Apprentissage évolutif de comportements éthiques

Rémy Chaput

LIRIS
Co-encadré par Salima Hassas (LIRIS) et Olivier Boissier (LaHC)

26 Juin 2019

Table des matières

Ethics.AI

Artificial constructivist agents that learn ETHICS
in humAn-Involved co-construction

La Région Auvergne-Rhône-Alpes

Financé par la région
Auvergne-Rhône-Alpes
(Pack Ambition Recherche, 2019-2023)

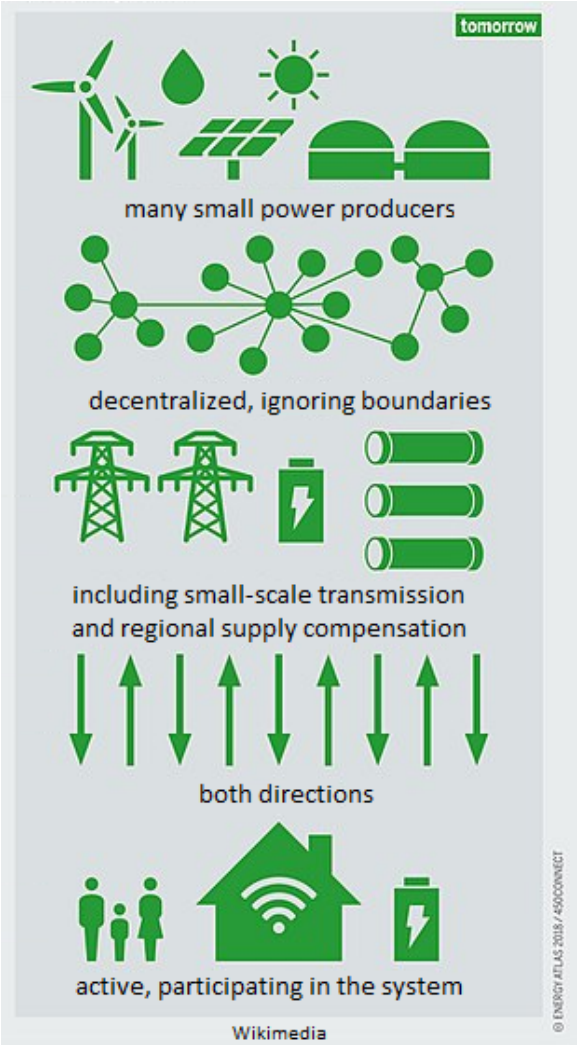
LIRIS — LABORATOIRE HUBERT CURIEN UMR • CNRS • 5516 • SAINT-ETIENNE — UCLy LYON CATHOLIC UNIVERSITY — ubiant

Motivations

- De plus en plus d'automatisation des décisions qui nous impactent
  - Conduite autonome
  - Transactions automatiques
  - Répartition des ressources dans un réseau
- Les approches existantes ne permettent pas l'évolution
  - Situation nouvelle

Cas d'étude

- Répartition de l'énergie dans les *Smart Grids* (Ubiant)
- Environnement sous contraintes
- Limitation de la consommation quand la grille est trop sollicitée
- Opposition de valeurs : confort personnel contre le bien-être de l'ensemble des participants



Problématique

Comment concevoir un **modèle d'agents artificiels intelligents**, capables d'apprendre des comportements que l'on considère comme **acceptables** par rapport à des **valeurs humaines**, et de **s'adapter** au fil de l'exécution, à partir de **mesures objectives et calculables** de leur environnement ?

Travaux existants

Éthique par conception

Choix éthiques faits par les concepteurs, intégrés en dur dans le système
Exemple : drone militaire [Dyndal et al. 2017]
Inconvénient : envisager toutes les situations possibles

Éthique par casuistique

Inférer des comportements à partir de jugements d'experts
Exemple : conduite autonome [Anderson et al. 2015]
Inconvénient : besoin de beaucoup de données annotées

Éthique par logique (dont architectures cognitives)

Formaliser des principes éthiques
Exemple : architecture Ethicaa : *trading* [Cointe 2017]
Inconvénient : monde limité à l'ensemble des règles formalisées

Approche proposée

Système Multi-Agent avec apprentissage par renforcement de comportements responsables

Environnement Multi-agent, agents hétérogènes, centrés sur l'humain

Constructivisme Basé sur les travaux de Piaget [Guerin 2008]
Apprentissage de la représentation / évolution

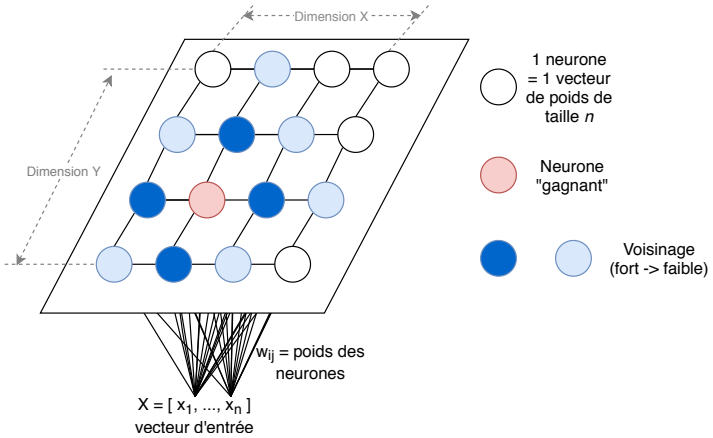
Apprentissage Basé sur une récompense (*Reinforcement Learning*)
Conséquences de décisions
Éventuellement multi-objectif

Monde continu Perceptions et actions $\in \mathbb{R}^n$
Richesse de l'environnement
Permet de dégrader les actions (compromis)

Techniques utilisées (1)

- *Q-Learning* [Watkins et al. 1992]
  - Apprentissage par essai/récompense
  - Recherche de la politique optimale π^* : *Etat* $\rightarrow$ *Action* qui maximise la récompense
  - Table des *Q-Valeurs* pour mémoriser l'intérêt d'une action dans un état (valeurs initialisées à 0, ensemble fini d'états et d'actions)

- Cartes Auto-Organisatrices [Kohonen 1998]
  - Réseau de neurones
  - Apprend la topologie des données en entrée
  - Pour simplifier, grille rectangulaire, distance euclidienne
  - Vecteurs initialisés aléatoirement

