Approche multi-agent combinant raisonnement et apprentissage pour un comportement éthique [1]

R. Chaput[a] remy.chaput@univ-lyon1.fr
J. Duval jeremy-duval1@hotmail.fr
O. Boissier[b] Olivier.Boissier@emse.fr
M. Guillermin[c] mguillermin@univ-catholyon.fr
S. Hassas[a] salima.hassas@univ-lyon1.fr

[a]Univ. Lyon, Université Lyon 1, LIRIS, UMR5205, F-69622, LYON, France

[b]Institut Henri Fayol,
MINES Saint-Etienne, France

[c]Sciences and Humanities Confluence Research Center, Lyon Catholic University

Résumé

Le besoin d'incorporer des considérations éthiques au sein d'algorithmes d'Intelligence Artificielle est de plus en plus présent. Combinant raisonnement et apprentissage, cet article propose une méthode hybride, où des agents juges évaluent l'éthique du comportement d'agents apprenants. Le but est d'améliorer l'éthique de leur comportement dans des environnements dynamiques multi-agents. Plusieurs avantages découlent de cette séparation : possibilité de co-construction entre agents et humains ; agents juges plus accessibles pour des humains non-experts ; adoption de plusieurs points de vue pour juger un même agent, produisant une récompense plus riche. Les expérimentations sur la distribution de l'énergie dans un simulateur de Smart Grid montrent la capacité des agents apprenants à se conformer aux règles des agents juges, y compris lorsque les règles évoluent.

Mots-clés : *Éthique, Machine Ethics, Apprentissage Multi-Agent, Apprentissage par Renforcement, Hybride Neural-Symbolique, Jugement Éthique*

Abstract

The need to imbue Artificial Intelligence algorithms with ethical considerations is more and more present. Combining reasoning and learning, this paper proposes a hybrid method, where judging agents evaluate the ethics of learning agents' behavior. The aim is to improve the ethics of their behavior in dynamic multi-agent environements. Several advantages ensue from this separation : possibility of co-construction between agents and humans ; judging agents more accessible for non-experts humans ; adoption of several points of view to judge the same agent, producing a richer feedback. Experiments on energy distribution inside a Smart Grid simulator show the learning agents' ability to comply with judging agents' rules, including when they evolve.

Keywords: *Ethics, Machine Ethics, Multi-Agent Learning, Reinforcement Learning, Hybrid Neural-Symbolic Learning, Ethical Judgment*

1 Introduction

Alors que le nombre d'applications utilisant des modèles d'Intelligence Artificielle (IA) augmente, il y a un débat de société et de recherche au sujet du moyen d'introduire des capacités éthiques dans ces modèles.

Le domaine des *Machine Ethics* s'intéresse à la conception d'"agents à impact éthique" [18] ayant un impact éthique sur des vies humaines ; en particulier, l'"éthique dans la conception" [15] vise à ce que ces agents prennent des décisions selon des considérations éthiques, ce que nous appellerons un "comportement éthique".

Cette demande de capacités éthiques pour des agents autonomes artificiels a été largement documentée [21]. Toutefois les moyens d'implémenter ces compétences ne sont pas clairs : certains travaux proposent des approches descendantes par raisonnement symbolique, tandis que d'autres préfèrent utiliser des approches ascendantes par apprentissage [1]. Les deux approches offrent différents avantages mais ont également des inconvénients ; ainsi, dans cet article, nous présentons une nouvelle approche, hybride [15], avec un apprentissage de comportements guidé par des récompenses issues de raisonnements symboliques.

[*]Cet article est basé sur la traduction d'un article soumis à AIES2021.

Cet article est structuré comme suit : nous présentons d'abord la littérature sur laquelle nous appuyons notre approche hybride ; cette proposition est ensuite détaillée ; des expérimentations sur un cas d'application des *Smart Grids* et leurs résultats démontrent la faisabilité ; finalement, la dernière section compare l'approche à la littérature, examine les limitations actuelles et présente des perspectives.

2 Fondements

Afin d'identifier les principes de conception qui sous-tendent notre approche, nous explorons d'abord la littérature des *Machine Ethics*. Nous considérons ensuite le champ de l'IA hybride (Neural-Symbolique), qui combine les méthodes symboliques et d'apprentissage.

