Rapport de stage :
Identification et comparaison d'algorithmes d'apprentissage par renforcement multi-agent appliqués à un contexte éthique

Najib El khadir
Encadré par Salima Hassas (LIRIS) et Rémy Chaput (LIRIS)

Université Claude Bernard Lyon 1, France

Résumé La forte croissance du domaine de l'Intelligence Artificielle a entraîné le développement d'algorithmes destinés à des applications de plus en plus complexes, où des contraintes morales et éthiques peuvent faire face à des besoins de performance et d'efficacité. Les algorithmes et les expérimentations sont nombreux, mais nous nous intéressons ici à la simulation de grilles électriques intelligentes (Smart Grids) par le biais d'un système multi-agent. Plus particulièrement, nous adressons l'apprentissage des agents de cette simulation grâce à des algorithmes d'apprentissage par renforcement, les freins rencontrés par ce mécanisme d'apprentissage, ainsi que les questionnements éthiques soulevés par cette simulation. En premier lieu, nous identifions les problématiques sous-jacentes à ce cas de figure ; ensuite nous passons en revue les différentes approches algorithmiques présentes dans la littérature, en portant une attention particulière à celles utilisant des techniques récentes d'apprentissage profond ; puis nous proposons les approches qui nous semblent les plus adaptées à notre cas d'application éthique ; avant de conclure sur une étude comparative de certaines de ces approches.
Mots-clés: Intelligence Artificielle, Systèmes Multi-Agents, Apprentissage profond, Apprentissage par Renforcement, Réseaux de distribution intelligents, Éthique, Non-stationnarité

Abstract. The strong growth of the field of Artificial Intelligence has led to the development of algorithms for increasingly complex applications, where moral and ethical constraints may face performance and efficiency needs. Algorithms and experiments are numerous, but here we focus on the simulation of Smart Grids through a multi-agent system. More specifically, we address the learning of agents in this simulation through reinforcement learning algorithms, the obstacles encountered by this learning mechanism, and the ethical issues raised by this simulation. First, we identify the issues underlying this case study; then we review the different algorithmic approaches present in the literature, paying particular attention to those using recent deep learning techniques; then we propose the approaches that seem to us the most adapted to our ethical case study; before concluding with a comparative study of some of these approaches.

Keywords: Artificial Intelligence, Multi-Agent Systems, Deep Learning, Reinforcement Learning, Smart Grids, Ethics, Non-stationarity

Table des matières

1 Introduction 3
2 Cas d'étude : *Smart Grids* 5
3 Fondements et définitions 8
4 État de l'art 10
4.1 Éthique et IA 10
4.2 Apprentissage adaptatif de comportements éthiques 12
5 Contributions 15
5.1 Positionnement 15
5.2 Identification des problèmes de coordination dans un système multi-agents coopératif, compétitif ou mixte 15
5.2.1 Problématique identifiée 15
5.2.2 Problèmes de coordination dans un SMA coopératif, compétitif ou mixte 16
5.2.3 Traitement actuel de ces problèmes 17
5.2.4 Contraintes inhérentes 18
5.3 Identification des algorithmes d'apprentissage par renforcement multi-agents 19
5.3.1 Approches *RL* et *MARL* 20
5.3.2 Approches *Deep MARL* 21
5.4 Approche se basant sur un mécanisme d'attention 25
6 Expérimentations et résultats 27
7 Conclusion 28

1 Introduction

Ce travail s'inscrit dans le cadre du projet Ethics.AI[1].

Le domaine de l'Intelligence Artificielle a connu une forte croissance durant ces dernières années, prenant la forme de progrès technologiques divers dont l'échelle peut varier ; allant de l'algorithme de recommandation de votre fournisseur de vidéos à la demande préféré, au projet de gestion intelligente de l'énergie dans des villes grâce aux grilles électriques intelligentes (*Smart Grids*).

La disparité de l'impact que peuvent avoir ou qu'ont déjà eu ces évolutions doit susciter chez la communauté scientifique ainsi qu'auprès de l'opinion publique des intérêts et des inquiétudes adéquats ; notamment lorsque celles-ci concernent des vies humaines et ont pour but de modifier et d'améliorer un élément vital de notre modèle de fonctionnement social.

Un cas de figure comme la gestion de l'énergie d'une ville soulève évidemment des questionnements éthiques et moraux et nécessite un certain regard pluri-disciplinaire, afin de prévenir au mieux les situations indésirables ; ce qui est l'une des préoccupations principales du domaine de l'éthique des machines [12]. C'est dans ce but qu'une multitude de travaux de recherches sont réalisés et qu'un pan de la communauté scientifique de l'Intelligence Artificielle se penche sur ces questions-ci.

Cette problématique est heureusement d'ors et déjà au centre de l'attention des organismes (scientifiques ou non) et nous avons pu constater ces dernières années l'ouverture de groupes de travail nombreux. Notamment un groupe dédié à l'éthique dans les systèmes autonomes et intelligents par l'IEEE (*Institute of Electrical and Electronics Engineers*), ou par exemple la CNIL (Commission Nationale de l'Informatique et des Libertés) qui a émis un document offrant un point de vue plus orienté sur le débat publique.

Ces questionnements concernent donc, en plus des domaines de la sociologie et de la philosohie, plusieurs sous-domaines de l'Intelligence Artificielle, tels que l'apprentissage automatique, les agents autonomes et leurs applications (systèmes multi-agent, apprentissage profond (Deep Learning), ...), etc.

Ceci dit, le but de ce travail ne sera pas de se pencher particulièrement sur la conception d'un modèle éthique pour le cas de figure précité, mais de considérer une modélisation existante de ce problème sous la forme d'un système multi-agents [8] et d'étudier le mécanisme d'apprentissage de ce modèle.

En effet, il s'agira ici de traiter les problématiques sous-jacentes aux mécanismes d'apprentissage autonome de ces dits agents grâce auxquels est

[1] Artificial constructivist agents that learn ETHICS in humAn-Involved co-construction.

modélisée la grille intelligente. Plus particulièrement, nous essayerons de répondre aux questions suivantes :

- QR1 : Quels sont les problèmes de coordination qui nuise à l'efficacité et l'optimalité d'un système multi-agents ?
- QR2 : Quels sont les algorithmes et les modèles d'apprentissage les plus adéquats pour un système multi-agents ?
- QR3 : Quelles considérations éthiques rentrent en jeu lors du choix d'un modèle d'apprentissage pour notre cas d'application ?

2 Cas d'étude : *Smart Grids*

Ce stage s'inscrit dans le cadre d'étude proposé par l'entreprise Ubiant, partenaire du projet Ethics.AI ; c'est le problème de distribution énergétique dans des réseaux intelligents (voir Figure 1).

Smart Grid est un terme qui désigne un réseau d'énergie dit intelligent, c'est-à-dire un réseau qui intègre des technologies de l'information et/ou des mécanismes de communication, visant à améliorer l'utilisation et l'exploitation de l'énergie et à développer de nouveaux usages tels que les véhicules électriques, l'autosuffisance énergétique, etc.

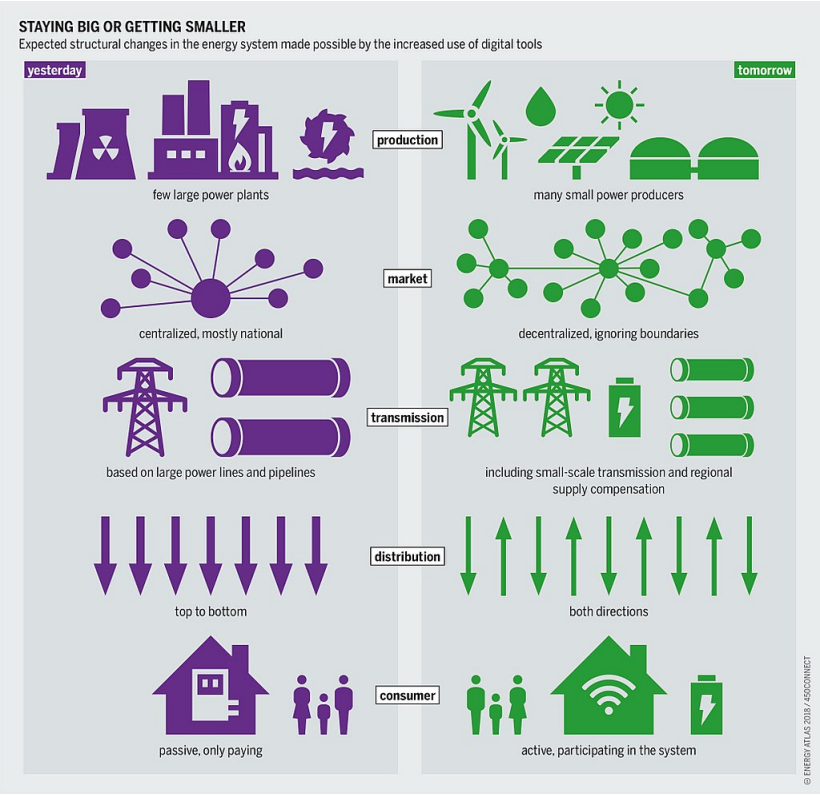

Figure 1. Caractéristiques d'un réseau intelligent - Source : Wikipédia

Plusieurs modélisations concrètes de ce problème existent tels que les centrales électriques virtuelles [11] (voir Figure 2) ou les micro-grilles (voir Figure 3) ; et chacune a ses défauts et ses avantages.

Les centrales virtuelles par exemple, peuvent plus facilement prévoir leur production et combiner un grand volume de sources énergétiques. Mais les prédictions, la coordination des sources, la prise en compte des logiques de marché et des contraintes du réseau est très coûteuse.

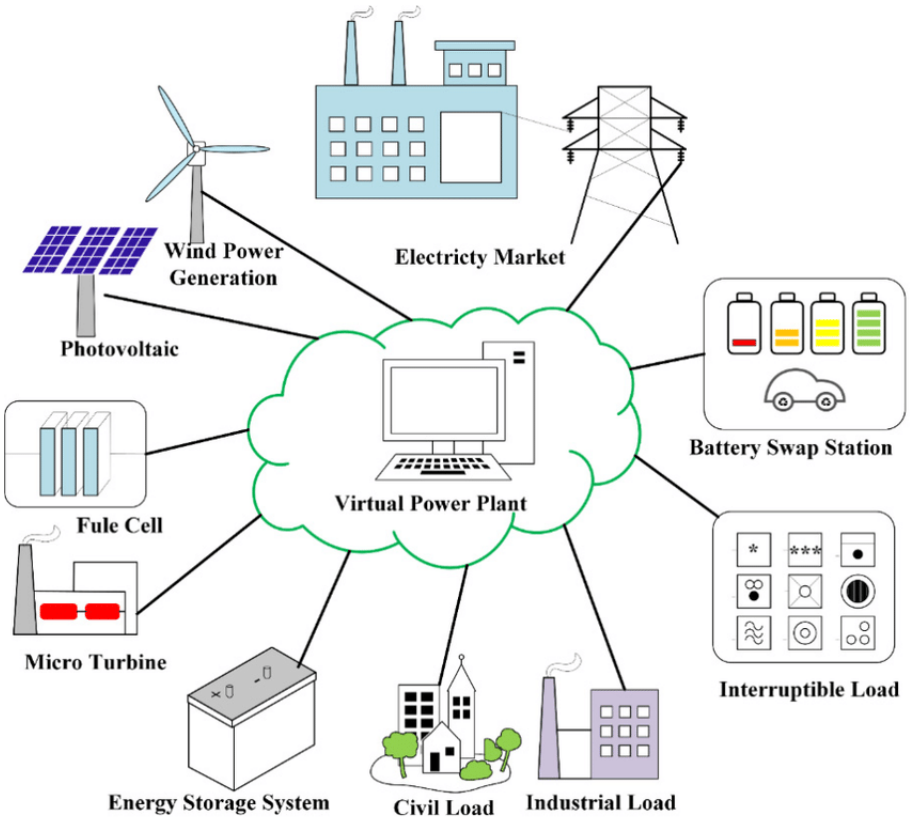

Figure 2. Éléments de base d'une centrale électrique virtuelle

En ce qui concerne les micro-grilles, elles se basent sur le principe de l'interconnexion des différents acteurs afin de minimiser la marge d'erreur. Les acteurs sont ici qualifiés de *prosumers* (producteurs et consommateurs, de l'énergie), c'est-à-dire que chaque habitation ou plus globalement chaque bâ-

timent peut consommer de l'énergie sur le réseau ainsi qu'y injecter de l'énergie qu'il produit grâce à des équipements (photo-voltaïques par exemple) et a des capacités de stockage. Plusieurs contraintes sont inhérentes à ce modèle et en particulier des exigences de planification et d'équité.

