T_{ri_i,i,n_i}(s)$$



Nouveau problème

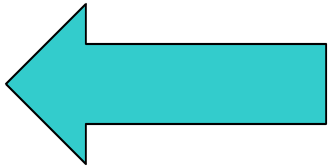
- Les agents peuvent
  - Agir
  - Interagir

- Objectif : déterminer
  - Politique d'action $\{\pi_i\}_i$
  - Politique de déclenchement $\{\pi_{trig,i}\}_i$
  - Politique de résolution $\{\Pi_{interac,i,j,I_k}\}_{i,j,k}$

- De manière décentralisée
- Pour maximiser une récompense perçue partiellement par les agents

Plan

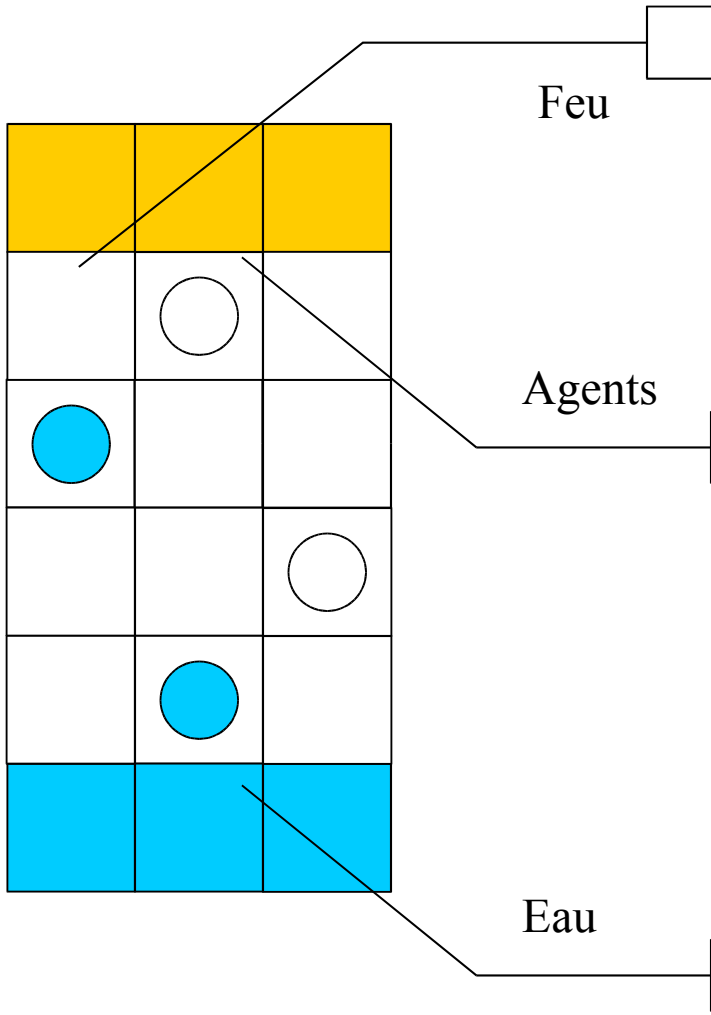
- Modèles markoviens
  - MDP
  - Extensions
- Notre proposition
  - Interac-DEC-MDP
  - Formalisme
- Exemples
- Résolution
- Conclusion



Exemples

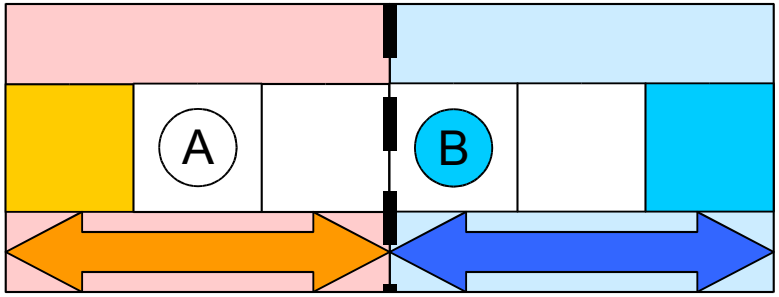
- Partage de nourriture
- Partage de ressources

- Pompiers
  - Chaque agent
    - Position
    - Possède seau plein/vide
  - Action individuelles
    - Les agents ne se gênent pas
    - T indépendants
  - Un agent reçoit une récompense
    - Met de l'eau dans le feu
    - R indépendant
  - Possibilité d'échanger des seaux
    - Interaction
    - Deux résultats: échange effectif / refusé
  - Intérêt de l'interaction
    - Plus vite dans les échanges