2.1 Éthique et IA

La plupart des travaux existants en *Machine Ethics* se focalisent sur un unique agent isolé dans son environnement [26]. Nous arguons qu'il est important de considérer plusieurs agents en interaction dans un environnement commun, comme il s'agit d'une situation plus réaliste, qui soulève le problème de la confrontation de plusieurs éthiques.

Comme en IA, les travaux en *Machine Ethics* sont divisés en trois catégories [1] : approches descendantes, ascendantes et hybrides.

Les approches descendantes s'intéressent à la formalisation de principes éthiques, tel que l'Impératif Catégorique de Kant. En utilisant un raisonnement logique sur des représentations symboliques, de telles approches peuvent s'appuyer sur des connaissances expertes et offrir une meilleure lisibilité et explication des décisions prises. Par exemple, l'*Ethical Governor* [3] vérifie l'adéquation des actions avec des règles préétablies comme les Règles d'Engagement ou le Droit de la guerre. Dans *Ethicaa* [13], des agents raisonnent sur plusieurs principes éthiques pour décider de leur comportement et juger les actions des autres agents. Cependant, ces approches descendantes, du fait de leur corpus de connaissances explicite mais figé, ne peuvent pas s'adapter, sans reconception, à des situations non prévues ou à une évolution de l'éthique.

Les approches ascendantes cherchent à apprendre un comportement à partir d'un jeu de données, e.g., des exemples étiquetés ou des expériences obtenues par interactions. Par exemple, *GenEth* [2] utilise des décisions d'éthicistes dans de multiples contextes pour apprendre un principe éthique ; une autre approche utilise l'apprentissage par renforcement (RL) en ajoutant à la récompense de la tâche une composante éthique sous forme de différence entre les comportements de l'agent et d'un humain moyen, supposé exhiber des considérations éthiques [25]. Ces approches, bien qu'utilisant de l'apprentissage, n'ont pas considéré la question de l'adaptation sur le long-terme en réponse à des situations pouvant évoluer. De plus, les approches ascendantes sont plus difficiles à interpréter que les approches descendantes.

Finalement, les approches hybrides couplent les approches descendantes et ascendantes, de telle sorte que les agents puissent apprendre des comportements éthiques par expérience tout en étant guidés par un cadre éthique existant afin de forcer des contraintes et les empêcher de diverger. Pour plus de détails, le lecteur peut se référer à [1, 15].

Nous discutons des différents moyens d'IA hybride Neural-Symbolique et comment les intégrer dans un agent éthique dans la prochaine section.

2.2 Approches hybrides

Les approches hybrides en IA visent à coupler le raisonnement symbolique avec l'apprentissage numérique pour bénéficier des avantages des deux approches en réduisant leurs inconvénients. Plusieurs manières pour les intégrer existent, voir par exemple [9]. Les auteurs avancent que les plans dans un agent BDI sont plus faciles à expliquer à un humain ; il est aussi admis qu'il est plus facile d'introduire des connaissances, par exemple d'un expert du domaine non-développeur, avec des règles symboliques. Des exemples d'approches hybrides incluent SOAR-RL [19] ou BDI-RL [10], qui intègrent des algorithmes d'apprentissage par renforcement avec du raisonnement. Plusieurs travaux ajoutent une couche de raisonnement symbolique, souvent BDI, par-dessus un agent artificiel [3, 11], et sont souvent qualifiés d'hybride.

Le projet Ethicaa propose un système multi-agent dans lequel les agents juges déterminent un jugement sur les actions d'autres agents, en utilisant des croyances sur une situation donnée [13]. À notre connaissance, l'intégration d'un jugement symbolique pour donner une récompense numérique aux agents apprenants n'a pas

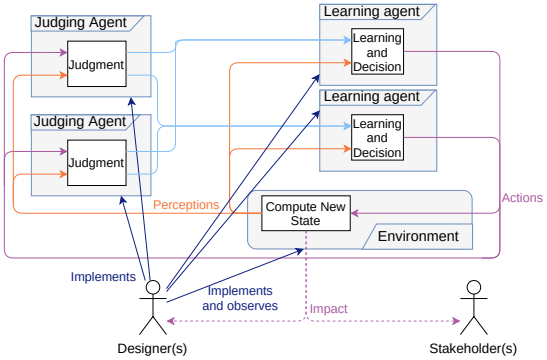

FIGURE 1 – Architecture de notre approche, comprenant des concepteurs humains implémentant des règles pour juger les agents apprenants. Les actions des apprenants modifient un environnement partagé, ce qui impacte les humains.