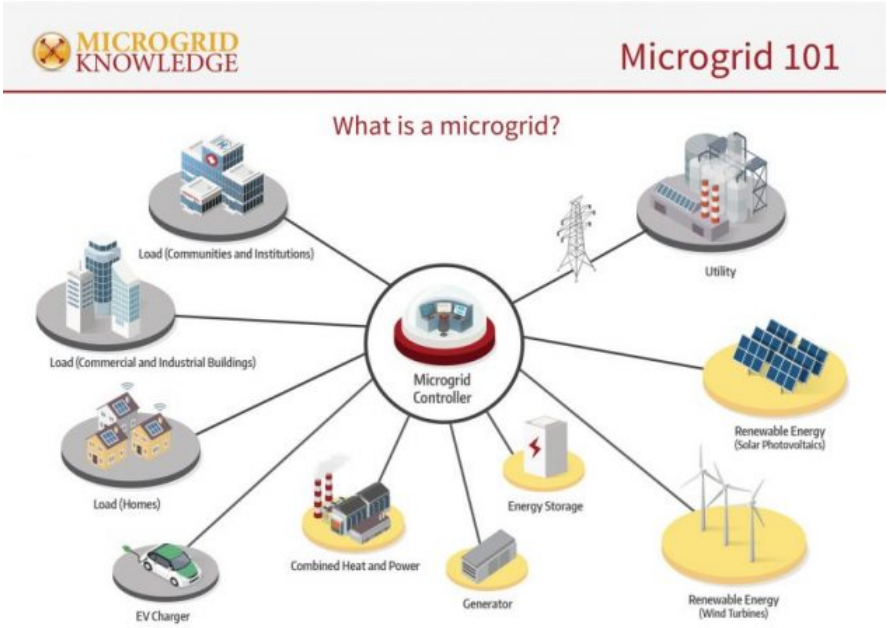

Figure 3. Qu'est-ce qu'une micro-grille ? - Source : https ://www.microgridknowledge.com/

Pour des raisons de passage à l'échelle et d'homogénéité, nous allons nous intéresser dans le cadre de nos travaux à ce dernier type de *Smart Grids*, donc les *Microgrids*. Nous considérons en particulier l'aspect coopératif de celles-ci, qui est mis en évidence par une implémentation qui sera détaillée et discutée plus bas.

Ce cas d'application permet de considérer chaque *prosumer* comme un agent du système et met en avant des problèmes évidents d'opposition de valeurs et d'intérêts (écologie, confort, etc.).

Le système globalement pose des problématiques évidentes de mise à l'échelle :

– Topologie et localisation des acteurs ;

- Coût des équipements ;
- Hétérogénéité des infrastructures matérielles et logicielles ;
- Implication des utilisateurs uniques

Mais également des enjeux de robustesse :

- Modularité et adaptabilité ;
- Communication centralisée ou décentralisée

Et surtout des questionnements éthiques :

- Opposition des valeurs des utilisateurs ;
- Respect de la vie privée ;
- Protection des données ;
- Garantie de l'équité ;
- Garantie de l'optimalité

En addition aux législations et aux débats nécessaires pour l'encadrement de la mise en œuvre réelle de telles infrastructures, les modèles d'apprentissage automatique destinés à ce type d'applications doivent prendre en compte toutes ces considérations dans leur conception et leur mise en œuvre. Et c'est dans cette lignée, que s'inscrivent les travaux de ce stage, ainsi que ceux auxquels il est lié.

3 Fondements et définitions

Pensée cognitiviste

Le cognitivisme est décrit dans [10] tel que le courant de pensée qui estime que l'humain (*Human Machine*) n'est que le fruit de raisonnements sur des informations et des règles concrètes fournies par son environnement, dont il est capable grâce à sa capacité de réflexion pure. Dans ce paradigme, les agents reçoivent des représentations symboliques prédéfinies de leur monde. La sémantique leur permettant de "comprendre" le monde leur est codée à l'aide de règles. Les systèmes cognitivistes ont tendance à être centralisés, à favoriser l'optimalité et la rationalité ainsi que l'absence de curiosité ; et sont le plus souvent destinés à la réalisation de tâches spécifiques.

Pensée constructiviste

L'apprentissage constructiviste [4], quant à lui s'inspire de la théorie de Piaget du développement cognitif chez l'enfant. L'intention est de faire émerger l'intelligence à travers l'apprentissage et l'éducation du système. L'agent n'a pas de représentation globale de son environnement, il apprend à partir de sa propre expérience. L'apprentissage est donc incrémental et les agents construisent leur représentation du monde en fonction des structures qui sont implémentées et de leurs patterns sensorimoteurs. Le focus ici est sur l'adaptation et l'évolution de l'agent.

Apprentissage par Renforcement

Dans le cas de l'Apprentissage par Renforcement [9], l'apprentissage de l'agent est basé sur l'interaction de celui-ci avec l'environnement, l'agent est donc actif. Il apprend par expérience en fonction de ses échecs (punitions) et de ses succès (récompenses) réalisés. L'agent apprend à associer des actions à des situations et a pour objectif de maximiser la récompense.

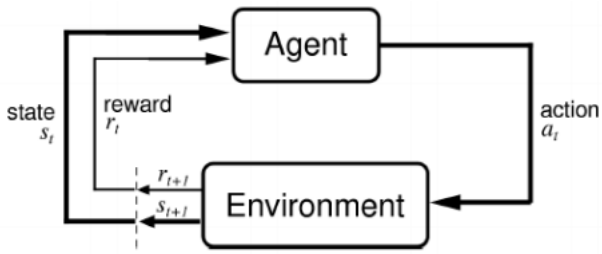

Figure 4. Boucle d'interaction de l'agent avec l'environnement dans le cas de l'Apprentissage par Renforcement

Apprentissage profond

L'apprentissage profond, ou *Deep Learning* est un sous-type de l'apprentissage automatique qui s'appuie sur un réseau de neurones artificiels composé de plusieurs couches (des dizaines, parfois des centaines) de neurones. Ces systèmes ont pour objectif d'apprendre à reconnaître des objets, des symboles réels ou à guider la prise de décisions des agents autonomes dans des contextes plus complexes comme la reconnaissance faciale, les jeux vidéos, le diagnostic médical, etc. Les mauvaises réponses sont éliminées et renvoyées par rétro-propagation vers les niveaux antérieurs pour réajuster le modèle et réorganiser les informations en blocs au fur et à mesure, jusqu'à stabilisation du modèle.

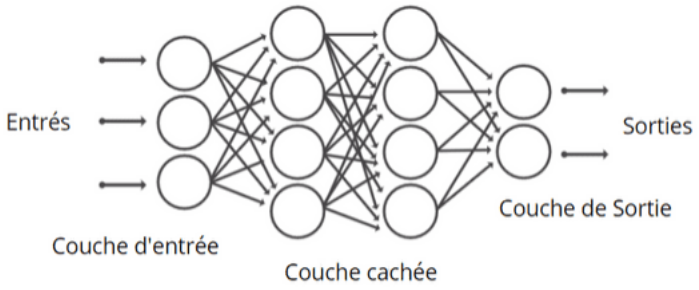

Figure 5. Entrées, sorties et couches dans un réseau de neurones profond

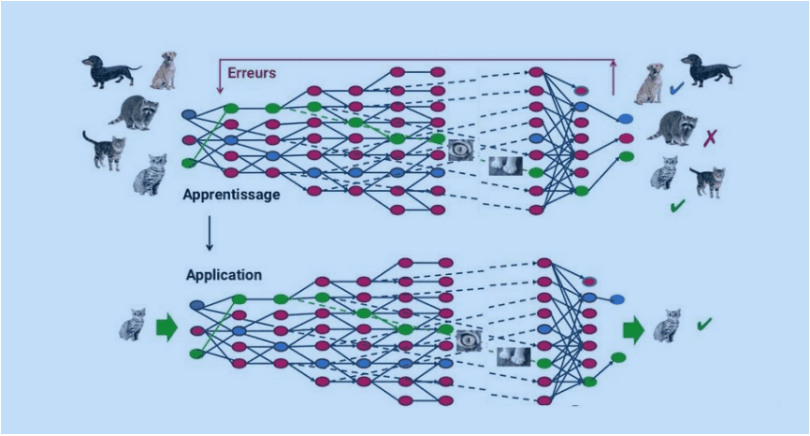

Figure 6. Réseau de neurones profond pour la reconnaissance de chats - Source : https ://www.superdatacamp.com

4 État de l'art

4.1 Éthique et IA

Les systèmes intelligents sont de plus en plus présents dans nos vies quotidiennes et les applications et infrastructures qui les intègrent se démultiplient de jour en jour, mais comme il est dit dans [5] : "Ce qui est technologiquement possible n'est pas toujours humainement ou socialement souhaitable." En effet, au vu de certaines ambitions et la recherche de résultats optimaux,

ces systèmes risquent de transgresser certaines règles morales et éthiques, voire de jouer avec les limites des cadres légaux ; certains le font déjà comme les systèmes se basant sur la collecte d'informations privées des utilisateurs. Il faut donc intégrer des règles éthiques dans les approches dont le but est de concevoir ces systèmes et viser à augmenter l'explicabilité de ceux-ci. Les travaux en *Machine Ethics* se divisent, entre autres classifications, en trois types selon [1] : approches descendantes, ascendantes et hybrides.

Approches descendantes (*top-down*)

Ces approches ont comme objectif la formalisation de principes éthiques. Elles se basent sur un raisonnement logique sur des représentations symboliques, et s'inscrivent donc dans le courant de pensée cognitiviste. Elles peuvent se baser sur des connaissances expertes comme le *Ethical Governor* [13] pour les drones autonomes létaux ou sur des systèmes de règles ou encore des schémas d'argumentation. Elles visent à créer des agents à éthique implicite, c'est-à-dire des agents qui s'appuient sur ce qui a été pré-codé par les concepteurs ; ou des agents à éthique explicite [14], possédant des règles générales (telles que l'Impératif Catégorique de Kant) leur permettant de raisonner sur l'éthique et parfois de justifier leurs choix. Ces agents utilisent des règles qui font que par définition ils ne peuvent pas s'adapter aux nouvelles situations imprévues par les concepteurs. L'un des exemples les plus notables est le projet Ethicaa [2] qui propose un système Multi-Agent dans lequel des agents juges émettent un jugement sur les actions des autres agents, en se basant sur des "croyances" et ce, en fonction de la situation.

Approches ascendantes (*bottom-up*)

Les approches ascendantes ont pour objectif d'apprendre un comportement à partir de jeux de données, des exemples, des *replay* ou des expériences vécues par l'agent grâce à ses interactions avec l'environnement (et donc possiblement les autres agents). Ces approches utilisent des mécanismes d'apprentissage s'inscrivant dans le courant constructiviste tel que l'Apprentissage par Renforcement *(RL)* se basant sur des récompenses qui visent à entraîner des comportements éthiques. Il y a eu plusieurs tentatives telles que GenEth [7] mais elles ne prennnent que peu en compte l'adaptation sur le long terme de l'agent.

Approches hybrides

Les approches hybrides quant à elles sont celles mêlant interactions (apprentissage par expériences) et contraintes éthiques. Elles visent à allier raisonnement symbolique et apprentissage numérique pour tirer profit des deux bords.