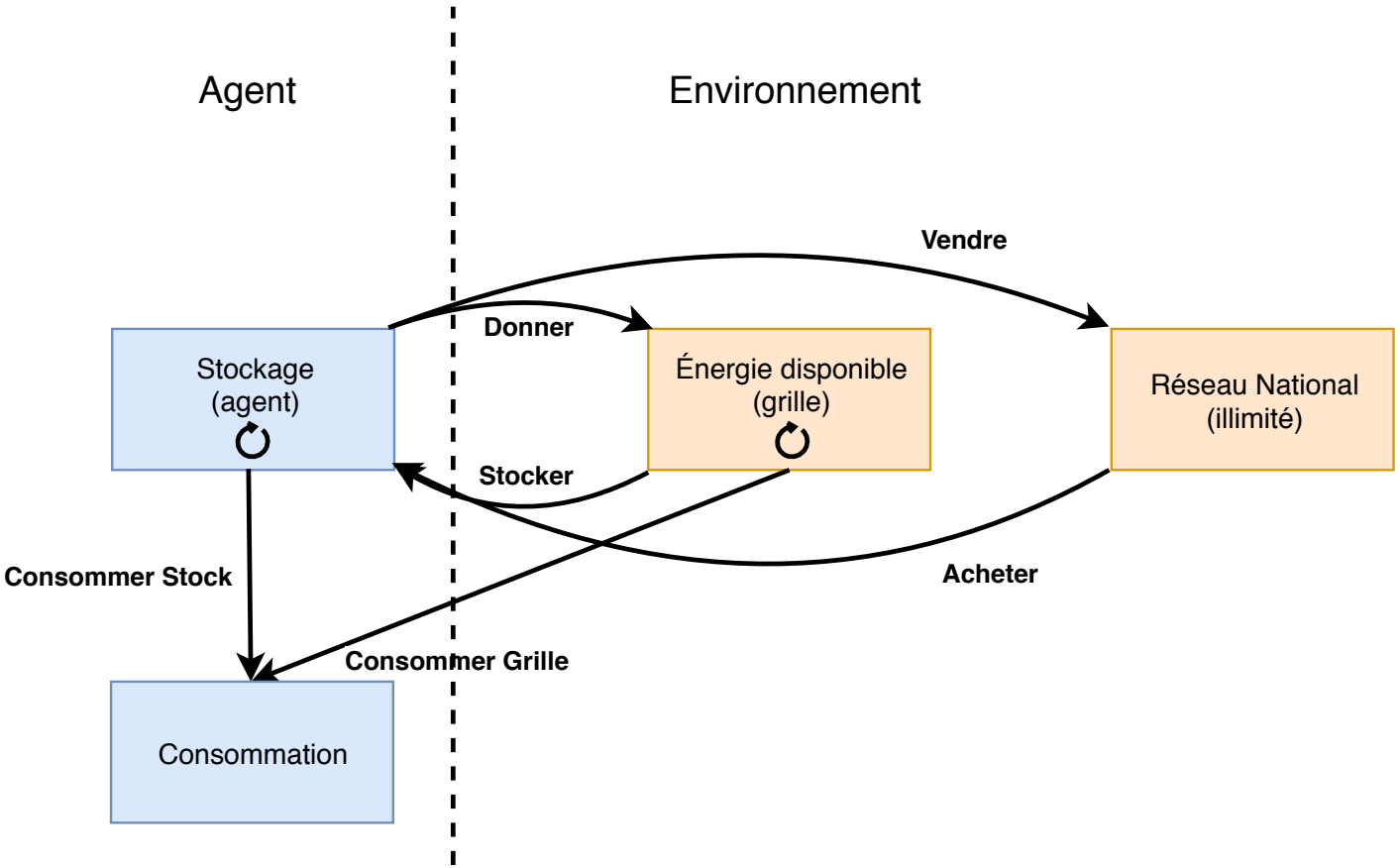


Techniques utilisées (2)

Choix de la récompense primordial pour spécifier les comportements voulus (*Credit Assignment Problem*)

- Contexte : multiples agents, état comme résultat des actions simultanées
- Problème : la récompense doit permettre à chaque agent d'apprendre
  - Récompense "globale" $\Longrightarrow$ récompense (resp. punition) même si le comportement est incorrect (resp. correct), si l'ensemble des agents compense
  - Récompense "locale" $\Longrightarrow$ maximiser pour soi donc comportements égoïstes
- Solution : *Difference Rewards* (DR) [Yliniemi et al. 2014]
  - On compare la valeur du monde actuel avec la valeur d'un monde hypothétique sans l'agent

Schéma du modèle



Production d'énergie dans la grille et dans chaque agent à chaque pas de temps

Simulateur multi-agent

- Temps discret
- Conception générique
- Env. stochastique, partiellement observable
- État suivant calculé comme résultat de l'influence des agents
- Chaque pas de temps :
  - Production d'énergie dans la grille et dans chaque agent
  - Les agents reçoivent des perceptions (capteurs)
  - Les agents choisissent les paramètres d'actions (répartition de l'énergie)
- Instance régulatrice pour empêcher les situations interdites par le concepteur

Environnement - Définitions

Nom	Description
$Consommation_{a,t}$	Qté. énergie utilisée par a à t
$Besoin_{a,t}$	Qté. énergie voulue par agent a à t
$Confort_{a,t}$	Satisfaction (selon $Consommation_{a,t}$ et $Besoin_{a,t}$)
$Effort_{a,t} = 1 - Confort_{a,t}$	Complémentaire du Confort
$EnergiePrise_{a,t}$	Qté. extraite de la grille (consommer + stocker)