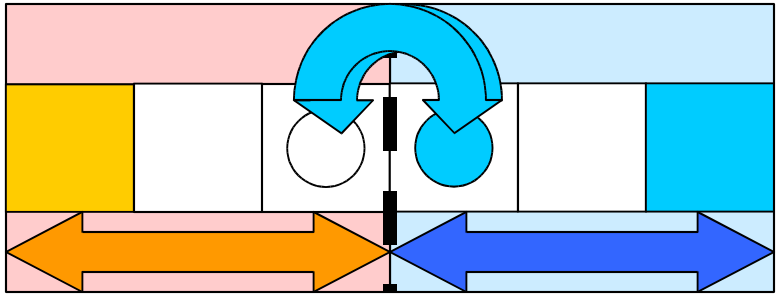


Exemple simple

- Deux agents
- Positions limitées



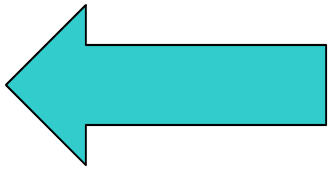
- Échanges possibles



- Conséquences
  - Agent A voit feu et récompense mais pas eau
  - Agent B voit eau mais pas le feu ni les récompenses

Plan

- Modèles markoviens
  - MDP
  - Extensions
- Notre proposition
  - Interac-DEC-MDP
  - Formalisme
- Exemples
- Résolution
- Conclusion



Résolution

- En cours

- Deux objectifs
  - Apprentissage individuel ↙ Collectif
  - Apprentissage collectif ↙ Individuel

- Représentation décentralisée des politiques
  - Apprentissage individuel ↙ Collectif
  - Utilise les apprentissages individuels
    - Maximiser somme des récompenses escomptées
    - Représentation décentralisée des résolutions d'interactions

Utilisation des $Q_{interac}$

- Chaque agent dispose de

$$Q_{interac,i} : S \times RI_{k,l} \times \{A, P\} \rightarrow \Re$$

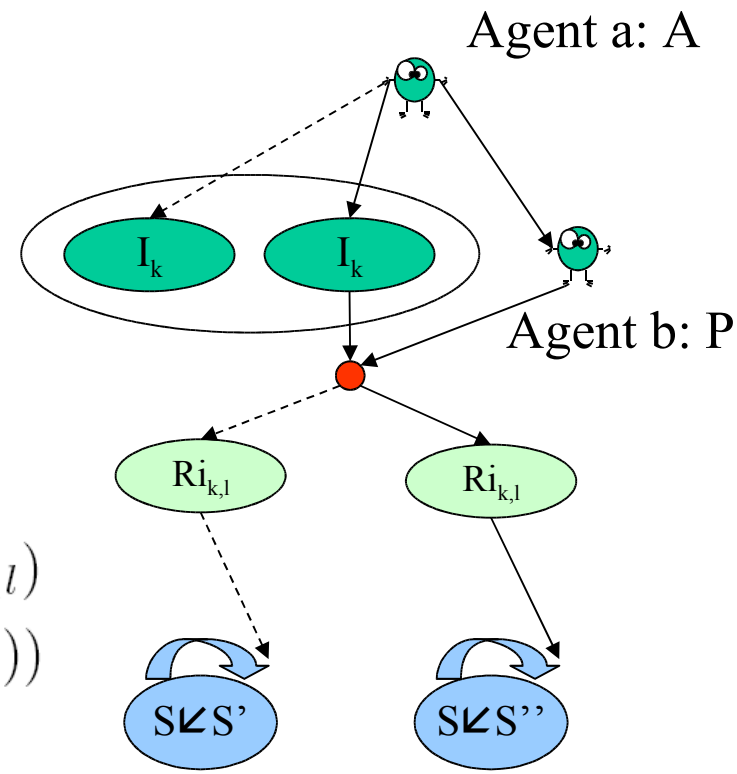
- Description
  - S : État du système
  - $RI_{k,l}$: Résultat d'interaction
  - {A,P} : Agent Actif ou Passif

- Interaction

Introduction du collectif

$$Q_r(RI_{k,l}) = Q_{I_k,a,A}(RI_{k,l}) + Q_{I_k,b,P}(RI_{k,l})$$

$$\Pi_{interac,a,b,I_k}(s) = argmax_{RI_{k,l}}(Q_r(RI_{k,l}))$$

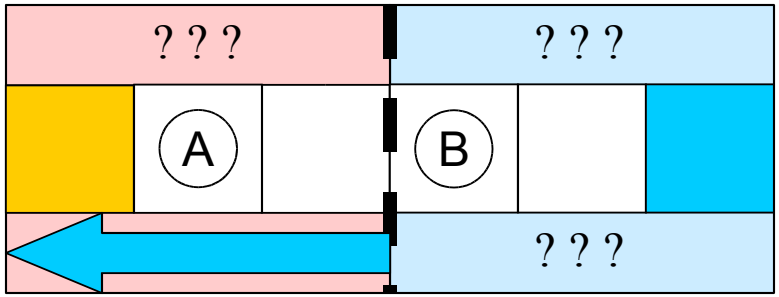


Approche naïve

- 3 apprentissages dépendants
  1. Apprentissage actions individuelles

$$Q(s, a_i) \leftarrow (1 - \alpha).Q(s, a_i) + \alpha.(r + max_{a'}(Q(s', a'))$$