4.2 Apprentissage adaptatif de comportements éthiques

En particulier, l'approche la plus aboutie du projet Ethics.AI de simulation de Smart Grids est l'approche proposée par Rémy Chaput & al. [3] ; dont le simulateur consiste en quelques éléments dénombrables :

- Un environnement avec certaines caractéristiques : pas de temps discret, une liste d'agents prédéfinie au début de la simulation, un historique d'actions passées et leurs résultats, des variables partagées, ainsi qu'une instance de régulation entre les agents et l'environnement (appelée Garde-fou) ;
- Des agents qui reçoivent un certain nombre de perceptions qu'ils utilisent pour déterminer une hypothèse d'état, puis décider de l'action à effectuer. Ces agents ont des profils (e.g. leur besoin par heure) différents ainsi que des variables personnelles (e.g. leur batterie) ;
- Des propriétés qui sont des mesures objectives calculées par l'environnement à chaque nouveau tour. Elles sont au nombre de six :
  - L'équité : mesure statistique de la dispersion des conforts ;
  - L'autonomie : le degré de non-interaction avec le marché national d'énergie ;
  - L'exclusion : mesure du taux d'agents qui font plus d'efforts que les autres ;
  - La perte : l'énergie gaspillée, donc non utilisée ;
  - Le bien-être : la moyenne du confort des agents ;
  - La surconsommation : l'énergie prise en excès
- Des perceptions fournies par l'environnement aux agents sous forme de vecteur à chaque nouvelle étape de la simulation. Elles se divisent en deux groupes :
  - Les perceptions relatives à l'environnement : l'heure, l'énergie (disponible dans la grille) et les propriétés ;
  - Les perceptions relatives à chaque agent : le stockage (état de la batterie de l'agent), le confort (satisfaction au pas de temps précédent) et le profit (énergie vendue - énergie achetée par l'agent)
- Des récompenses calculées par l'environnement pour chaque agent

Modèle d'apprentissage et de prise de décision

Le modèle utilisé dans [3] est une fusion du Q-Learning (un des algorithmes d'Apprentissage par Renforcement les plus populaires) pour l'apprentissage de l'intérêt d'une action dans un état donné et les Cartes auto-organisatrices de Kohonen pour la correspondance entre états discrets et vecteurs de perceptions continus. Les figures (7) & (8) résument le processus de décision, d'action et d'apprentissage des agents dans le simulateur :

1. L'environnement envoie des perceptions de l'état actuel aux agents au début de chaque étape de la simulation, ainsi qu'une récompense si nous ne sommes pas à l'étape initiale ;

2. L'agent discrétise l'état en utilisant la carte SOM des entrées ;

3. L'agent associe l'hypothèse d'état à la récompense pour déterminer si l'action précédente a été fructueuse et modifie la carte des actions et la Q-table en conséquence ;

4. Pour décider de la prochaine action, l'agent reçoit un vecteur de perception X et détermine le neurone U_i dont le vecteur associé est le plus proche de X dans la carte des entrées ; U_i donne une hypothèse d'état e_i, on regarde donc dans la Q-table la ligne i ; l'agent choisit son action via une politique de Boltzmann avec température, l'action choisie a_j indique le vecteur associé W_j dans la carte des actions ; du bruit est également rajouté pour faciliter l'exploration des actions. En résulte un vecteur d'action perturbé ;

5. L'instance de régulation (Garde-fou) intervient si nécessaire, afin de limiter les actions abusives qui sortent du cadre des permissions du simulateur (consommations aberrantes, etc.) ;

6. Une fois que tous les agents ont émis leur action, l'environnement les traite et passe dans un nouvel état ;

7. L'environnement calcul les récompenses pour chaque agent, les nouvelles perceptions et les transmet aux agents ;

8. Si l'action perturbée a été plus efficace que l'action proposée, on met à jour la carte des actions, les Q valeurs et la carte des entrées.

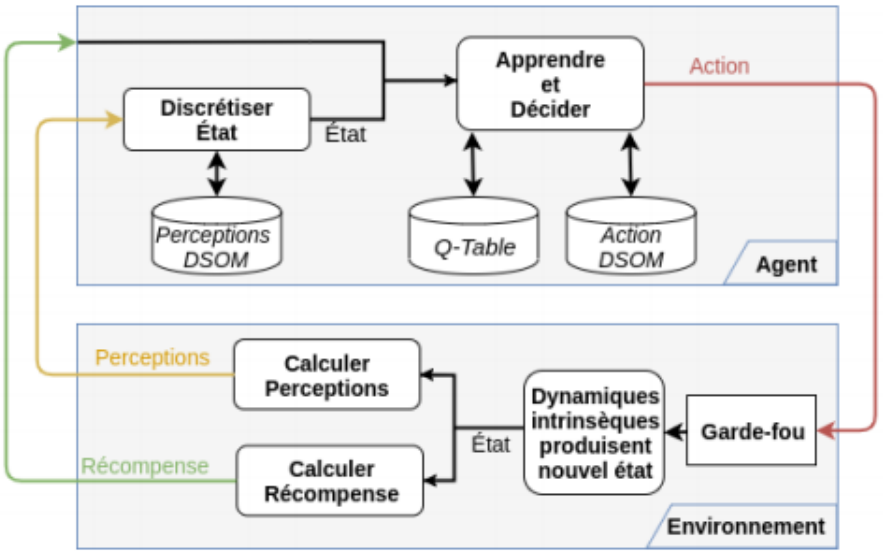

Figure 7. Cycle d'apprentissage de chaque agent interagissant avec l'environnement - *Chaput & al. 2020*

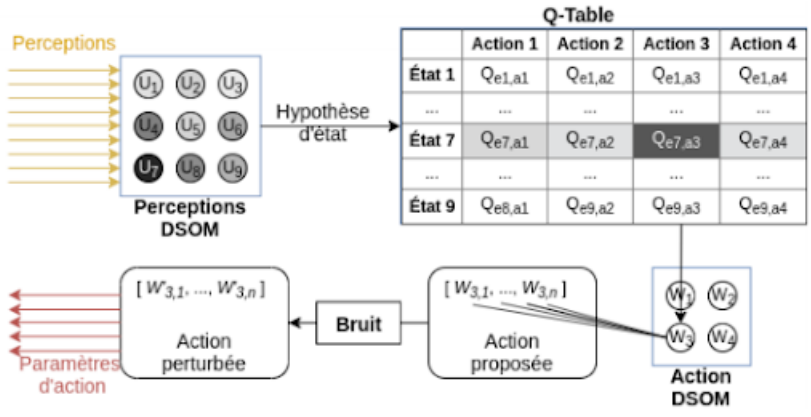

Figure 8. Processus de décision de chaque agent - *Chaput & al. 2020*

Injection éthique

L'éthique dans cette simulation est de l'éthique par conception. Elle est injectée grâce à la fonction de récompense des agents qui prend en compte leur implication dans l'évolution des propriétés de l'environnement et des agents : cette fonction a été déclinée sous plusieurs exemples permettant de plus ou moins accorder d'importance au facteur d'équité dans la simulation, à la minimisation de la surconsommation et à la prise en compte du confort personnel, du bien être des autres agents, etc. Le lecteur est invité à consulter [3] pour plus de détails sur ces fonctions de récompense.

5 Contributions

5.1 Positionnement

Ce stage s'inscrit dans la continuité des travaux de Rémy Chaput & al. [3]. Pour l'instant les standards de la simulation sont le modèle décrit plus haut mêlant *Q-learning* et cartes auto-organisatrices, nommé *Q-SOM* ; ainsi que la fonction de récompense *Adaptability2* [3] qui est une fonction qui change au cours du temps permettant bien de vérifier la notion d'adaptabilité du processus d'apprentissage, cette fonction vise à minimiser la surconsommation au sein de la simulation, puis rentre en compte au bout de 2000 pas de temps l'équité des actions réalisées par les agents, et enfin le confort au bout de 6000 pas de temps.

Le but de ce stage est de se focaliser sur le modèle d'apprentissage et comment l'améliorer ; et ce à travers trois objectifs établis :

- L'identification des problèmes qui ralentissent l'apprentissage des agents ;
- L'identification des algorithmes d'apprentissage adéquats expérimentés dans la littérature, avec un focus sur les technologies récentes d'apprentissage profond ;
- La proposition d'un ou plusieurs modèles pour améliorer le processus d'apprentissage et de prise de décision des agents.

5.2 Identification des problèmes de coordination dans un système multi-agents coopératif, compétitif ou mixte

5.2.1 Problématique identifiée

Approcher les *Smart Grids* par un *MARL* est prometteur, mais nécessite d'explorer des approches plus complexes algorithmiquement.

En effet, l'apprentissage isolé de chaque agent ne garantit pas la convergence vers une solution optimale. L'un des objectifs principaux des *Smart*

Grids étant d'améliorer et d'optimiser la production et la distribution énergétique, cela soulève une réelle problématique.

Ce manque d'efficacité s'explique par la nature même de l'environnement qui se veut semi-coopératif. En effet, chaque agent (prosumer) a des objectifs propres aux valeurs éthiques et au mode de vie qu'il soutient : confort, écologie, vie privée, etc. mais dépend de la dynamique de la *Smart Grid* et de ses autres composants, ici les autres *prosumers*. Et donc, les choix et les actions des autres agents ainsi que leur impact sur l'environnement peut être un frein à l'apprentissage efficace de chaque agent.

Cette problématique est l'un des sujets phares de la recherche dans le domaine des SMA dans un environnement coopératif, compétitif, collaboratif ou mixte. Nous allons par la suite dresser un inventaire des problèmes qui guident la recherche ainsi qu'un état des lieux des modèles de *MARL* qui ont été proposés ces dernières années, notamment ceux qui intègrent des mécanismes de *Deep Learning*.

5.2.2 Problèmes de coordination dans un SMA coopératif, compétitif ou mixte

Non-stationnarité de l'environnement [15]

Dans le modèle actuel de Chaput & al. [3], chaque agent détermine une hypothèse d'état à partir de perceptions qui décrivent l'état de l'environnement puis décide de l'action à effectuer. Les pas de temps sont discrets, à chaque pas de temps les agents perçoivent, apprennent et agissent les uns après les autres et leur ordre de passage est tiré aléatoirement à chaque pas de temps.

Dans un tel environnement, la pertinence des observations réalisées par chaque agent se trouve remise en question puisque les dynamiques de l'environnement sont dépendantes des comportements de tous les agents ainsi que leur évolution dans le temps. L'environnement est donc perçu comme non stationnaire par chaque agent et l'apprentissage de ceux-ci peut se trouver biaisé.

De plus, dans un environnement réaliste où les *prosumers* agissent de manière asynchrone sur la *Smart Grid*, la propriété de stationnarité de l'environnement qui est essentielle pour garantir la convergence du système est vivement compromise. Du point de vue d'un agent, le processus local n'est pas stationnaire, car il doit potentiellement changer sa politique pour améliorer celle des autres ; et les politiques des autres agents ne sont également pas stationnaires et nécessitent de s'y adapter.

Stochasticité de l'environnement

Du point de vue d'un agent, l'état de l'environnement n'est pas gouverné par l'impact ses actions, mais par les actions jointes de tous les agents de

la *Smart Grid*. Plus explicitement : en réalité, si à un pas de temps un prosumer décide qu'il est adéquat de consommer et d'injecter telles quantités d'énergie dans la grille, car l'état dans laquelle elle serait après-coup semble convenable ; ce jugement se trouve totalement faussé par les actions des autres agents.

Si cela n'est pas pris en compte dans la conception, les paramètres d'évolution de la grille deviennent aléatoires et l'environnement est perçu comme stochastique par chaque agent.

Pareto-optimalité

Quand ce type de SMA atteint une solution stable, l'ensemble des politiques résultant ne constitue pas directement une solution Pareto-optimale, c'est-à-dire une solution où aucune politique n'est améliorable sans porter préjudice à un autre agent.

Dans le cas des *Smart Grid*, ce problème est d'autant plus flagrant, car l'optimalité de la distribution énergétique entre les *prosumers* n'est pas garantie ; et mérite d'autant plus de s'y pencher.

Alter-exploration

Dans la littérature, le bruit de l'exploration est défini comme étant la probabilité qu'au moins un agent explore et peut être quantifiée grâce aux taux d'exploration individuels de chaque agent.