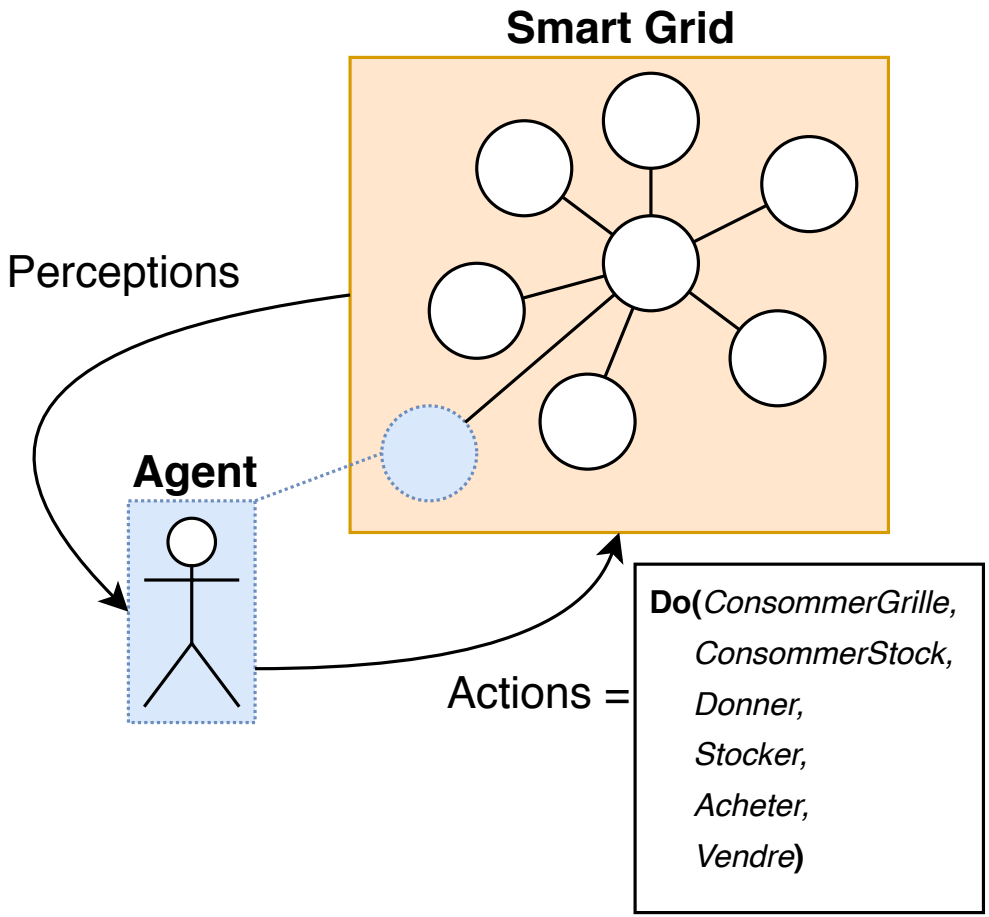
Environnement - Propriétés

Ensemble de mesures objectives sur l'environnement à un instant t

Nom	**Description**
Équité	Mesure statistique de dispersion (Gini) des conforts
Perte en énergie	Énergie disponible non utilisée (gaspillée)
Autonomie	Ratio vente + achat / énergie totale
Exclusion	Proportion d'agents dont effort $>$ 150% médiane
Bien-Être	Médiane du confort des agents
Surconsommation	Quantité d'énergie prise en excès

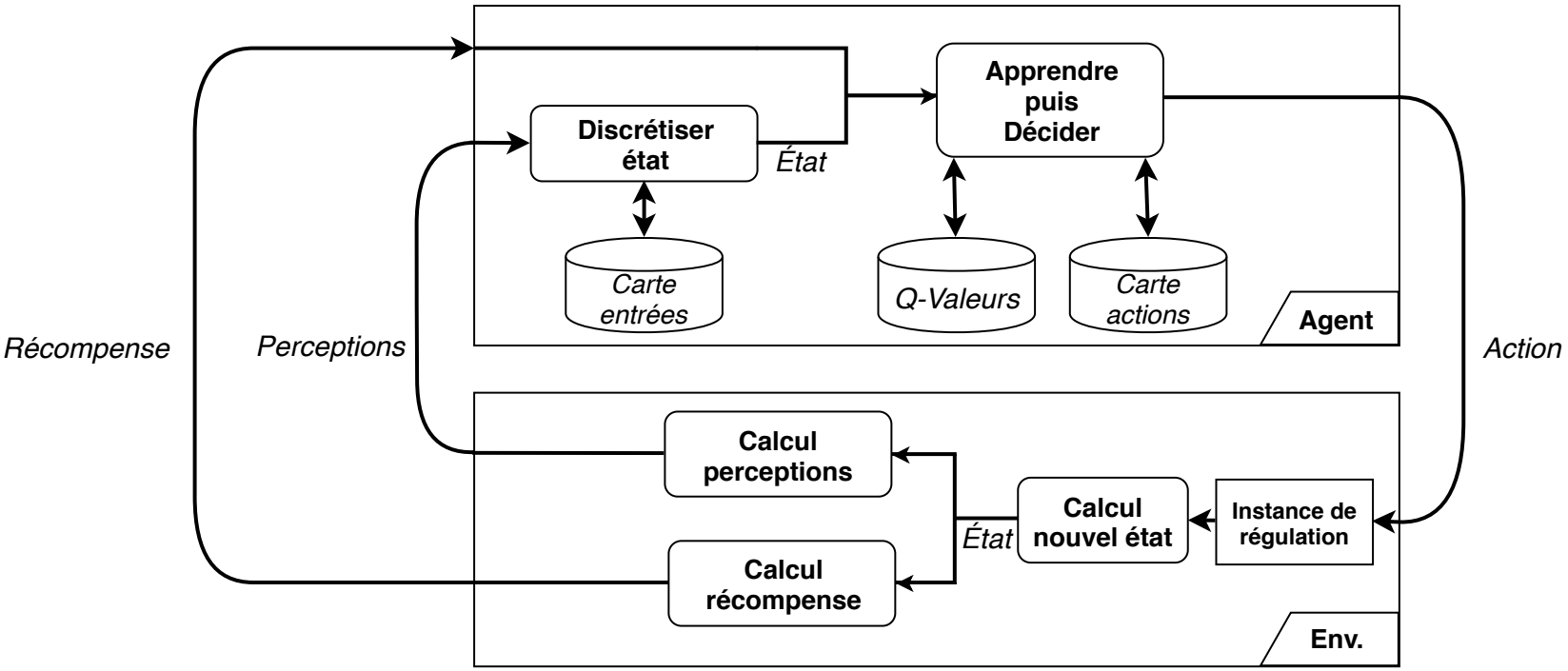
Agent

	Nom	Description
Env.	Heure	Heure de la journée
	Énergie disponible	Qté. disponible dans la grille
	Propriétés	Ensemble des propriétés de l'env.
Soi	Stockage	Batterie de l'agent
	Confort	Confort au pas de temps précédent
	Profit	Énergie vendue - énergie achetée



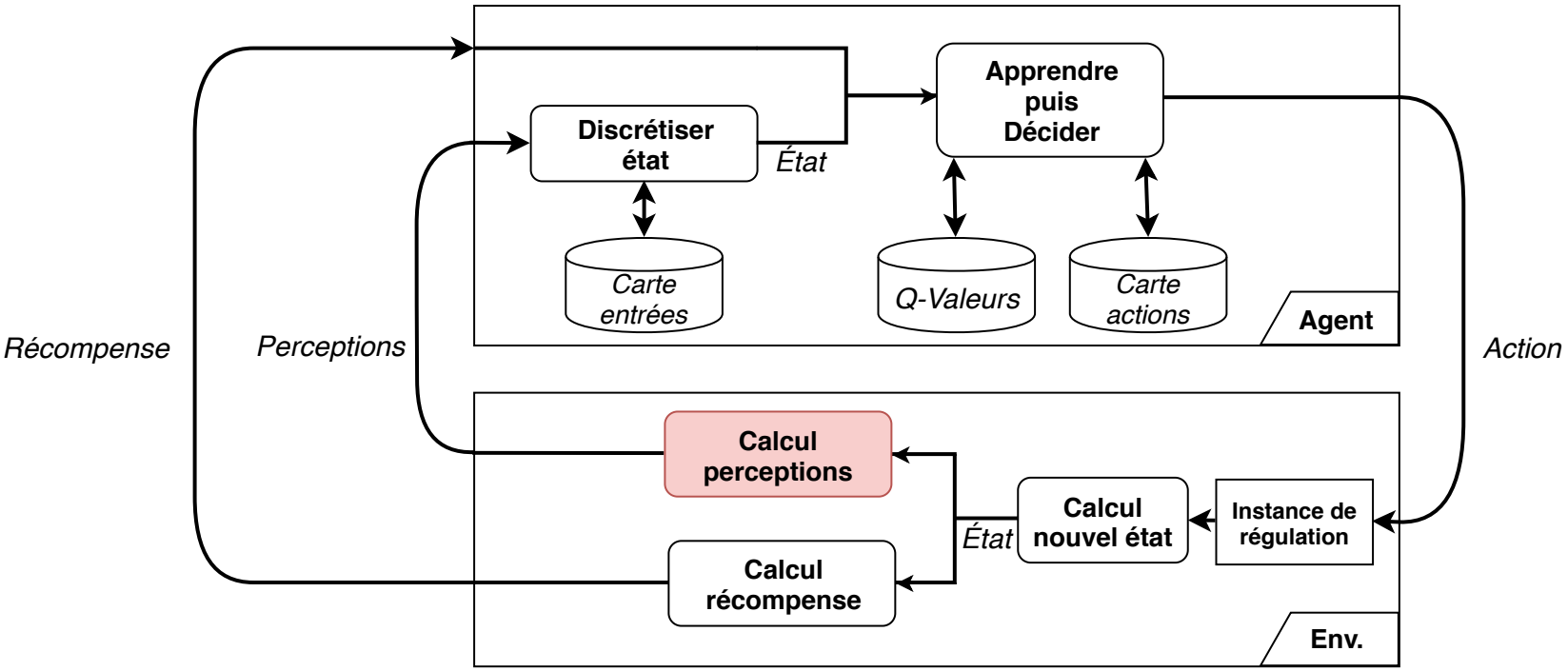
Plusieurs profils d'agent (Habitation, Hôpital) : change le besoin et la capacité d'action