  – Q-learning individuel



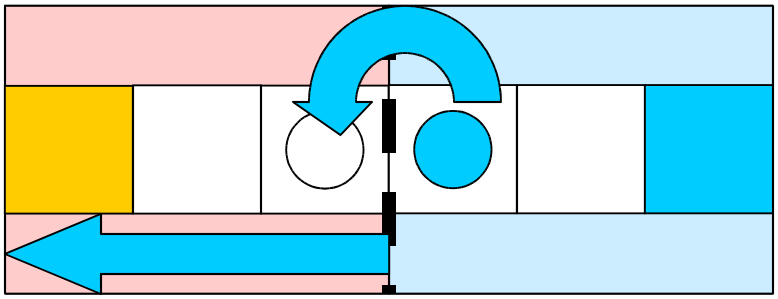
Approche naïve

- 3 apprentissages dépendants
  1. Apprentissage actions individuelles

$$Q(s, a_i) \leftarrow (1 - \alpha).Q(s, a_i) + \alpha.(r + max_{a'}(Q(s', a'))$$

  4. Apprentissage des interactions

$$Q_I(s, RI_i) \leftarrow (1 - \alpha).Q_I(s, RI_i) + \alpha.(max_{a'}(Q(s', a'))$$



Approche naïve

- 3 apprentissages dépendants
  1. Apprentissage actions individuelles

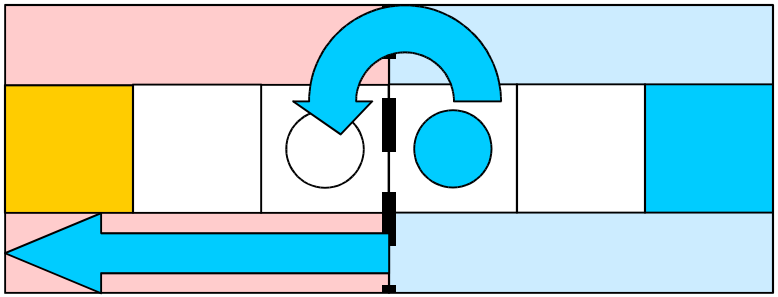
$$Q(s, a_i) \leftarrow (1 - \alpha).Q(s, a_i) + \alpha.(r + max_{a'}(Q(s', a'))$$

  4. Apprentissage des interactions

$$Q_I(s, RI_i) \leftarrow (1 - \alpha).Q_I(s, RI_i) + \alpha.(max_{a'}(Q(s', a'))$$

  7. Apprentissage des déclenchements

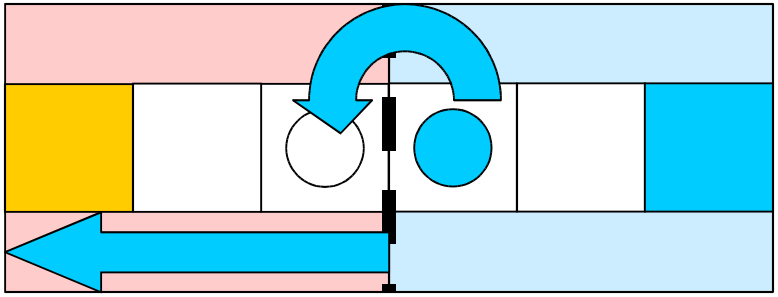
$$Q_{trig}(s, I) \leftarrow (1 - \alpha).Q_{trig}(s, I) + \alpha.(max_{a'}(Q(s', a'))$$



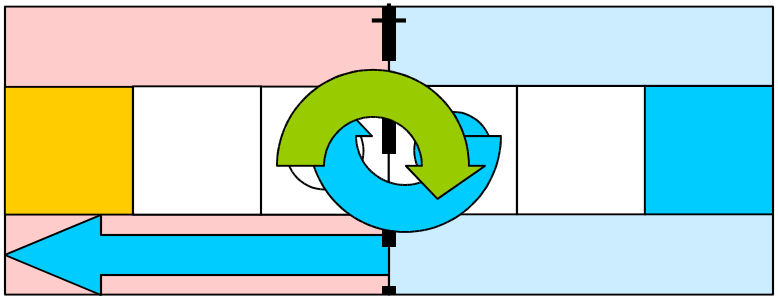
Apprentissage Égoïste → Gestion Du collectif