L'exploration d'un ou de plusieurs agents apprenants dans un SMA coopératif peut influencer, changer, voir détruire les politiques des autres agents ou leur induire des pénalités. L'adaptation des autres agents à ce changement peut entrainer des instabilités et ainsi ralentir ou empêcher la convergence du système.

Multi-Agent Credit Assignment

Dans un modèle totalement adapté, il est important de récompenser chaque agent de manière relative à l'effort fourni dans le système (e.g. quantité d'énergie transmise à la grille, minimisation de la surconsommation, minimisation du confort personnel, etc.) sans pénaliser les autres agents qui ont suivi leur politique optimale.

C'est un problème à prendre en compte lors de la conception de modèles d'Apprentissage par Renforcement Multi-Agents en général ; et cela semble encore plus pertinent dans un cas d'application qui intègre des propriétés éthiques.

5.2.3 Traitement actuel de ces problèmes

L'impact de ces problèmes est actuellement toléré, car le comportement des agents est assumé comme non nocif pour les autres agents. En particulier,

un garde-fou a été implémenté avec comme rôle de simuler une transaction entre la grille locale et le réseau national en cas de surconsommation de la part d'un agent.

Également l'injection éthique proposée dans ces travaux se faisant grâce aux fonctions de récompense des agents, permet selon chaque fonction de récompense de prendre en compte l'équité dans la distribution des conforts ou de minimiser la sur-consommation dans la grille. Les expérimentations antérieures se sont également focalisées sur la combinaison de plusieurs fonctions de récompense pour faire de l'apprentissage multi-objectifs, ainsi que sur l'apprentissage adaptatif où les fonctions de récompense des agents changent au cours du temps ; et ce avec des résultats plus ou moins convaincants.

De plus, il est important de noter que dans les travaux actuels les *prosumers* ont des profils de consommation différents (trois catégories) mais suivent la même fonction de récompense à chaque expérimentation. Il est donc intéressant pour les travaux futurs de garder en perspective qu'en situation réelle les *prosumers*, donc les agents, peuvent avoir des objectifs différents, donc des calculs de récompense différents.

5.2.4 Contraintes inhérentes

Nous allons maintenant lister quelques contraintes qui doivent être prises en compte lors du choix des modèles à approfondir dans le cadre de ce stage ainsi que leur conception. À noter que la relaxation de ces contraintes par certains modèles ne doit pas nécessairement être un obstacle total à l'implémentation de ceux-ci à des fins expérimentales si cela vise à améliorer les performances dans la simulation de *Smart Grid*.

Scalabilité

L'extensibilité d'un modèle est toujours une propriété à avoir en tête lors de la conception de celui-ci. D'autant plus lorsqu'il s'agit d'un projet à grande échelle comme celui des *Smart Grids*.

Beaucoup des algorithmes de *MARL* dans la littérature gardent cette contrainte au cœur de la recherche, car elle limite souvent les performances algorithmiques.

Des exemples dont nous reparlerons plus tard sont les algorithmes se basant sur le mécanisme de *Joint Action Learning* ou ceux se basant sur la communication inter-agents. En effet, l'espace des actions possibles avec ce premier type d'algorithme grandit exponentiellement avec le nombre d'agents, et il va de même pour le coût de la communication avec ce second ; pour certains leur extensibilité est quasiment nulle.

Adaptabilité

Une des propriétés les plus recherchées dans un système Multi-Agent est sa robustesse face aux changements ; dans le cas des Smart-Grids, les agents doivent théoriquement être capables de s'adapter si une ligne du réseau est obstruée, si un agent est supprimé, etc.

Le modèle d'apprentissage des agents peut également intégrer la notion d'adaptabilité. En particulier, il est ici question de s'adapter au changement dans les objectifs du prosumer et donc de l'agent ; et ce, suite à un déménagement, un changement de conviction, une évolution dans les valeurs, etc.

Observabilité

Beaucoup de modèles d'Apprentissage par Renforcement visant à optimiser les performances d'un SMA coopératif nécessitent que chaque agent puisse observer les historiques d'actions des autres agents et/ou leurs récompenses. Or, dans une structure comme les *Smart Grid*, cet accès aux informations des autres agents n'est pas forcément garanti parce que les utilisateurs peuvent refuser la réutilisation de leurs informations de consommation par des algorithmes qui leur sont opaques, ou parce que le partage de ces informations peut représenter un coût conséquent, voir nécessiter une couche architecturale supplémentaire difficile à mettre en œuvre. Il est important donc de distinguer les algorithmes de *MARL* en fonction de leur positionnement au niveau de cette propriété.

Pour plus de détails et d'explications générales concernant ces problématiques, le lecteur est invité à consulter [16] pour le cas des agents apprenants indépendants, [15] pour le cas des systèmes multi-agents, ainsi que [17] qui traite du contexte du *Deep MARL*.

5.3 Identification des algorithmes d'apprentissage par renforcement multi-agents

Dans cette partie nous allons discuter les différents algorithmes d'apprentissage par renforcement qui ont été conçus pour les systèmes multi-agents (*MARL : Multi-Agent Reinforcement Learning*) présents dans la littérature.

En particulier, nous allons dénombrer les approches ainsi que quelques détails de leurs conceptions, nous allons également les distinguer en fonction de leur positionnement sur certaines propriétés et des cas d'expérimentations dans lesquels ils ont été testés.

Cette partie de la contribution se présente sous forme d'une revue et l'annexe A (7) en est une synthèse. Dans un premier temps, les approches sont séparées en deux catégories :

- Celles qui s'apparentent à de l'apprentissage par renforcement multi-agents classique ;
- Celles qui utilisent des technologies d'apprentissage profond (*Deep Learning*) : les approches de type *Deep MARL*. Ce sont ces approches sur lesquelles se sont le plus concentrées nos recherches.

5.3.1 Approches *RL* et *MARL*

Ces approches représentent la base des recherches qui ont été effectuées dans le domaine de l'apprentissage par renforcement avec plusieurs agents interagissant dans et avec l'environnement, elles sont dérivées du Q-learning pour la plupart et sont pour la majorité assez archaïques et ne représentent donc que peu d'intérêt pour cette contribution. Parmi ces algorithmes nous trouvons le *Decentralized Q-learning* [18] qui ne traite aucun problème en particulier car il se base sur du *Q-learning* décentralisé classique.

Il a également été théorisé le *Distributed Q-learning* [19] qui est une tentative de réponse aux problèmes d'alter-exploration et de pareto-optimalité du contexte multi-agents [16] en utilisant un principe de récompense jointe calculée par l'environnement et distribuée entre les agents, cela n'est pas sans rappeler les récompenses au mérite calculées et envoyées aux agents dans [3]. Malheureusement cette approche ne passe pas bien à l'échelle, car la récompense jointe explose avec le nombre d'agents.

D'autres approches ont également été formalisées. Nous retrouvons les *Hysteretic Learners* de Matignon & al. [20] qui sont des agents qui apprennent avec deux taux d'apprentissage différents en fonction de si l'agent est récompensé ou puni ; ici rentre en jeu un mécanisme d'optimisme dans le calcul des nouvelles Q-valeurs. WoLF-PHC [21] est également un algorithme qui fonctionne grâce à une heuristique se basant sur deux taux d'apprentissage.

Quant à l'algorithme Hyper-Q [22], il se base sur une Q-fonction qui dépend de l'estimation de la stratégie mixte de tous les agents afin de répondre à la problématique de non-stationnarité. Mais comme nous pouvons nous y attendre, le passage à l'échelle d'une telle approche se trouve très compliqué de par l'explosion de l'espace des stratégies mixtes possibles avec l'augmentation du nombre d'agents.

Il existe également d'autres approches qui s'éloignent du *Q-learning*. L'une des plus notables et des plus récentes (2014) dans le domaine du *RL* basique étant RL-CD [15] qui, dans l'effort de répondre au problème de non-stationnarité, considère les environnements stationnaires au cours de l'apprentissage comme des contextes pour lesquels des modèles partiels sont

appris ; ce qui fait que les changements dans les dynamiques de l'environnement entraînent un passage d'un contexte stationnaire à un autre (avec chacun ses propres dynamiques distinctes) du point de vue des agents.

5.3.2 Approches *Deep MARL*

Ces approches sont le fruit du croisement des domaines de la recherche en apprentissage profond (*Deep Learning*) et celui de l'apprentissage par renforcement multi-agents (*MARL*). Elles se divisent en plusieurs catégories dont les principales sont :

- Les approches IQL (pour *Independent Q-Learning*) qui, comme leur nom l'indique, sont des approches qui ne considèrent pas le système multi-agents en entier, mais bien l'apprentissage indépendant de chaque agent ;
- Les approches *Fully Observable Critic* qui peuvent être basées sur du *Q-Learning*, sur une architecture *Actor-Critic* [23] ou les deux ;
- Les approches se basant sur des fonctions de décomposition de valeur (*Value Decomposition Functions*) ;
- Les approches se basant sur un mécanisme de consensus entre les agents.

Approches *IQL*

Ces approches considèrent que les autres agents font partie de l'environnement. Ainsi, elles ne souffrent pas de la scalabilité, mais en contrepartie faiblissent face aux problèmes inhérents aux architectures des systèmes multi-agents. La seule approche utilisant des agents indépendants qui pourrait nous intéresser est l'approche combinant *IQL* et Q-learning profond (*DQL*) de Tampuu & al. [24] ; cette approche utilise en effet un réseau de neurone profond de Q-learning *DQN* au niveau de chaque agent, ce qui la rend très prometteuse. En revanche, elle nécessite d'utiliser une *Replay Memory* qui éprouve des difficultés à se stabiliser à cause de la non-stationnarité de l'environnement. Des extensions de cet algorithme ont été étudiées afin de réussir à stabiliser la Replay Memory, en utilisant des agents *Hysteretic*, des réseaux de neurones profonds récursifs (*DRQN*), etc. Cet algorithme a eu des résultats satisfaisants dans des expérimentations sur des jeux compétitifs ou coopératifs simples avec deux agents.

Approches *Fully Observable Critic*

Dans ces approches nous retrouvons le très populaire *MADDPG* [25] (pour *Multi-Agent Deep Deterministic Policy Gradient*). Cet algorithme est une amélioration de *DDPG* utilisant un *Critic* centralisé par agent ; ce qui fait qu'elle passe difficilement à l'échelle, mais obtient de meilleurs résultats que *DDPG* dans des expérimentations de communication ou de *predator-prey*.

Il existe des extensions de *MADDPG* telle que *MADDPG-GCPN* [26] pour *Generative Cooperative Policy Network*, où chaque agent apprend une politique déterministe qui prend en compte lors de son apprentissage des échantillons d'actions des autres agents (passés en entrée au *Critic*) générés par un réseau d'*Actor* supplémentaire. De part cette dernière innovation, cette solution est assez coûteuse et passe très difficilement à l'échelle même s'il y a eu des extensions très prometteuses pour des problèmes à N agents, où les problèmes restent simples.

Une des pistes les plus intéressantes est l'*Actor-Attention-Critic for MARL* de Iqbal & al. [27] qui ajoute un mécanisme d'attention (qui sera décrit plus en détail plus tard dans cette contribution) permettant de répondre à plusieurs des problématiques inhérentes à un SMA (non-stationnarité, Pareto-optimalité, scalabilité, robustesse) dans des environnements coopératifs.

Un autre algorithme très étudié dans la littérature est *COMA* [28] *(COunter factual Multi Agent policy gradients)*, qui utilise des *Actor* entraînés localement et un seul *Critic* centralisé ; cette approche se base pour le calcul des actions des agents sur une méthode appelée *Counterfactual Baseline* et obtient des résultats meilleurs qu'une grande partie de l'état de l'art *Actor-Critic* dans des jeux compétitifs et coopératifs tels que StarCraft, qui est l'un des benchmarks les plus utilisés.

D'autres approches de ce type sont également intéressantes, mais demandent une certaine refonte de l'infrastructure du projet telles que CM3 [29] (*Cooperative Multi-goal Multi-stage Multi-agent RL*) où les agents sont d'abord entraînés sur leurs objectifs (e.g. objectifs locaux de confort ou de non-surconsommation) puis sur l'objectif commun (e.g. équité dans la *Smart Grid*) ; ou GCRL [30] (*Graph Convolutional RL*) où le SMA est représenté sous forme d'un graphe et chaque agent partage ses observations à ses voisins directs, cette approche a prouvé son efficacité dans plusieurs expérimentations dans des environnements coopératifs en grille ; ou encore *Mean Field RL* [31] où chaque agent prend en compte l'effet moyen de son voisinage dont la limite est définie, ce qui passe mieux à l'échelle qu'une communication exhaustive entre tous les agents.