Schéma du modèle d'apprentissage



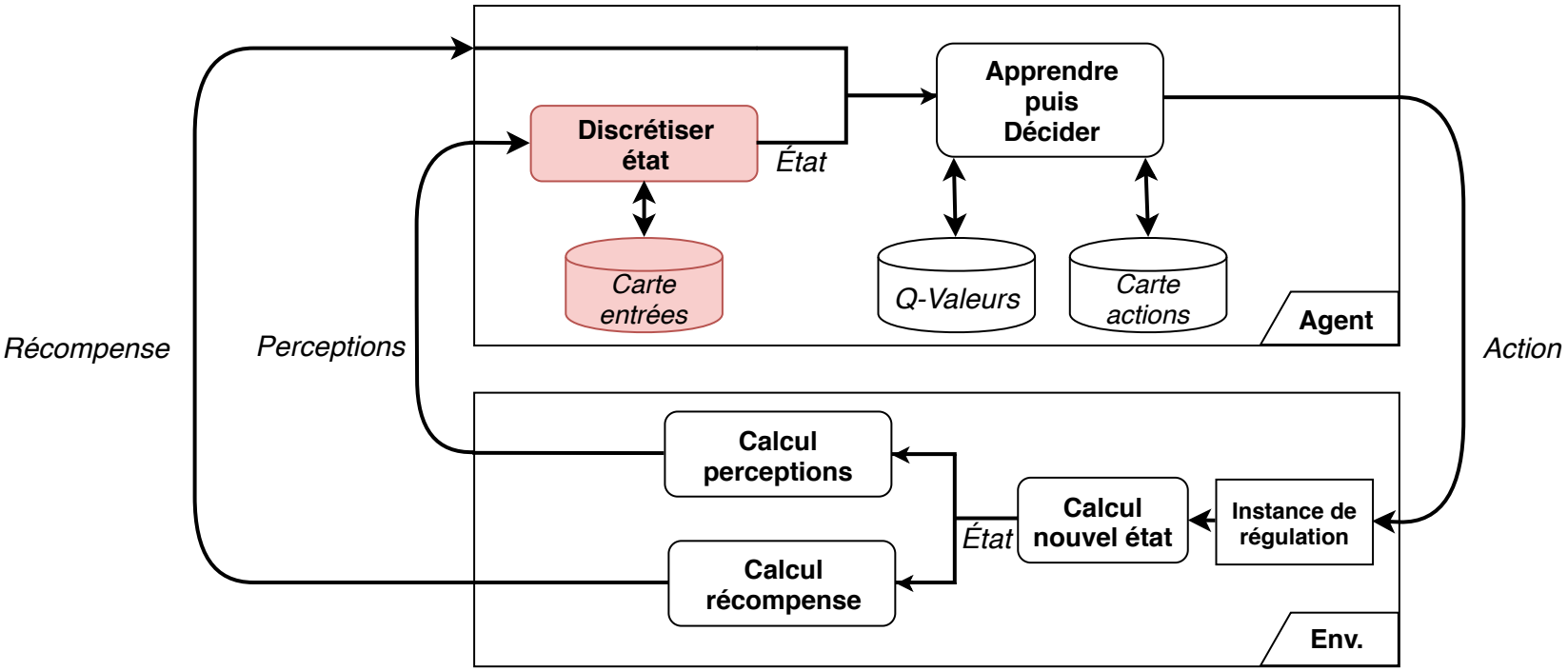
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage



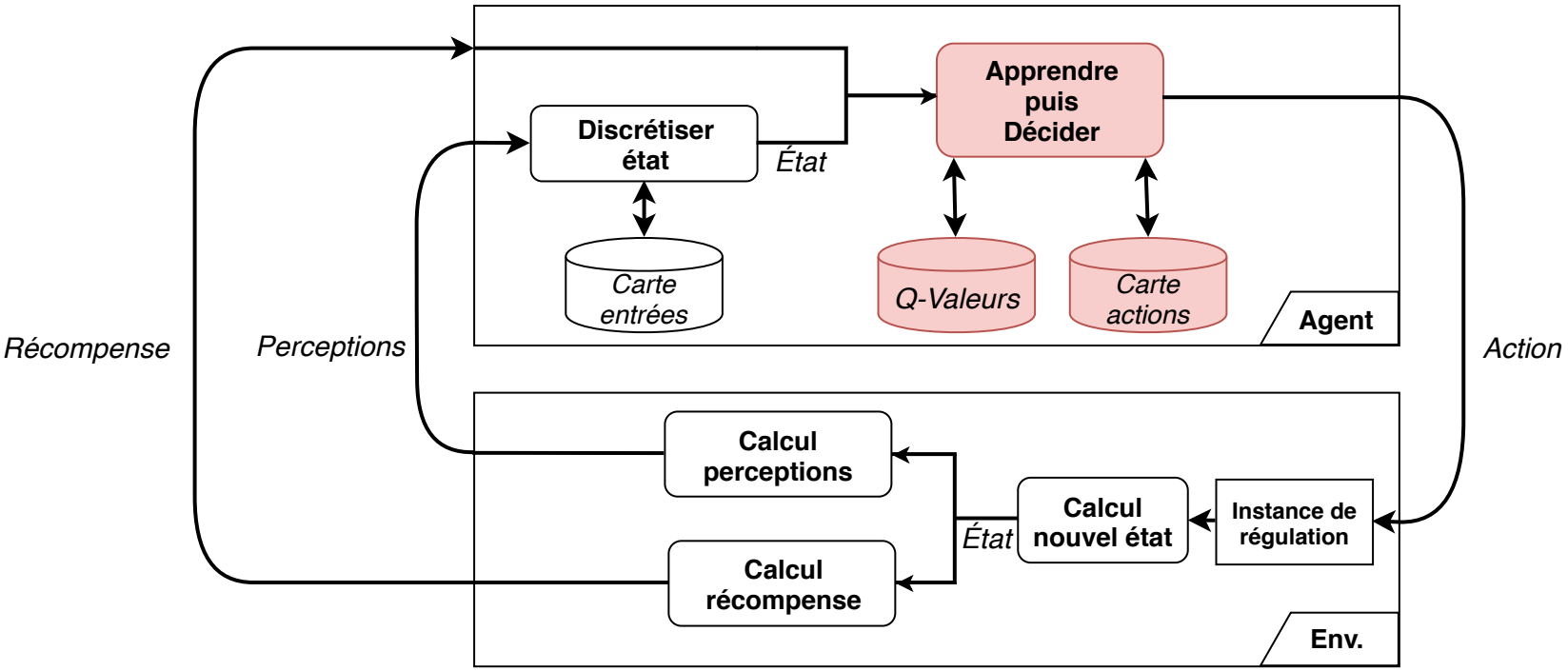
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage



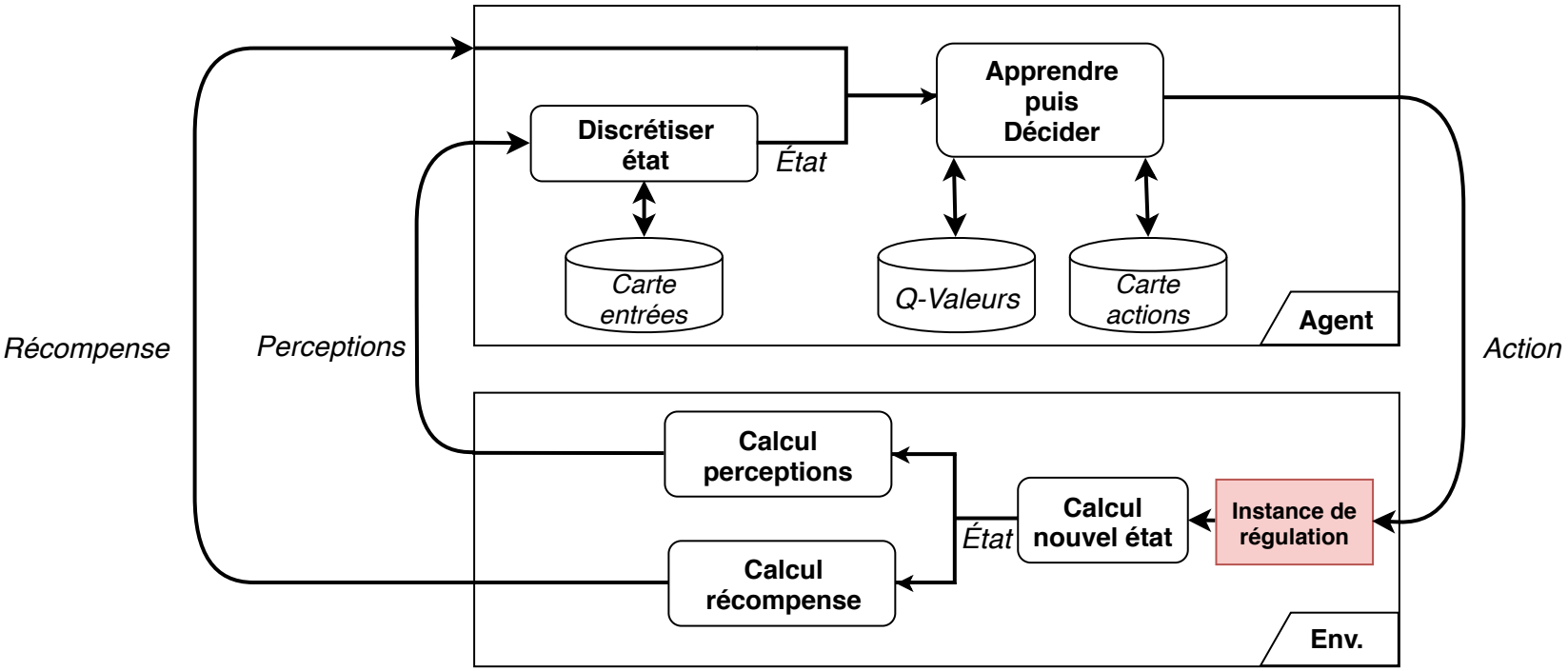
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage



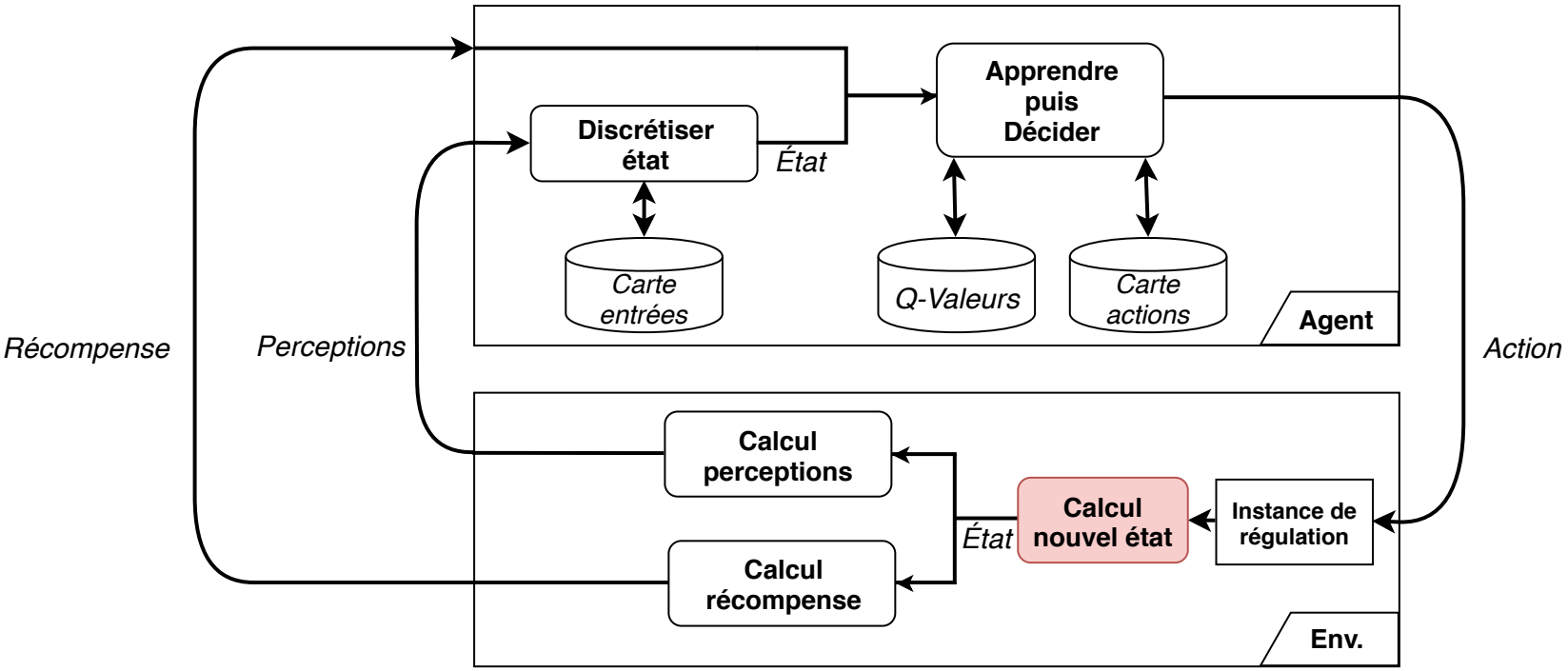
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage



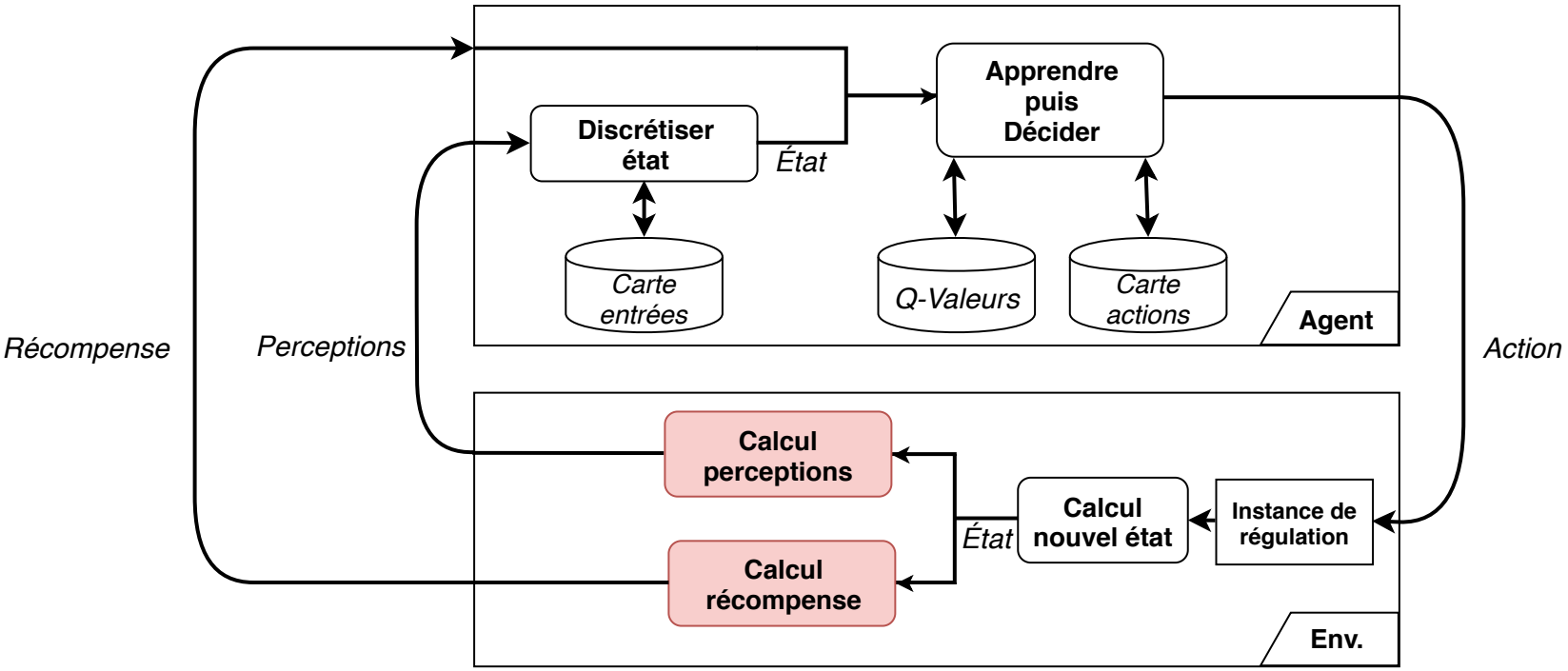
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage



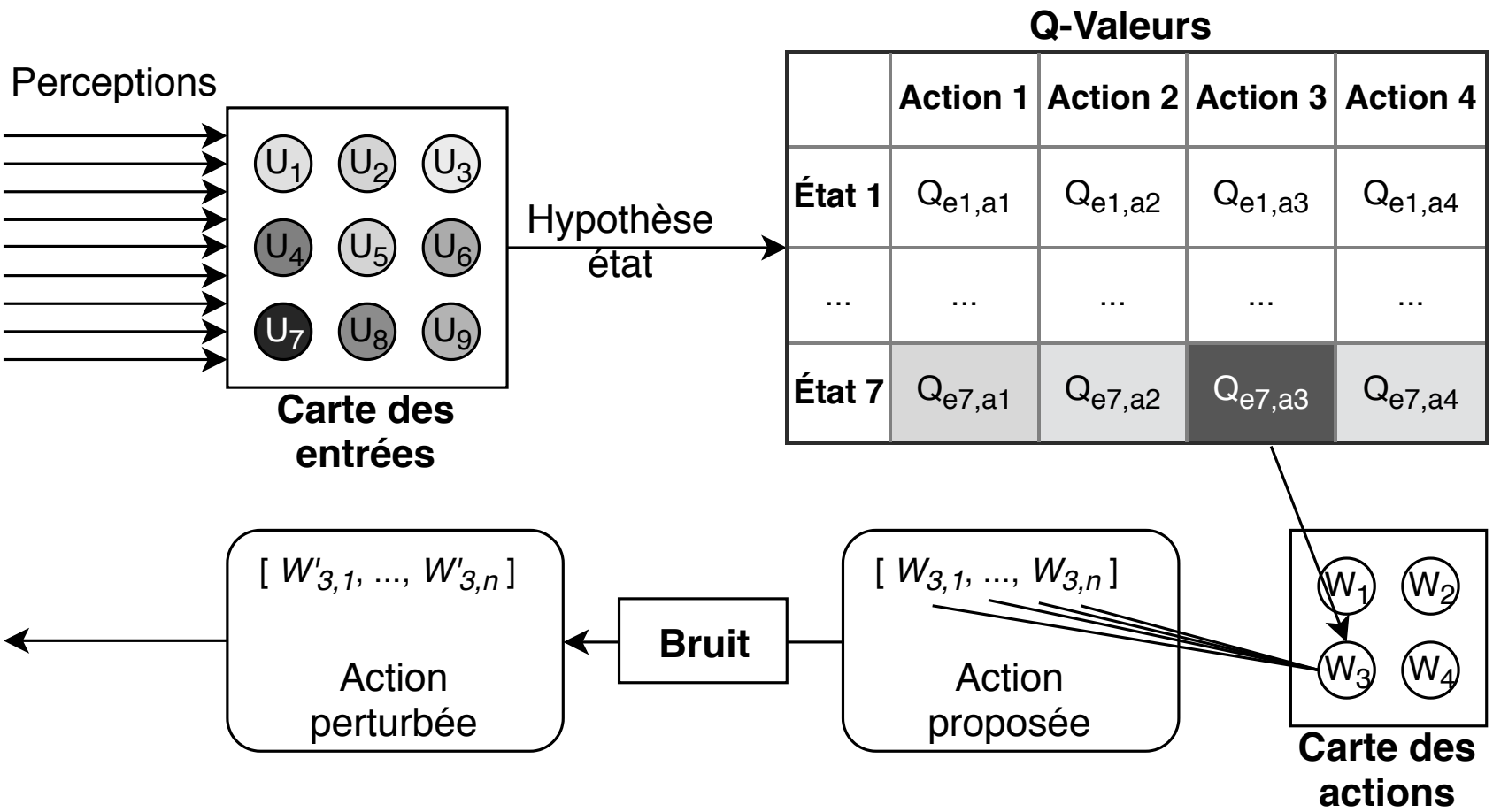
Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Schéma du modèle d'apprentissage