Problème à résoudre

- Il reste à remettre à jour comportement individuel



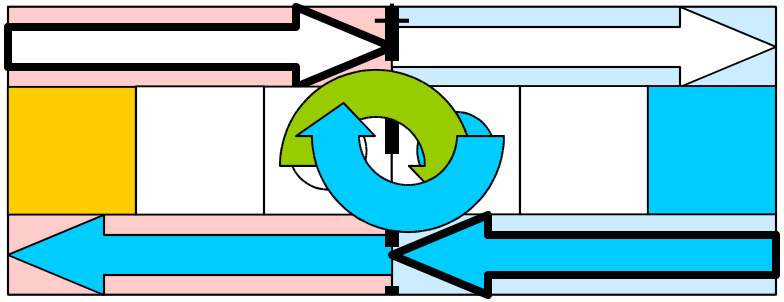
- B n'a rien appris
  - Solution : transfert de récompense



Apprentissage Égoïste ← Gestion Du collectif

Essais

- Forcer la Q-valeur de l'autre agents

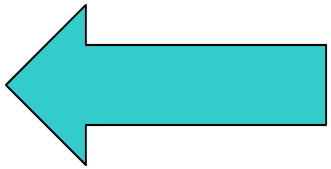


- Donne des résultats
  - Pour l'instant fait à la main
  - Apprentissages simultanés
  - Converge souvent

- Reste à analyser plus finement ce passage.
  - Références au MDP faiblement couplés

Plan

- Modèles markoviens
  - MDP
  - Extensions
- Notre proposition
  - Interac-DEC-MDP
  - Formalisme
- Exemples
- Résolution
- Conclusion



Conclusion

- Un nouveau modèle Interac-DEC-MDP
  - Actions
  - Interactions
  - Problème collectif perçu partiellement
- Séparer les décisions collectives / individuelles
  - Actions:
    - Conséquences locales
  - Interactions:
    - Conséquences plus globales
    - Décisions prises à plusieurs
- Définit une nouvelle entité
  - Ensemble d'agents
  - Transfert de récompense

Perspectives

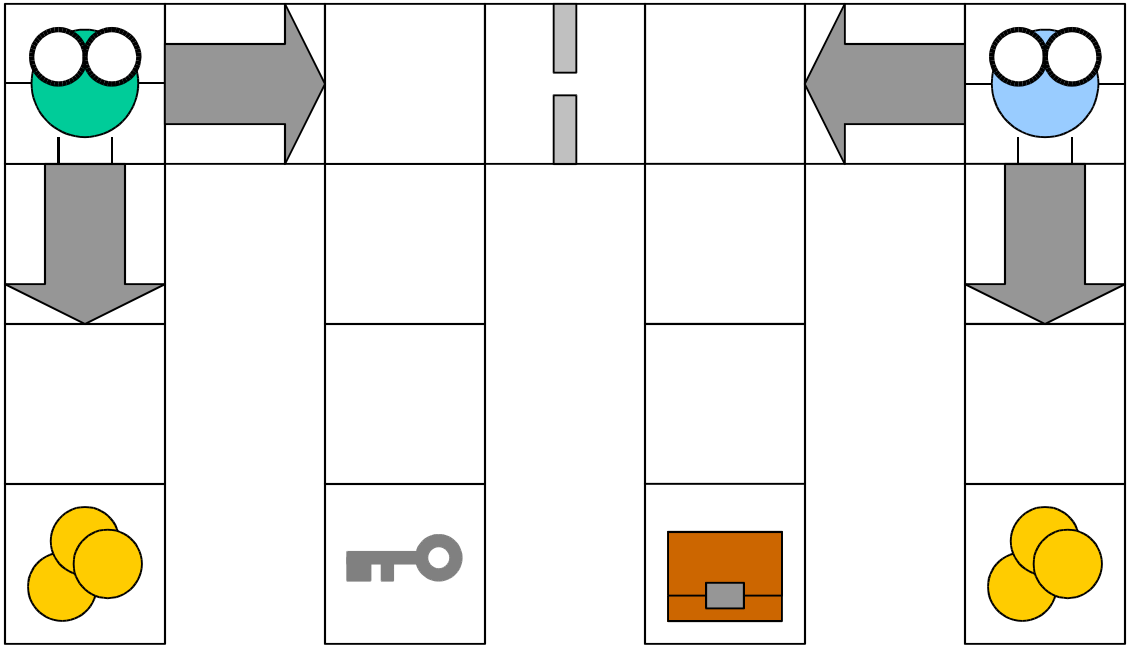
- Un exemple très simple
  - 2 agents
  - Perception globale
  - Mais algorithmique non triviale

- Première étape
  - Résoudre à deux agents

- Par la suite
  - Changer d'échelle (plus d'agents)
  - Perceptions partielles
  - DEC-MDP (couplages supplémentaires)

} Apprentissage Dans des systèmes Réels

Exemple



R_1	R_2	R_3	
5	5	10	Peu importe
8	1	10	Clef et coffre
8	3	10	Individuelles