été étudié dans le domaine des Machine Ethics, mais nous pouvons appliquer les travaux d'Ethicaa au jugement d'agents utilisant de l'apprentissage numérique.

3 Modèle

Dans cette section, nous décrivons notre proposition, basée sur les principes de conception retenus de la littérature.

3.1 Architecture Abstraite

Considérons un système multi-agent comprenant des humains et des agents artificiels, représenté dans la Figure 1. Les concepteurs créent un environnement partagé et des agents autonomes afin qu'ils accomplissent des tâches ; les actions effectuées pour ce faire vont impacter l'environnement partagé et les humains. Le but des concepteurs est d'intégrer des *considérations éthiques* dans ces agents, afin de contraindre leur "impact éthique", en accord avec un ensemble de *valeurs morales* sélectionnées par les concepteurs.

Nous nous focalisons sur une approche hybride et plus particulièrement sur une séparation du jugement et de l'apprentissage, afin de permettre une co-construction. Pour cela, nous introduisons des agents juges et apprenants séparés, qui pourront évoluer indépendamment, par exemple pour mettre à jour les règles des juges par les concepteurs humains, tandis que les apprenants adaptent leur comportement en accord avec ces nouvelles règles.

De plus, nous proposons de rendre le jugement plus riche en confrontant plusieurs valeurs morales, que nous représentons par des agents juges séparés afin de clarifier le modèle ; cela ouvre également la voie à des interactions plus complexes entre juges tels que des processus d'argumentation, et offre un moyen simple de changer les règles en ajoutant ou supprimant des agents juges. Ces agents juges sont basés sur les agents Ethicaa [13], manipulant un ensemble de règles morales pour raisonner et juger les actions des autres agents dans l'environnement. Les juges utilisent ces règles pour déterminer un jugement (e.g., "moral", "immoral"), qui sont transformés en récompenses pour les apprenants. Ceux-ci l'intègrent dans leur processus d'apprentissage pour apprendre à effectuer de meilleures actions. La pluralité d'apprenants permettra d'évaluer leur impact dans un environnement partagé, plutôt qu'un agent isolé.

Finalement, le comportement attendu peut évoluer au cours du temps, du fait des dynamiques de notre société, les agents devraient donc être capables de s'adapter à des règles changeantes. Les concepteurs humains observent les actions des agents apprenants dans l'environnement et rectifient ces comportements indirectement en ajustant les règles à la base du jugement des agents juges. Il est ainsi possible d'envisager une approche d'IA centrée sur l'humain, avec un contrôle humain dans la boucle, comme le préconise le rapport Européen HLEG AI [1].

Les valeurs et règles morales à la base des considérations éthiques sont clairement visibles dans ce modèle car choisies explicitement par les concepteurs. Nous faisons l'hypothèse que cela améliore l'intelligibilité du jugement et donc du comportement attendu, ce qui est reconnu comme important, notamment par le rapport du HLEG. Bien que ce point ne sera pas évalué dans les expérimentations, il sert en partie de motivation à notre approche ; les humains, y compris les utilisateurs, devraient pouvoir vérifier la compatibilité avec leurs propres principes éthiques.

3.2 Modèle Formel

Considérons l'ensemble J des agents *juges* : chaque agent $j \in J$ est associé à une seule valeur morale et un ensemble de règles morales permettant de décider si une action *supporte* ou *trahit* cette valeur morale (e.g., la Justice, l'Inclusivité, la Sécurité). Dans le second ensemble L d'agents *apprenants*, chaque agent

1. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

$l \in L$ apprend un comportement et effectue des actions dans l'environnement, en se basant sur l'évaluation F_l, agrégation des jugements $Jugement_j(l)$ des $j \in J$ sur le comportement de $l : \forall l \in L : F_l = \{\forall j \in J : Jugement_j(l)\}$.