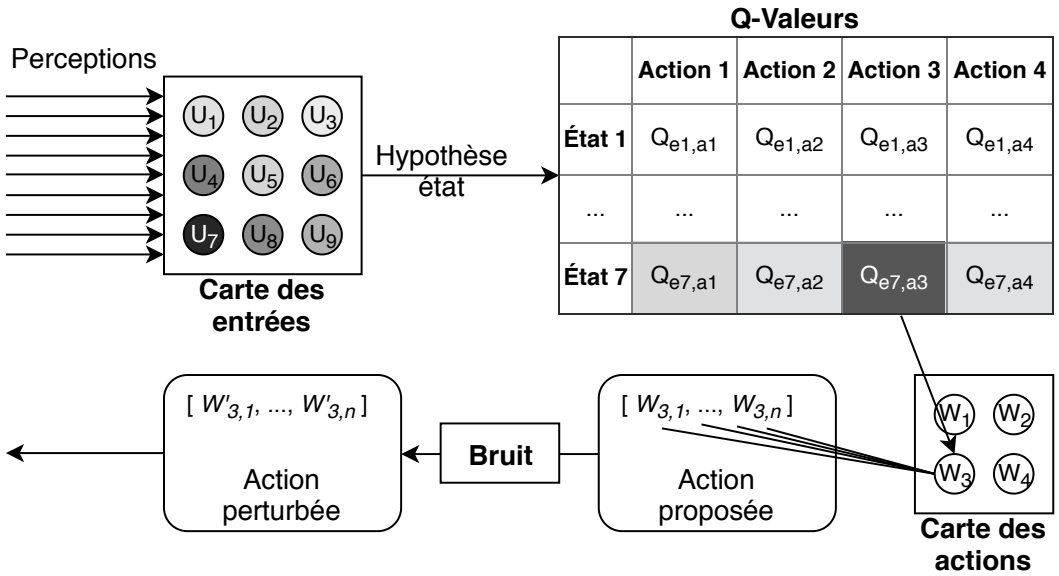


Q-Learning avec SOM pour l'aspect continu des perceptions et actions
Inspiré de [Smith 2002 ; Uang et al. 2012]

Processus de décision



Mise à jour des poids



- L'agent reçoit une récompense et un nouveau vecteur de perceptions
- Si l'action était intéressante,
  l'agent met à jour sa carte des actions
- L'agent met à jour *toutes* ses Q-Valeurs
  selon la carte des entrées et des actions
- L'agent met à jour sa carte des entrées

Fonctions de récompense

équité : Utilise le Gini

équité 2 : Utilise la moyenne

surconsommation : Minimiser la propriété de surconsommation

multi-objectif produit : surconsommation $\times$ confort

multi-objectif somme : surconsommation + confort

adaptabilité 1 :
$$\begin{cases} \text{équité} & \text{si étape} < 3000 \\ \text{multi-objectif somme} & \text{sinon} \end{cases}$$

adaptabilité 2 :
$$\begin{cases} \text{équité} & \text{si étape} < 2000 \\ \text{équité + surconsommation} & \text{sinon} \end{cases}$$

adaptabilité 3 :
$$\begin{cases} \text{équité} & \text{si étape} < 2000 \\ \text{équité + surconsommation} & \text{si étape} < 6000 \\ \text{équité + surconsommation + confort} & \text{sinon} \end{cases}$$

Scénarios de simulation

10 agents "Habitation" + 1 agent "Hôpital"

Aléatoire

Processus de décision remplacé par une fonction aléatoire uniforme

SoloLearner

Un seul agent apprenant (les autres sont aléatoires)
Hypothèse : perfs. légèrement meilleures que Aléatoire

Apprentissage

Tous les agents apprennent

InitZero

Les poids des vecteurs des cartes d'action sont initialisés à 0
Permet dans certains cas d'apprendre plus finement les paramètres d'actions

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	**SoloLearner**	**Apprentissage**	**InitZero**
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

Résultats

Score : Moyenne des récompenses "globales" (au niveau de l'env.)

	Aléatoire	SoloLearner	Apprentissage	InitZero
équité	0.76	0.78	0.99	0.62
équité 2	0.76	0.78	0.99	0.89
surconsommation	0.18	0.19	0.56	0.94
MO-produit	0.57	0.57	0.58	0.55
MO-somme	0.33	0.34	0.32	0.53
adaptabilité 1	0.46	0.47	0.52	0.44
adaptabilité 2	0.53	0.54	0.67	0.58
adaptabilité 3	0.59	0.60	0.72	0.64