Approches *VD (Value Decomposition)*

Les approches discutées dans cette section sont des approches qui s'attaquent au cas centralisé : les informations sont partagées entre les agents et il n'y a pas de limite de communication ; ce qui en fait des approches relativement naïves lorsqu'elles sont appliquées en apprentissage par renforcement. Mais la solution proposée est de déterminer le rôle de chaque agent dans ce qui est appelée la récompense jointe, et ce, soit grâce à des récompenses différentiées : $R_i = R - R_{-i}$, où R_i est la récompense de l'agent, R la récompense unifiée de tout les agents et R_{-i} la récompense unifiée des

autres agents ; ou des récompenses modelées en se basant sur le potentiel (*Potential Based Reward Shaping*) : on remplace R par $R + \Phi(s') - \Phi(s)$ où $\Phi(s)$ désigne la désirabilité de l'agent d'être dans l'état s.

Sunehag & al. proposent donc VDN [32] (*Value Decomposition Network*) qui se base sur une unique récompense partagée dans des contexte d'apprentissage multi-agents coopératifs, ce qui en fait une approche qui passe difficilement à l'échelle, puisque le réseau apprend une Q_{tot} globale qui est calculée par additivité, puis décomposée en N Q-fonctions pour N agents. De par sa mauvaise scalabilité, cette approche n'a montré des résultats que dans des contextes impliquant deux agents, mais elle a pavé la route pour d'autres approches.

Ainsi, en 2018 fut proposée QMIX [33], cette approche répond à plusieurs problèmes trouvés dans les SMA et permet de représenter une classe plus vaste de fonctions action-valeur que VDN. Chaque agent apprend grâce à un *DRQN*, ce qui en fait une approche à la scalabilité réduite, et grâce à un réseau commun de "mix" permettant de les combiner. Elle a notamment eu d'excellents résultats sur StarCraft.

D'autres approches existent, telles que QTRAN [34] qui obtient de meilleurs résultats dans des tâches plus variées que ses prédécesseurs grâce à l'utilisation de trois réseaux combinés pour l'apprentissage de chaque agent : un pour les Q_i individuelles, un pour Q_{tot} et un réseau de régulation. Ou encore le méta-agent proposé par Mguni & al. [35] qui introduit un agent entraîné par un *Actor-Critic* dont le rôle est de modifier les récompenses des autres agents ; mais dont l'entraînement se trouve être très coûteux même si les expérimentations se sont conclues sur d'excellents résultats dans une tâche de réorganisation de 2000 agents dont la position désirée changeait au fil de l'expérience.

Approches basées sur le consensus

Il existe également des approches se basant sur le consensus des agents, dont le but est de trouver un accord entre les agents voisins, et ce, bien évidemment, en limitant le voisinage de communication afin de garder la quantité de communication linéaire par rapport au nombre de voisins. Parmi ces approches, deux se démarquent particulièrement.

Le *Diffusion-Based Ditributed Actor-Critic* ou Diff-DAC, qui a été proposé en 2020 [36], implique l'existence d'un chemin entre chaque agent. Le but avec cette approche est de trouver une politique moyenne qui permet aux agents de performer sur un certain nombre de tâches similaires ou différentes les unes des autres, en entraînant des agents en parallèle sur les tâches jusqu'à obtention d'un consensus dit "moyen". Cette approche passe très bien à l'échelle, mais son utilisation est réservée à des contextes de multitâches.

Egalement il existe une approche mêlant *Actor-Critic* et Consensus du nom de *Networked Agents for Fully Decentralized MARL* [37] où les agents décident en se basant sur leurs observations et sur les communications qu'ils entretiennent avec leurs voisins. Dans ce type d'approches il est important de noter que la problématique de vie privée discutée précédemment se retrouve au centre de la conception puisqu'il devient nécessaire d'encrypter les données transmises d'un agent à l'autre, car elles peuvent être plus ou moins sensibles.

Approches *Learn 2 Communicate*

Il existe également un bon nombre d'approches basées sur la communication dans la littérature. Ces approches, telles que ComNet [40], BiCNet [39], IC3Net [38], TarMAC [41](*Targeted Multi-Agent Communication*) et ATOC [42](*Attentional Communication*), sont des protocoles de communication utilisant des technologies de *Deep Learning* qui peuvent être combinés à de l'apprentissage profond par renforcement afin de structurer la communication entre les différents agents dans un contexte SMA ; mais elles représentent une couche architecturale additionnelle, donc un certain coût, lors de leur conception et de leur mise en œuvre.

Les approches qui ont été détaillées ci-dessus représentent une liste non exhaustive des approches d'apprentissage par renforcement multi-agent ; et constituent une base solide pour l'identification des approches les plus adéquats à notre cas de figure de simulation des *Smart Grids*.

Dans ce but, une synthèse (7) a été réalisée et détaille les critères explicités et expliqués dans la première page de celle-ci.

De par les différents critères relevés et synthétisés que sont l'observabilité (des états, des actions et des récompenses), la nature de l'environnement, les particularités algorithmiques de l'approche, les problèmes qu'elle traite, ses défauts et ses cas d'applications notables ; les approches les plus prometteuses sont *MADDPG* [25], *COMA* [28] et l'approche se basant sur un mécanisme d'attention (*Actor-Attention-Critic for MARL* de Iqbal & al. [27]).

5.4 Approche se basant sur un mécanisme d'attention

Dans le cadre des travaux antérieurs à ce stage, il a déjà été effectué une implémentation de *MADDPG* [25] dans le module d'apprentissage du simulateur [3] donc nous n'allons pas détailler la conception de cette approche. Également, il a été tenté sans réussite l'implémentation de *COMA* [28]. Nous allons donc présenter plus en profondeur l'approche *MAAttC (Multi-agent Actor-Attention-Critic)* [27] et ses particularités dans cette partie.

Cette approche est une approche où chaque agent entraîne un *Actor* décentralisé pour sa politique et un *Critic* centralisé ; le mécanisme d'Attention rentre en jeu afin d'identifier les informations les plus pertinentes pour chaque agent à chaque pas de temps.

Les approches dites simples violent la règle de stationnarité de l'environnement. *MADDPG* [25] et *COMA* [28] combinent les forces des approches centralisées (QMIX, *Joint Action Learners*, etc.) et des approches décentralisées (IQL, etc.) ; elles constituent une bonne évolution dans le domaine du *MARL* mais ne suffisent pas à répondre aux problématiques de scalabilité, de non-stationnarité, etc. inhérentes aux SMA.

C'est dans le but d'étendre ces modèles et de les améliorer qu'a été proposé *MAAttC*[27], et ce, en utilisant le mécanisme d'Attention qui a déjà prouvé son efficacité dans les domaines de la vision par ordinateur [44] et du traitement automatique des langues *(NLP)* [43]. L'algorithme proposé dans [27] est très polyvalent et adaptable en matière de récompenses, d'espaces d'actions, de natures d'environnement et de passage à l'échelle ; ce qui en fait un bon candidat pour intégrer le module d'apprentissage du simulateur [3].

MAttAC - Fonctionnement

Les Q_i sont calculées après que chaque agent a encodé ses observations et ses actions puis les a envoyées à la "centrale d'Attention" (9) ; puis chaque agent reçoit une somme pondérée des encodages des autres agents transformés par une matrice de normalisation V.

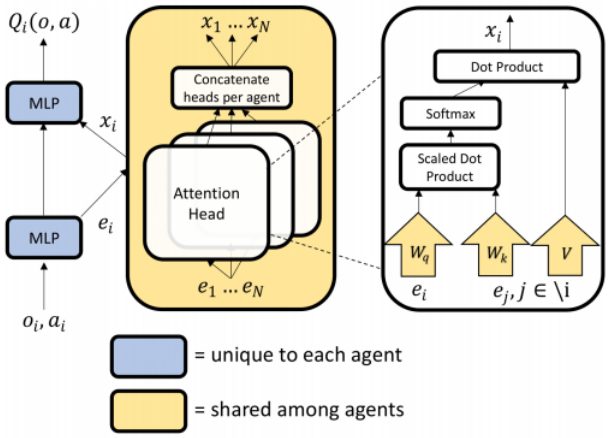

Figure 9. Calcul de la Q valeur ${Q_i}^{\psi}(o,a)$ pour un agent i [27]

Fonction de valeur (${Q_i}^{\psi}(o,a)$) et contribution des autres agents (x_i) :

${Q_i}^{\psi}(o,a) = f_i(g_i(o_i,a_i),x_i)$
$x_i = \sum_{j \neq i} \alpha_j v_j = \sum_{j \neq i} \alpha_j h(V_{g_j}(o_j,a_j))$

Plus précisement :

- Chaque agent à chaque pas de temps interroge les autres agents sur leurs observations et leurs actions, ici les agents interrogent sur les quantités d'énergies qui ont été consommées, produites, stockées et injectées dans le réseau. Ces informations sont incorporées dans la propre fonction de valeur de l'agent ;
- Le *Critic* de l'agent a_i reçoit les observations et les actions des autres agents indexés par i ;
- Ici, f_i est un MLP (*Multi-Layer Perceptron*) à deux couches ;
- g_i est une fonction d'embedding de l'agent a_j encodée puis tranformée par la matrice de normalisation partagée V ;
- h est un élément de non-linéarité, ici un Leaky ReLU (10) ;
- Les têtes d'Attention sont multiples afin de varier les possibilités dans la simulation ;
- Chaque tête d'Attention utilise un set de paramètres (W_k,W_q,V) séparé et permet d'obtenir une agrégation des contributions des autres agents à l'agent a_i ;

– Ces contributions sont calculées par chaque tête puis concaténées sous la forme d'un vecteur unique. Ainsi, chaque tête peut se focaliser sur une pondération différente des participations des autres agents.

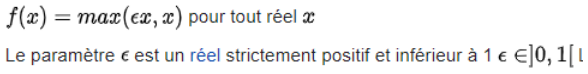

Figure 10. Fonction de Leaky ReLU - Source : Wikipédia

6 Expérimentations et résultats

Expériences réalisées

Beaucoup de difficultés ont été rencontrées lors de l'intégration de l'algorithme *MAttAC*. La difficulté principale étant la complexité technique de l'intégration d'une telle approche dans le module d'apprentissage du simulateur. Au-delà de la documentation sur les différentes approches, l'intégration de ceux-ci aux travaux déjà réalisés constitue une charge de travail remarquable qui n'a malheureusement pas encore aboutie.

Ainsi, pour cette partie, nous nous contenterons de dresser une brève étude comparative entre les performances de *MADDPG* [25] et *Q-SOM* [3] ; afin de voir si nous obtenons de meilleurs résultats en utilisant un algorithme plus adapté au contexte multi-agent et qui plus est, une technologie plus récente.

Dans ce but, nous avons donc réalisé une série d'expériences dans le simulateur de Chaput & al. [3]. En particulier, nous avons réalisé :

– 20 expériences de 10000 étapes avec la fonction de récompense R_{equity} décrite dans [3]. Dont 10 en utilisant l'algorithme *Q-SOM* et 10 en utilisant l'algorithme *MADDPG* ;
– Ainsi que 20 expériences de 10000 étapes avec la fonction de récompense $R_{adaptability2}$ décrite dans [3] ; cette fonction prend en compte la surconsommation de l'agent avant l'étape 2000, puis introduit la récompense basée sur l'équité $(R_{equity} + R_{consumption})/2$ jusqu'à l'étape 6000, avant d'introduire le confort de l'agent :
$(R_{equity} + R_{consumption} + R_{comfort})/3$. Dont 10 en utilisant l'algorithme *Q-SOM* et 10 en utilisant l'algorithme *MADDPG* bien évidemment.