Ce modèle d'apprentissage correspond conceptuellement à un jeu Markovien (ou *Stochastic Game*), une extension du Processus de Décision Markovien (MDP) à plusieurs agents. Formellement, il s'agit d'un n-uplet $(S, P, A_0, ..., A_n, T, R_0, ..., R_n)$. S est l'ensemble de tous les états possibles, sous forme de vecteurs de nombres réels (états multi-dimensionnels et continus) pour permettre des cas d'applications riches. P, de taille n, est l'ensemble de tous les agents ; les agents juges n'agissant pas dans l'environnement, il correspond à L, l'ensemble des agents apprenants. A_l est l'ensemble des actions possibles pour l'agent l, également sous forme de vecteurs de nombres réels (actions paramétrées). L'ensemble des actions jointes $A = A_0 \times ... \times A_n$ regroupe la combinaison des actions possibles des différents agents. T est la fonction de probabilité de transition, définie par $T : S \times A \times S \rightarrow [0, 1]$, i.e., $T(s, a, s')$ est la probabilité de passer de l'état s à s' en effectuant l'action a. R_l est la fonction de récompense pour l'agent l, définie par $R_l : S \times A_l \times S \rightarrow \mathbb{R}$, i.e., $R_l(s, a_l, s')$ est la récompense de l'agent l pour avoir effectué l'action a_l dans l'état s, résultant en l'état s'.

Les MDPs et jeux Markovien peuvent être résolus avec de l'apprentissage par renforcement [22], une méthode pour apprendre la probabilité $\pi(s, a)$ de sélectionner chaque action a dans chaque état s. Pour chacun des agents apprenants, le but de l'algorithme d'apprentissage est d'apprendre la stratégie optimale, qui maximise l'espérance des récompenses reçues.

Traditionnellement, la fonction de récompense est une fonction mathématique qui indique si l'action exécutée était bonne, i.e. un objectif à optimiser. Nous voulons utiliser le jugement symbolique calculé par les agents juges ; pour cela, la fonction de récompense agrège et transforme ces jugements en une valeur numérique. Nous décrivons d'abord les agents apprenants et le processus par lequel ils apprennent une stratégie optimale π, en mettant de côté les détails de R que nous décrirons ensuite.

3.3 Agents Apprenants

Les agents apprenants doivent apprendre comment sélectionner une action dans un état donné,

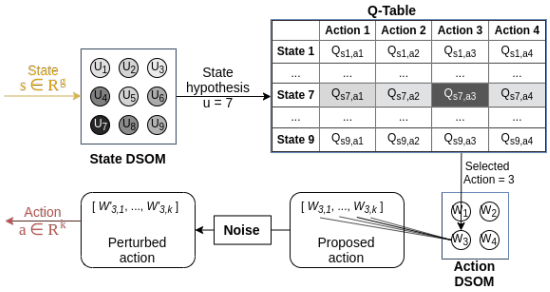

FIGURE 2 – Exemple de décision : l'agent reçoit un état, vecteur de réels, comparé à la State-DSOM. Le 7^{eme} neurone, qui a le vecteur prototype le plus proche, est choisi comme hypothèse d'état. À partir de la Q-Table et de ce 7^{eme} état, la 3^{eme} action est choisie. L'action obtenue est le résultat de la perturbation par un bruit aléatoire du vecteur associé au 3^{eme} neurone de l'Action-DSOM.

afin de maximiser la récompense reçue sur l'ensemble des pas de temps. Nous utilisons l'algorithme Q-DSOM pour sa capacité à manipuler des états et actions multi-dimensionnels et continus [12]. Cet algorithme utilise deux *Dynamic Self-Organizing Map* (DSOM) [20], inspiré des Cartes Auto-Organisatrices de Kohonen [17], afin d'apprendre les espaces d'états (State-DSOM) et d'actions (Action-DSOM).

Les neurones des deux DSOMs sont liés à une Q-Table [24], de telle sorte que chaque neurone corresponde à un état ou une action discrets, i.e., une ligne ou une colonne dans la Q-Table. La Q-Table permet d'apprendre l'intérêt, ou Q-Value, d'une paire état-action, afin que l'agent puisse choisir la meilleure action pour chaque état.

Les agents peuvent donc représenter n'importe quel état ou action multi-dimensionnel et continu comme un identifiant discret via les DSOMs et utilisent la Q-Table pour déterminer l'intérêt associé. Se reporter à l'article originel [12] pour une description détaillée de l'algorithme. Un exemple est présenté dans la Figure 2.

3.4 Agents Juges

L'architecture BDI des agents juges (cf. Figure 3) s'appuie sur [13] en simplifiant le mécanisme d'évaluation morale (des travaux futurs seront dédiés à enrichir cette composante).