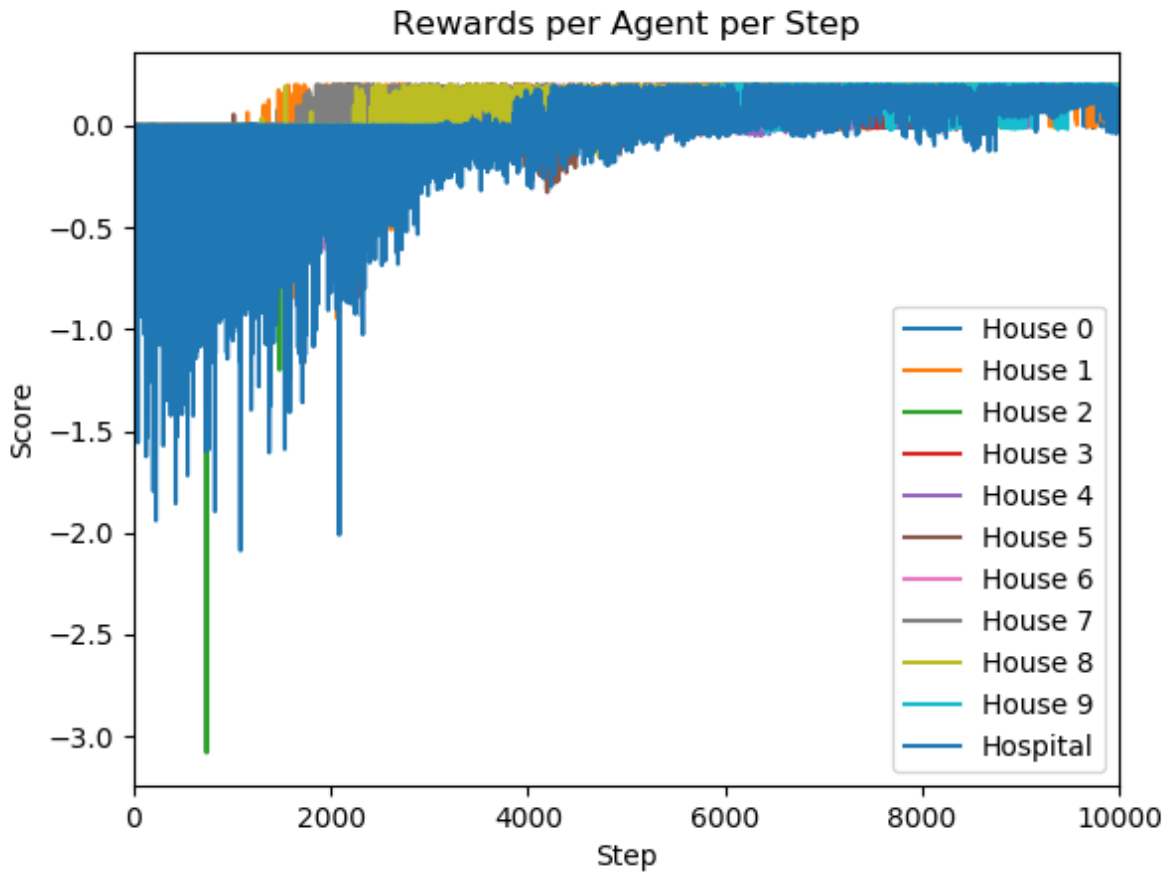
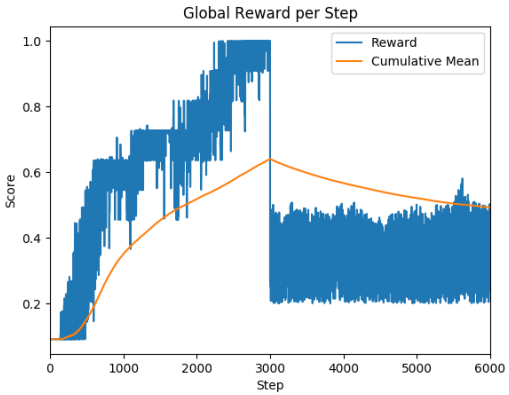
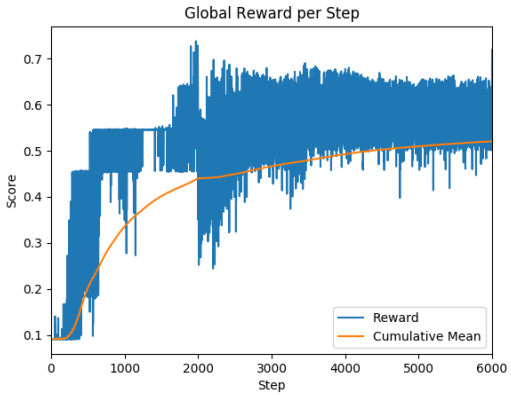


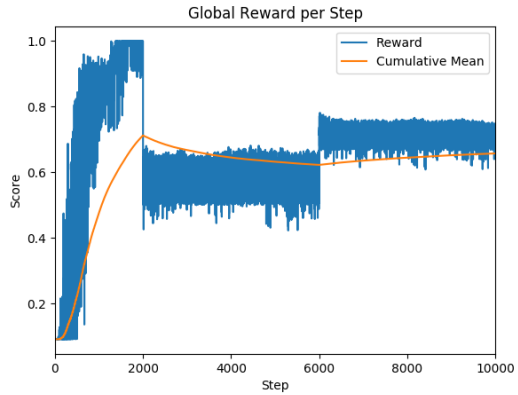
Figure 1 – Récompenses de chaque agent sur le temps - Scénario "InitZero" - Récompense "MultiObjectif Somme".



(a) Adaptabilité 1



(b) Adaptabilité 2



(c) Adaptabilité 3

Figure 2 – Récompenses pour les différentes récompenses "Adaptabilité" - Scénario "Apprentissage".

Rappel de la contribution

Apprentissage
- Utilisation d'une approche existante
- Expérimentations avec une méthode d'exploration différente
- Utilisation de récompenses multi-objectives et différenciées

Simulateur
- Création d'un simulateur multi-agents orienté *Smart Grids*
- Environnement stochastique, partiellement observable
- Temps discrétisé

Résultats
- Mise en évidence de l'apprentissage de comportements
- Ces comportements exhibent des valeurs humaines selon la fonction de récompense

Limitations et perspectives

Constructivisme Manque l'impact de la récompense sur la discrétisation

Exploration Perturbation des actions non contrôlée par la récompense donc exploration aléatoire

Récompense Apprentissage fortement dépendant de la fonction de récompense
Peut complètement échouer dans certains cas (e.g. MultiObjectif Produit)

Adaptabilité Peu mise en évidence dans les résultats (peut être un coup de chance ?)

Merci de votre attention

Questions ?

Références I

Anderson, Michael et Susan Leigh Anderson (2015). « Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm ». In : *Industrial Robot : An International Journal* 42.4, p. 324–331.

Cointe, Nicolas (2017). « Ethical Judgment for decision and cooperation in multiagent systems ». Theses. Université de Lyon.

Dyndal, Gjert Lage, Tor Arne Berntsen et Sigrid Redse-Johansen (2017). « AUTONOMOUS MILITARY DRONES-No Longer Science Fiction-. ». In : *Romanian Military Thinking* 2.

Guerin, Frank (2008). « Constructivism in AI : Prospects, Progress and Challenges ». In : *AISB 2008 Convention : Communication, Interaction and Social Intelligence, 1st-4th April 2008, University of Aberdeen, UK*. Sous la dir. de Frank Guerin et Wamberto Weber Vasconcelos. T. 12. Computing & Philosophy. AISB, p. 20–27.

Kohonen, Teuvo (1998). « The self-organizing map ». In : *Neurocomputing* 21.1-3, p. 1–6.

Smith, Andrew James (2002). « Applications of the self-organising map to reinforcement learning ». In : *Neural Networks* 15.8-9, p. 1107–1124.

Références II

Uang, Chang-Hsian, Jiun-Wei Liou et Cheng-Yuan Liou (2012). « Self-Organizing Reinforcement Learning Model ». In : *Intelligent Information and Database Systems - 4th Asian Conference, ACIIDS 2012, Kaohsiung, Taiwan, March 19-21, 2012, Proceedings, Part I*. Sous la dir. de Jeng-Shyang Pan, Shyi-Ming Chen et Ngoc Thanh Nguyen. T. 7196. Lecture Notes in Computer Science. Springer, p. 218–227.

Watkins, Christopher J. C. H. et Peter Dayan (1992). « Q-learning ». In : *Machine Learning* 8.3, p. 279–292.

Yliniemi, Logan et Kagan Tumer (2014). « Multi-objective Multiagent Credit Assignment Through Difference Rewards in Reinforcement Learning ». In : *Simulated Evolution and Learning*. Sous la dir. de Grant Dick et al. Lecture Notes in Computer Science. Springer International Publishing, p. 407–418.