Le choix de ces fonctions de récompense se justifie par le fait que la fonction $R_{adaptability2}$ est considérée dans les travaux de Chaput & al. sur l'apprentissage adaptatif de comportements éthiques [3] comme la fonction dont les résultats sont les plus intéressants à étudier. Là où la fonction R_{equity} est normalement une des plus simples à apprendre pour les modèles testés dans le simulateur au fil des travaux.

Résultats

Les résultats de cette étude sont observables à l'annexe B (7). Pour chaque métrique, nous avons pris la moyenne des résultats obtenus sur les 10 expériences concernées.

En premier lieu, il nous est clair que les résultats sont décevants peu importe que ce soit dans le cas de $R_{Adaptability2}$ ou R_{Equity}.

Pour $R_{Adaptability2}$ Nous pouvons voir sur la figure (11) que les récompenses par étape lors de l'utilisation de *MADDPG* connaissent une énorme chute passée la 2000ème étape, signe que l'algorithme d'apprentissage peine à s'adapter ; là où *Q-SOM* s'en sort à merveille et trace doucement son chemin vers une convergence. Le graphe de la récompense cumulée moyenne (12) confirme cela et affiche encore plus nettement l'échec d'adaptation et de stabilisation de *MADDPG*.

Dans le cas de R_{Equity} (13 & 14), comme prévu l'algorithme *Q-SOM*, qui a été amélioré depuis [3], démontre d'excellents résultats. L'algorithme *MADDPG* montre des résultats convaincants, car il améliore sa performance au fur et à mesure d'une expérience.

Également, en matière de score final (moyenne absolue des récompenses), la disparité des résultats dans le cas de *MADDPG* est flagrante (15) et les résultats restent nettement inférieurs qu'avec *Q-SOM* même dans le cas de R_{Equity} (16).

7 Conclusion

Les résultats obtenus lors des expérimentations avec *MADDPG* ne sont pas encourageants pour la suite. Mais ce n'est pas si surprenant qu'un modèle d'apprentissage profond ne réussisse pas à ne serait-ce faire qu'aussi bien qu'un modèle développé sur-mesure. Les modèles d'apprentissage profond sont réputés pour être difficile à "*tuner*", il faudrait donc passer énormément de temps en plus à faire jouer les hyperparamètres de l'algorithme et essayer de trouver une version de ceux-ci plus adéquate.

Il faut donc apporter un soin particulier si nous désirons utiliser des techniques d'apprentissage profond et essayer au maximum de faire varier les

tentatives d'implémentations de celles-ci pour espérer obtenir des résultats convaincants.

Tout un questionnement se pose également quant au biais d'auto-confirmation dont pourrait souffrir le simulateur ; peut être que le modèle mis en place s'autofavorise et ne permet pas une modularité du mécanisme d'apprentissage ?

L'une des pistes possibles pour l'avenir du projet est la restructuration de l'architecture du simulateur afin de permettre des fonctionnements relativement différents, ouvrant la porte à l'expérimentation avec des approches beaucoup plus complexes et plus propices au réalisme et à la diversité des contextes éthiques.

Par exemple CM3 [29], GCRL [30] ou *Mean Field RL* [31]. Avec ces approches-ci, nous pourrions beaucoup plus traiter un cas de *Smart Grid* vraisemblable où les bâtiments (agents) ont non seulement des profils de consommation différents, mais également des fonctions de récompenses différentes selon leurs priorités morales, exposant ainsi des problématiques éthiques plus riches.

Nous pourrions rajouter un mécanisme de communication entre les agents permettant d'intégrer un module supplémentaire à activer pour faciliter la transmission d'informations, mais surtout la rendre plus semblable au graphe d'interconnexions que sont les réseaux numériques réels ; facilitant par la même occasion la coopération au sein du système multi-agents.

En bref, il y a une multitude d'options pour l'évolution de ce simulateur. Et du point de vue de ces travaux, la richesse des technologies d'apprentissage par renforcement profond est à disposition de notre imagination et des modules que nous aimerions développer et rajouter au projet. L'amélioration de l'apprentissage doit passer par l'amélioration de la structure du simulateur (par exemple en le passant sous OpenAI Gym). Parce qu'en l'état, un algorithme maison aussi abouti que *Q-SOM* suffit à répondre convenablement aux besoins actuels du simulateur.

Remerciements

Ce travail a été financé par la région Auvergne-Rhône Alpes, dans le cadre du projet Ethics.AI.

Je tiens également à remercier Salima Hassas et Rémy Chaput pour leur soutien tout au long de ce stage, Roxane B. pour ses encouragements dans ce contexte socio-sanitaire plus que trouble, ainsi que ma famille qui ne quitte jamais mes pensées.

Références

1. Allen, C., Smit, I. & Wallach, W. Artificial Morality : Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology* **7,** 149-155. issn : 1388-1957, 1572-8439. http://link.springer.com/10.1007/s10676-006-0004-4 (2021) (sept. 2005).
2. Cointe, N., Bonnet, G. & Boissier, O. Jugement éthique dans les systèmes multi-agents, 10.
3. Chaput, R. *et al.* Apprentissage adaptatif de comportements éthiques, 10.
4. Guerin, F. Constructivism in AI : Prospects, Progress and Challenges, 8.
5. Ethique et agents autonomes, ETHICAA, 49.
6. Duval, J. Jugement de l'éthique de comportement pour l'apprentissage, 36.
7. Anderson, M., Anderson, S. L. & Berenz, V. A Value Driven Agent : Instantiation of a Case-Supported Principle-Based Behavior Paradigm, 8.
8. Dorri, A., Kanhere, S. S. & Jurdak, R. Multi-Agent Systems : A Survey. *IEEE Access* **6,** 28573-28593. issn : 2169-3536. https://ieeexplore.ieee.org/document/8352646/ (2021) (2018).
9. Sutton, R. S. & Barto, A. G. Reinforcement Learning : An Introduction, 352.
10. Arponen, V. P. J. The extent of cognitivism. *History of the Human Sciences* **26,** 3-21. issn : 0952-6951, 1461-720X. http://journals.sagepub.com/doi/10.1177/0952695113500778 (2021) (déc. 2013).
11. Bai, H., Miao, S., Ran, X. & Ye, C. Optimal Dispatch Strategy of a Virtual Power Plant Containing Battery Switch Stations in a Unified Electricity Market. *Energies* **8,** 2268-2289. issn : 1996-1073. http://www.mdpi.com/1996-1073/8/3/2268 (2021) (23 mar. 2015).
12. Allen, C., Wallach, W. & Smit, I. Why Machine Ethics ? *IEEE Intelligent Systems* **21,** 12-17. issn : 1541-1672. http://ieeexplore.ieee.org/document/1667947/ (2021) (juil. 2006).
13. Arkin, R. *Governing Lethal Behavior in Autonomous Robots* isbn : 9780429150227 (mai 2009).
14. Moor, J. Four Kinds of Ethical Robots. *Philosophy Now* **72,** 12-14 (2009).
15. Hernandez-Leal, P., Kaisers, M., Baarslag, T. & de Cote, E. M. A Survey of Learning in Multiagent Environments : Dealing with Non-Stationarity. *CoRR* **abs/1707.09183.** arXiv : 1707.09183. http://arxiv.org/abs/1707.09183 (2017).
16. Matignon, L., Laurent, G. & Fort-Piat, N. Independent reinforcement learners in cooperative Markov games : A survey regarding

coordination problems. *The Knowledge Engineering Review* **27,** 1-31 (mar. 2012).

17. Papoudakis, G., Christianos, F., Rahman, A. & Albrecht, S. V. Dealing with Non-Stationarity in Multi-Agent Deep Reinforcement Learning. *CoRR* **abs/1906.04737.** arXiv : 1906.04737. http://arxiv.org/abs/1906.04737 (2019).
18. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Machine Learning* **8,** 279-292. issn : 1573-0565. https://doi.org/10.1007/BF00992698 (mai 1992).
19. Lauer, M. & Riedmiller, M. An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems, 535-542 (2000).
20. Matignon, L., Laurent, G. & Fort-Piat, N. Hysteretic Q-learning : an algorithm for Decentralized Reinforcement Learning in Cooperative Multi-Agent Teams. *IEEE International Conference on Intelligent Robots and Systems,* 64-69 (déc. 2007).
21. Bowling, M. & Veloso, M. Multiagent learning using a variable learning rate. *Artificial Intelligence* **136,** 215-250. issn : 0004-3702. https://www.sciencedirect.com/science/article/pii/S0004370202001212 (2002).
22. Tesauro, G. Extending Q-Learning to General Adaptive Multi-Agent Systems. (Jan. 2003).
23. Konda, V. & Tsitsiklis, J. Actor-Critic Algorithms, 1008-1014 (2000).
24. Tampuu, A. *et al.* Multiagent cooperation and competition with deep reinforcement learning. *PLOS ONE* **12,** 1-15. https://doi.org/10.1371/journal.pone.0172395 (avr. 2017).
25. Lowe, R. *et al.* Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *CoRR* **abs/1706.02275.** arXiv : 1706.02275. http://arxiv.org/abs/1706.02275 (2017).
26. Ryu, H., Shin, H. & Park, J. *Multi-Agent Actor-Critic with Generative Cooperative Policy Network* oct. 2018.
27. Iqbal, S. & Sha, F. *Actor-Attention-Critic for Multi-Agent Reinforcement Learning* in *Proceedings of the 36th International Conference on Machine Learning* (éd. Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, sept. 2019), 2961-2970. https://proceedings.mlr.press/v97/iqbal19a.html.
28. Foerster, J. N., Farquhar, G., Afouras, T., Nardelli, N. & Whiteson, S. Counterfactual Multi-Agent Policy Gradients. *CoRR* **abs/1705.08926.** arXiv : 1705.08926. http://arxiv.org/abs/1705.08926 (2017).
29. Yang, J., Nakhaei, A., Isele, D., Zha, H. & Fujimura, K. CM3 : Cooperative Multi-goal Multi-stage Multi-agent Reinforcement Lear-

ning. *CoRR* **abs/1809.05188.** arXiv : 1809.05188. http://arxiv.org/abs/1809.05188 (2018).

30. Jiang, J., Dun, C. & Lu, Z. Graph Convolutional Reinforcement Learning for Multi-Agent Cooperation. *CoRR* **abs/1810.09202.** arXiv : 1810.09202. http://arxiv.org/abs/1810.09202 (2018).
31. Yang, Y. *et al.* Mean Field Multi-Agent Reinforcement Learning. *CoRR* **abs/1802.05438.** arXiv : 1802.05438. http://arxiv.org/abs/1802.05438 (2018).
32. Sunehag, P. *et al.* Value-Decomposition Networks For Cooperative Multi-Agent Learning. *CoRR* **abs/1706.05296.** arXiv : 1706.05296. http://arxiv.org/abs/1706.05296 (2017).
33. Rashid, T. *et al.* QMIX : Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *CoRR* **abs/1803.11485.** arXiv : 1803.11485. http://arxiv.org/abs/1803.11485 (2018).
34. Son, K., Kim, D., Kang, W. J., Hostallero, D. & Yi, Y. QTRAN : Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. *CoRR* **abs/1905.05408.** arXiv : 1905.05408. http://arxiv.org/abs/1905.05408 (2019).
35. Mguni, D. *et al. Learning to Shape Rewards using a Game of Switching Controls* mar. 2021.
36. Valcarcel Macua, S. *et al.* Diff-DAC : Distributed Actor-Critic for Multitask Deep Reinforcement Learning (oct. 2017).
37. Zhang, K., Yang, Z., Liu, H., Zhang, T. & Başar, T. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents (fév. 2018).
38. Singh, A., Jain, T. & Sukhbaatar, S. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks. *CoRR* **abs/1812.09755.** arXiv : 1812.09755. http://arxiv.org/abs/1812.09755 (2018).
39. Henriques, R. & Madeira, S. BicNET : Flexible module discovery in large-scale biological networks using biclustering. *Algorithms for Molecular Biology : AMB* **11** (2016).
40. Gao, X., Jin, S., Wen, C.-K. & Li, G. Y. ComNet : Combination of Deep Learning and Expert Knowledge in OFDM Receivers. arXiv : 1810.09082 [eess.SP] (2018).
41. Das, A. *et al. TarMAC : Targeted Multi-Agent Communication* in *Proceedings of the 36th International Conference on Machine Learning* (éd. Chaudhuri, K. & Salakhutdinov, R.) **97** (PMLR, sept. 2019), 1538-1546. https://proceedings.mlr.press/v97/das19a.html.
42. Jiang, J. & Lu, Z. *Learning Attentional Communication for Multi-Agent Cooperation* in *Proceedings of the 32nd International Conference*

on Neural Information Processing Systems (Curran Associates Inc., Montréal, Canada, 2018), 7265-7275.