À chaque pas de temps, les agents juges génèrent des croyances (B) à partir de leurs perceptions de l'environnement (les mêmes que les agents

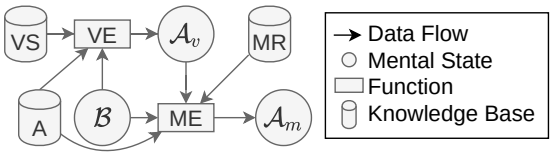

FIGURE 3 – Architecture des agents juges, adaptée d'Ethicaa [13].

apprenants) et des actions effectuées par ces apprenants. Les agents juges traitent de manière séparée chacun des composants des vecteurs réels, correspondant aux paramètres des actions. En d'autres termes, les agents juges reçoivent un ensemble d'actions $\{\forall i \in [[1, k]] : a_{l,i}\}$, tel que a_l est un vecteur de k composants et $a_{l,i}$ est le i-ème composant du vecteur, i.e., un nombre réel.

L'agent juge utilise un ensemble fixé de valeurs et des règles morales associées (VS et MR) pour déterminer si chaque composant de l'action trahit ou supporte la valeur. L'Évaluation Morale (ME) utilise les croyances générées (B) et les actions (A) pour produire une valuation morale (Am), parmi l'ensemble $V = \{moral, immoral, neutral\}$. À chaque pas de temps, le jugement par un agent juge j de l'action a_l d'un agent apprenant l est l'évaluation morale de chacun des k paramètres de cette action a_l : $Jugement_j(l) = \{i \in [[1, k]] : ME_j(B, a_{l,i})\}$. Chaque agent juge calcule un jugement différent pour chaque apprenant, de sorte que la fonction d'évaluation finale $F : L \rightarrow \left(V^k\right)^{|J|}$ retourne une liste de listes de valuations, soit $\forall l \in L : F_l = \{\forall j \in J : Jugement_j(l)\}$.

3.5 Jugement pour l'apprentissage

Dans cette section, nous faisons le lien entre les symboles (jugements et croyances) manipulés par les agents juges et les nombres réels (perceptions, actions, et récompenses) utilisés dans l'algorithme Q-DSOM.

À la fin de chaque étape de la simulation, après que les agents apprenants aient choisi leur action, les agents juges perçoivent plusieurs données et les transforment en croyances : liste des agents, les actions de chacun, les propriétés globales représentant l'état de l'environnement et les propriétés individuelles par agent.

Finalement, la fonction de récompense R_l : $S \times A_l \times S \rightarrow \mathbb{R}$ retourne un nombre réel. Plusieurs méthodes sont possibles pour transformer les valuations symboliques en une récompense numérique ; dans cette première étape, nous proposons la solution suivante. Nous transformons le jugement de chaque agent juge, i.e., une liste de valuations, en un nombre unique, en comptant le nombre de valuations *moral* et en divisant par le nombre de valuations *moral* et *immoral*, de telle sorte que les actions morales tendent vers 1, tandis que les actions immorales tendent vers 0. Comme cas spécial, si la liste ne consiste que de valuations *neutral*, nous considérons que l'action était ni bonne ni mauvaise, et nous la mettons à 0.5. La récompense finale est calculée comme la moyenne des récompenses de chaque agent juge.

On peut remarquer que cette méthode permet de résoudre de manière simple les conflits entre les agents juges ; par exemple, le premier agent juge peut juger que le premier composant de l'action est moral selon sa propre valeur morale, tandis qu'un second juge peut déterminer que ce même premier composant est immoral, en accord avec sa valeur morale (différente).

4 Expérimentations

Afin de tester la validité de notre approche, nous reprenons le cas d'application présenté dans [12] : il s'agit d'une micro-grille électrique hypothétique, dans laquelle la production d'énergie est décentralisée au lieu de reposer uniquement sur le réseau national. La grille possède une source d'électricité principale (e.g., une station hydro-électrique, ou une ferme à éoliennes) ; les utilisateurs, ou *prosumers* (producteurs-consommateurs), peuvent eux-mêmes produire une petite quantité d'énergie (e.g., via des panneaux photovoltaïques). Considérant la difficulté de stocker de grandes quantités d'énergie sur une longue période, et que la production et la demande peuvent fluctuer sur de courtes périodes, les *prosumers* peuvent échanger de l'énergie afin de ne pas la gaspiller. De tels échanges supposent une forme de coopération pour éviter les situations d'inégalité ; de manière similaire, quand la source principale est trop sollicitée, les *prosumers* doivent réduire leur consommation temporairement, et ainsi réduire leur confort, afin d'éviter des coupures.