43. Galassi, A., Lippi, M. & Torroni, P. Attention, please ! A Critical Review of Neural Attention Models in Natural Language Processing. *CoRR* **abs/1902.02181.** arXiv : 1902.02181. http://arxiv.org/abs/1902.02181 (2019).
44. Yang, X. An Overview of the Attention Mechanisms in Computer Vision. *Journal of Physics : Conference Series* **1693,** 012173. https://doi.org/10.1088/1742-6596/1693/1/012173 (déc. 2020).

Annexe A : Synthèse de l'état de l'art des algorithmes d'apprentissage par renforcement multi-agents

Synthèse de l'état de l'art des algorithmes de *MARL*

Colonnes du tableau :

- **Type** : Décrit si le modèle est un modèle d'apprentissage par renforcement ou d'apprentissage par renforcement profond.
- **Sous-type** : Décrit si le modèle est un modèle basé sur le Q-learning (QL, IQL), le Fully Observable Critic (AC-based ou Q/value-based), le Value Decomposition (VD), le Consensus ou la Communication.
- **Architecture** : Décrit si l'architecture est décentralisée (D), centralisée (C) ou distribuée (Di) pendant l'entrainement (colonne T) et pendant l'exécution (colonne E).
- **Nature de l'Env.** : Décrit si les environnements considérés sont coopératifs (Coop.), collaboratifs (Collab), compétitifs (Comp.) ou mixtes (Mixte).
- **OE** : Décrit l'observabilité des états des agents, c'est-à-dire si leur état est observable localement uniquement (L), ou par tous les autres agents donc globalement (G).
- **OA** : Décrit l'observabilité des actions des agents.
- **OR** : Décrit l'observabilité des récompenses des agents.
- **Particularités** : Décrit les innovations proposées par le modèle et ses particularités algorithmiques.
- **Problèmes traités** : Décrit les problèmes considérés lors de la conception du modèle tels que la non-stationnarité, l'alter-exploration, etc.
- **Défauts** : Décrit les points négatifs identifiés dont souffre le modèle tels que le manque de scalabilité, etc.
- **Cas d'applications notables** : Décrit les cadres d'expérimentation des recherches concernés et potentiellement la qualité des résultats obtenus.

Symboles particuliers :

- « - » : colonne non applicable
- « ? » : manque d'informations

Type	*Sous-type*	*Algorithme*	*Architecture*		*Nature de l'Env.*	*OE*		*OA*		*OR*		*Particularités*	*Problèmes traités*	*Défauts*	*Cas d'applications notables*
			T	*E*		*T*	*E*	*T*	*E*	*T*	*E*				
RL	**QL**	**Decentralized Q-learning (Watkins & Dayan, 1992)**	D	D	-	L	L	L	L	G	G	Q-learning décentralisé classique	Aucun en particulier	Ne converge pas vers des politiques optimales dans des environnements complexes	Succès relatif dans quelques cas d'applications
		Distributed Q-learning (Lauer & Riedmiller, 2000)	Di	Di	Coop.	G	G	L	L	G	G	Récompense jointe distribuée entre les agents. Agents optimistes : Mise à jour des Q-valeurs ssi la nouvelle évaluation est meilleure	Alter-exploration, Pareto-optimalité	Mauvais passage à l'échelle. Ne converge pas vers des politiques optimales dans les env. stochastiques	Conçu pour les jeux Markoviens coopératifs en env. déterministe mais applicable aux tâches coopératives
		Hysteretic Learners (Matignon & al., 2007)	D	D	-	G	G	G	L	G	G	Deux taux d'apprentissage avec optimisme, surestimation des Q-valeurs	Stochasticité	Paramétrage difficile	Expérimenté sur les Matrix Games
		WoLF-PHC (Bowling & Veloso, 2002)	D	D	Comp.	G	G	L	L	G	G	Heuristique avec deux taux d'apprentissage différents (selon si l'agent gagne ou perd)	Alter-exploration, Stochasticité	Risque de convergence vers un optimum local	Conçu pour les jeux stochastiques compétitifs mais applicables aux tâches coopératives
		Hyper-Q (Tesauro, 2003)	D	D	Comp.	G	G	G	G	G	G	La Q-fonction dépend de l'état, de l'estimation de la stratégie mixte des autres agents et la stratégie	Non-stationnarité	Passage à l'échelle très compliqué	Conçu spécifiquement pour les jeux stochastiques

												mixte actuelle de l'agent			
	?	**RL-CD (Da Silva et al., 2006; Hadoux et al., 2014a)**	D	D	?	?	?	G	G	L	L	Les environnements stationnaires sont des « contextes » pour lesquels un modèle partiel est appris	Non-stationnarité (partiellement)	?	Conçu pour les environnements qui changent de manière périodiques
Deep RL	**IQL**	**IQL + DQL (Tampuu & al., 2017)**	D	D	Comp. et Coop.	G	G	L	L	G	G	Intègre un DQN au niveau de chaque agent. Nécessite de stabiliser la Replay Memory avec un algorithme style Fingerprint	?	La non-stationnarité empêche la stabilisation de la Replay Memory	Expérimentations sur des jeux compétitifs et coopératifs avec deux agents
	Fully Obs. Critic (AC)	**MADDPG (Lowe & al., 2017)**	C	D	Comp et Coop.	G	L	G	L	L	L	Amélioration de DDPG proposant l'entraînement d'un Critic centralisé par agent.	Non-stationnarité	Passage à l'échelle de l'entraînement difficile	Très bons résultats vs DDPG dans des environnement coopératifs et compétitifs (communication, navigation, predator-prey…)

		MADDPG – GCPN (Generative Cooperative Policy Network) (Ryu & al., 2018)	D	D	Coop.	G	L	G	L	G ou L	G or L	Chaque agent apprend une politique déterministe. Un Actor Network supplémentaires génère des samples d'actions des autres agents qui sont passés en input au Critic.	Bonne adaptabilité aux actions des autres agents	Probablement assez coûteux à mettre en œuvre et à passer à l'échelle. Néanmoins des extensions existent pour des problèmes à N agents	Résultats convaincants sur des simulations de contrôleurs de systèmes de stockages d'énergie
		Actor-Attention -Critic for MARL (Iqbal & al., 2019)	C	D	Coop.	G	L	G	L	L	L	Ajout d'une couche d'attention au réseau de Critic pour le calcul des Q-valeurs	Non-stationnarité, Pareto-optimalité, Scalabilité et Robustesse	Modèle non étendue aux espaces d'actions discrets et les environnement compétitifs	Résultats très convaincants dans des env. de communication coopérative, de predator-prey et de routage de paquets
		COMA (COunter factual Multi Agent policy gradient) (Foerster & al., 2018)	C	D	Coop.	G	L	G	L	G	G	Utilise des Actors entraînés localement et un seul Critic centralisé. Utilise une méthode appelée Counterfactual Baseline pour calculer l'action d'un agent.	Stochasticité, Optimalité locale de la solution	N'adresse pas les espaces d'actions continus en environnement compétitif. L'exploration des agents pose problème pour la scalabilité	Résultats meilleurs qu'une grande partie de l'état de l'art des modèles AC sur StarCraft

		CM3: Cooperative Multi-goal Multi-stage Multi-agent RL (Yang & al., 2018)	C	D	Coop.	L	L	L	L	L	L	Les agents sont d'abord entraînés sur leurs objectifs personnels puis sur l'objectif commun	Non-stationnarité, Multi-Agent credit assignment	?	Résultats convaincants comparés à COMA et Independent Actor-Critic pour la négociation des voitures autonomes
		A3C Extension for Multi-Agent RL (Sartoretti & al., 2019)	C	D	Coop.	G	G	L	L	L	L	Pendant l'entraînement, les agents partagent les poids (comme dans A3C) pour les modèles d'Actor et de Critic ; et entraînent un agent de manière asynchrone. Durant l'execution chaque agent exécute une copie de la politique apprise	Non-stationnarité, Alter-exploration	Adaptabilité du modèle à des tâches et des environnements plus complexes à prouver	Résultats convaincants dans des tâches de planification de construction collaborative (TERMES)

												sans communication avec les autres ; et chaque agent voit les autres agents comme une partie de l'environnement			
	Fully Obs. Critic (Q)	**GCRL (Graph Convolutional Reinforcement Learning) (Jiang & al. 2018)**	C	D	Coop.	L	L	L	L	L	L	Le SMA est traité comme un graphe où chaque agent partage ses observations à ses voisins	Non-stationnarité	?	Expérimentations sur des environnements coopératifs en grille
	Fully Obs. Critic (Q ou AC)	**Mean Field Reinforcement Learning (Yang & al., 2018)**	C	D	Coop. et Mixte	G	G	G	L	G	L	Chaque agent prend en compte l'effet moyen de son voisinage au lieu d'une communication exhaustive entre tous les agents	Non-stationnarité, Pareto-optimalité, Scalabilité	?	Très bons résultats dans trois scenarios : Gaussian Squeeze task, Ising Model, Battle Game
	VD	**VDN (Sunehag & al., 2017)**	-	-	Coop.	L	L	G	G	G	G	Apprend une Q-tot globale calculée par additivité à décompenser en N Q-fonctions pour N agents	Stochasticité, Non-stationnarité, Multi-Agent credit assignment	Ne s'applique que dans le cadre d'une récompense unique partagée. Mauvaise scalabilité ?	Bons résultats dans des environnements de tâches 2D avec deux agents
		QMIX (Rashid & al., 2018)	C	D	Coop.	G	L	G	L	G	G	QMIX peut représenter une classe plus riche de fonctions action-valeur que VDN. Utilise des réseaux d'agents Qa et un réseau	Non-stationnarité, Multi-Agent credit assignment	Mauvaise scalabilité ?	Excellents résulats sur StarCraft

											de mix pour les combiner. Utilise un DRQN pour l'apprentissage de chaque Qi			
		QTRAN (Kyunghwan Son & al., 2019)	C	D	Coop.	L	L	L	L	G	G	Considère une action-valeur jointe Q'tot alternative, factorisable par décomposition additive. Utilise trois réseaux : un pour les Qi individuelles, un pour Qtot, et un pour gérer le régulateur joint Vtot (qui stabilise la différence entre Qtot et Q'tot)	Non-stationnarité, Multi-Agent credit assignment, Scalabilité relativement traitée. Gestion d'une plus grande variété de tâches et meilleurs résultats dans des contextes non-monotoniques que ses prédécesseurs	?	Résultats significativement meilleurs que QMIX et VDN dans des environnements de Multi-domain Gaussian Squeeze et de Predator-Prey modifié
		Meta-Agent reward shaping (Mguni & al., 2018)	?	D	Non-Coop.	L	L	L	L	G	G	Introduit un méta-agent qui modifie les récompenses des autres agents, entraîné par un Actor-Critic	Non-stationnarité, Convergence vers un équilibre de Nash. Scalabilité ?	Entraînement du méta-agent par un algo gradient-based très coûteux. Convergence et scalabilité à vérifier sur des tâches plus complexes	Résultats convaincants sur une tâche de réorganisation de 2000 agents où leur localisation désirée change avec le temps

	Con-sen-sus	**Diff-DAC (Diffusion-based Distributed Actor-Critic) (Macua & al., 2020)**	Di	Di	Coop. /Indé.	L	L	L	L	L	L	Limite le voisinage de communication mais implique l'existence d'un chemin entre chaque agent. Entraîne plusieurs agents en parallèle sur des tâches différents et/ou similaires pour trouver une politique moyenne qui performe bien sur toutes les tâches et donc obtenir un consensus	Bonne scalabilité. Traite le cas du multi-tâches.	La politique moyenne peut performer très bien sur certaines tâches mais mal sur d'autres	Expérimentations réalisées dans des environnements Cart-Pole différents
		Networked Agents for Fully Decentralized MARL (Zhang & al., 2018)	D	D	Collab.	G	G	G	L	L	L	Approches AC-Based. Les agents font des choix basés sur leurs observations et sur les messages reçus de la part de leurs voisins.	Scalabilité et vie privée mise au cœur de la proposition		Expérimentations dans des tâches de navigation coopérative et d'allocation distribuée de ressources

												Décliné en variantes pour les espaces d'états et d'actions discrets et continus			
	Learn 2 Com-mu-ni-cate	**CommNet**													
		CommNet Ext.													
		BiCNet													
		IC3Net													
		TarMAC													
		ATOC													

Annexe B : Résultats de l'étude comparative entre MADDPG et Q-SOM

Annexe B - 1

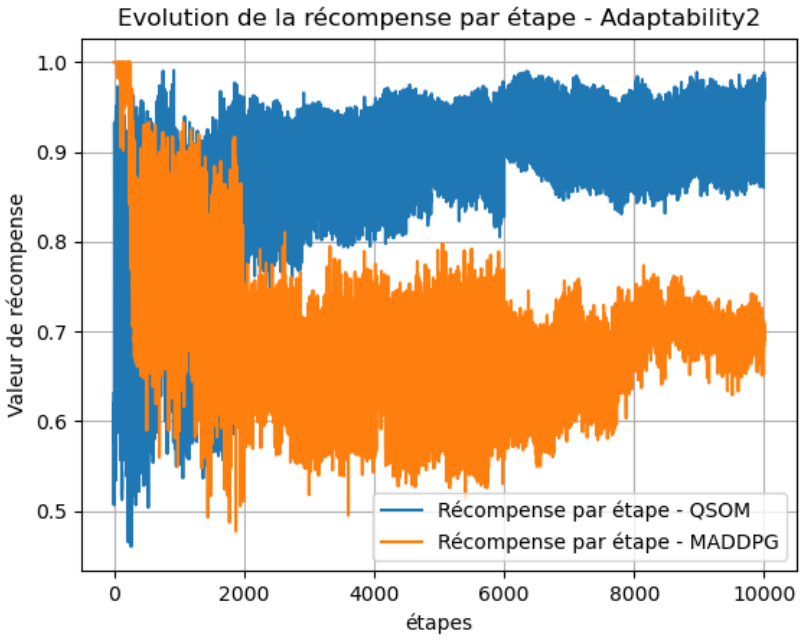

FIGURE 11.

Annexe B - 2

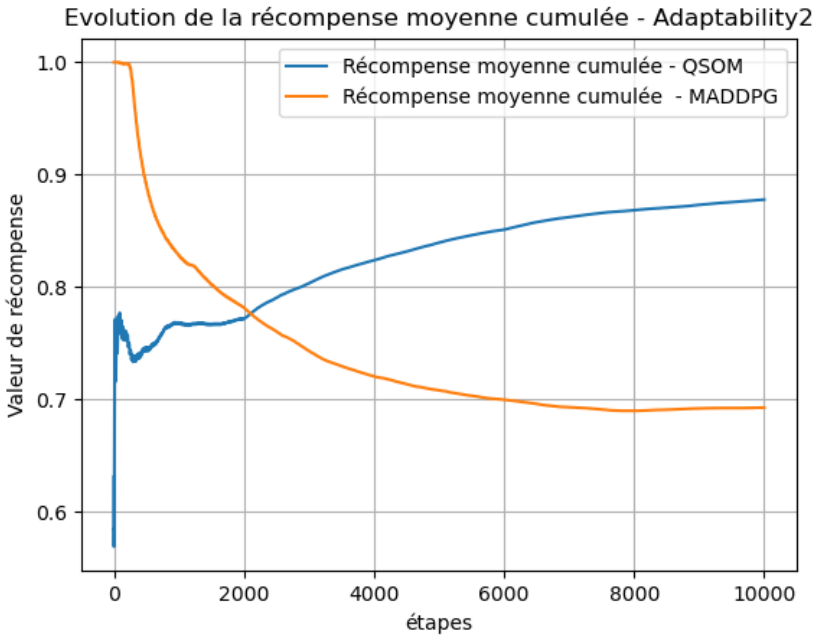

FIGURE 12.

Annexe B - 3

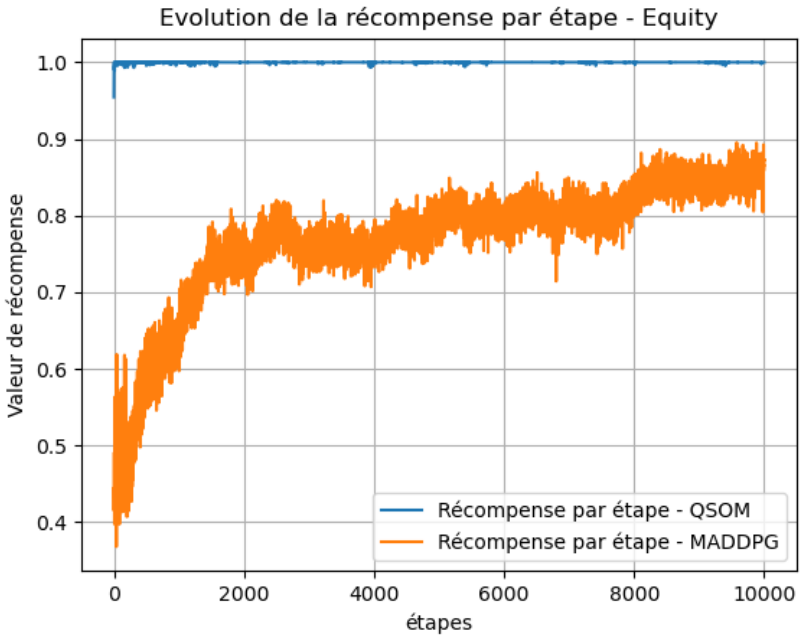

FIGURE 13.

Annexe B - 4

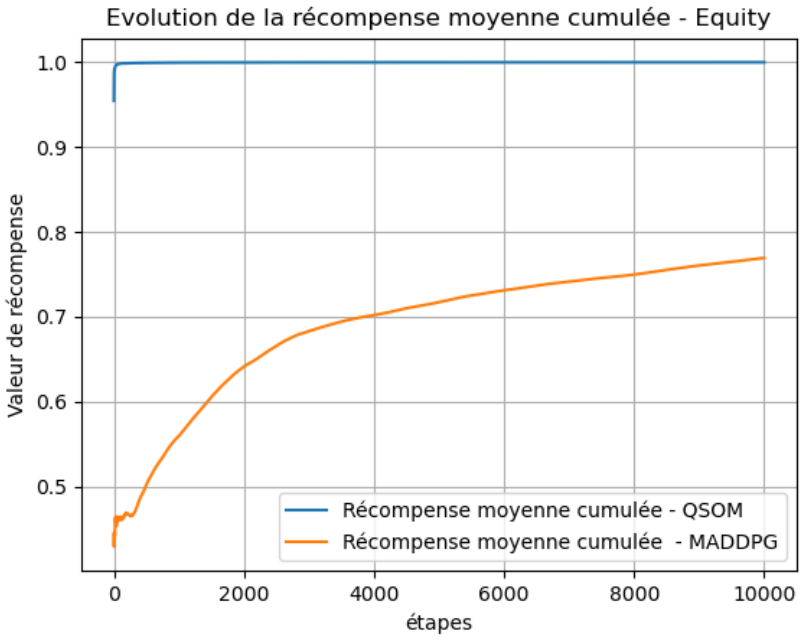

FIGURE 14.

Annexe B - 5

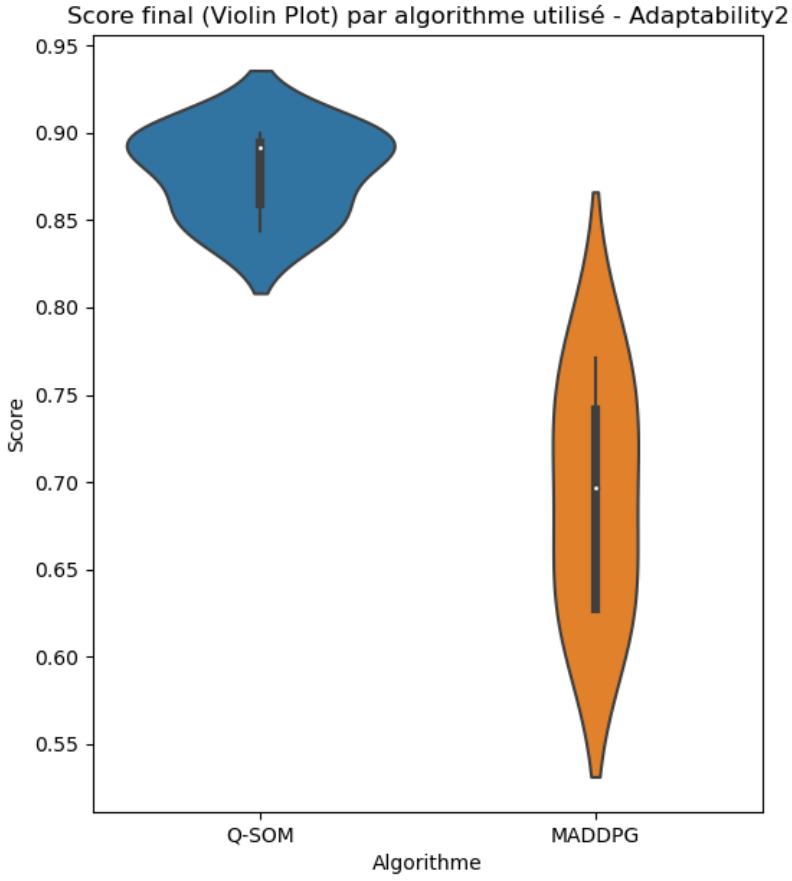

FIGURE 15.

Annexe B - 6

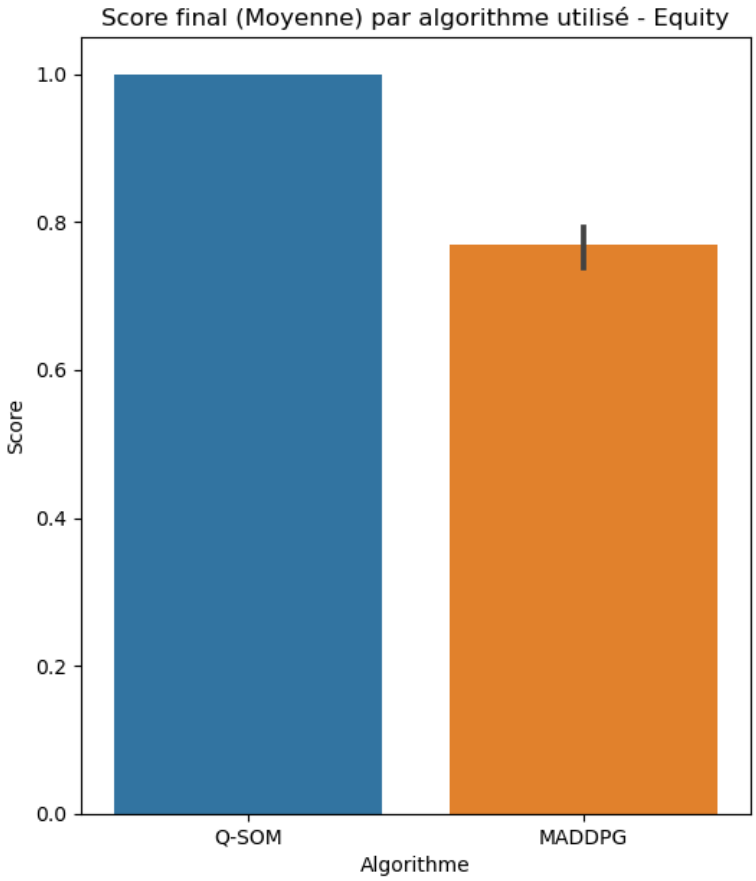

Figure 16.