Les simulations considèrent un ensemble de bâtiments (type habitations, bureaux et écoles ; voir la Figure 4) : la gestion de l'énergie de chaque bâtiment est prise en charge par un agent apprenant. Il doit apprendre à consommer et échanger de l'énergie pour satisfaire le besoin en confort de ses occupants, tout en considérant les intérêts

des autres *prosumers* de la grille.

Nous considérons ce simulateur simplifié et l'opposition d'intérêts entre les différents participants comme étant suffisamment plausibles et un cadre intéressant pour des comportements éthiques.

4.1 Règles et Valeurs Morales

Nous avons choisi des valeurs morales à partir de la littérature des réseaux électriques intelligents [4] et traduit ces valeurs afin de refléter le point de vue de citoyens participant à un tel système, et prenant des décisions pour allouer de l'énergie. En effet, les agents apprenants agissent en tant que mandataires pour ces *prosumers* et doivent donc soutenir les mêmes valeurs morales. Nous proposons quatre valeurs morales et les règles associées comme références communes pour tous les agents de la simulation : MR1 — Assurance de confort : une action permettant à un *prosumer* d'améliorer son confort est morale ; MR2 — Affordabilité : une action qui coûte trop cher à un *prosumer*, par rapport au budget configuré par l'utilisateur, est immorale ; MR3 — Inclusion sociale : une action qui améliore l'équité des conforts entre les *prosumers* est morale ; MR4 — Viabilité Environnementale : une action qui limite les échanges avec le réseau national est morale.

Ces règles sont partiellement en conflit : par exemple, les agents consomment de l'énergie pour satisfaire le confort des utilisateurs, en accord avec MR1, mais il n'y a pas assez d'énergie pour satisfaire tous les agents, ce qui trahit la valeur associée à MR3 ; acheter de l'énergie transgresserait MR4. En d'autres termes, chaque action implique une transgression d'au moins une règle morale, ce qui classe cette simulation comme un dilemme éthique selon la définition de Bonnemains [6].

Puisque les actions a_l sont des vecteurs de réels, le but des agents apprenants est de déterminer les bons paramètres, i.e., les composants du vecteur, afin de minimiser leur regret. En d'autres termes, la question pourrait être, par exemple, "Quelle quantité d'énergie devrais-je acheter afin de minimiser la transgression de MR4 tout en me permettant d'améliorer mon confort, en accord avec MR1 ?".

4.2 Simulateur

Le simulateur que nous utilisons est illustré dans la Figure 4 ; nous résumons ses composants ci-après.

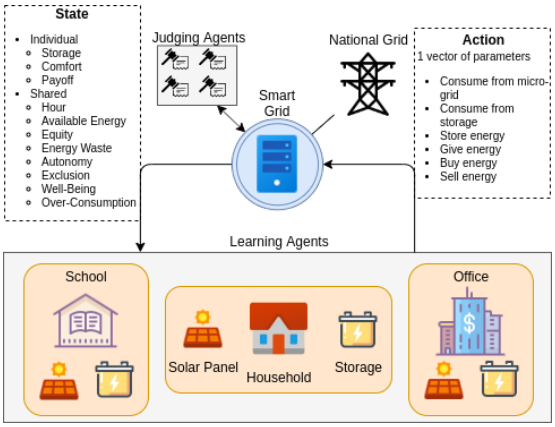

FIGURE 4 – Schéma du simulateur de *Smart Grid*. Une micro-grille, liée à la grille nationale, contient plusieurs agents représentant des bâtiments.

Agents Apprenants. Trois profils d'agents ont été créés pour répondre aux types de bâtiment et introduire de la variété entre les agents : Habitation, Bureau et École. Chaque profil est constitué : d'un profil de consommation, i.e. la quantité d'énergie dont il a besoin à chaque heure ; d'une courbe de confort, i.e. la fonction qui calcule son confort pour une consommation et un besoin donnés ; d'une capacité d'action, e.g. la quantité maximale d'énergie qu'il peut consommer ; et d'une capacité de stockage personnelle.

Nous utilisons un jeu de données publiques de consommation d'énergie [2] comme source des profils de consommation. Trois bâtiments ont été sélectionnés : *Residential*, *Small Office* et *Primary School* ; chacun dans la même ville (Anchorage) afin de minimiser le risque de biais dans la consommation recensée, par exemple à cause d'une différence de température qui nécessiterait plus de chauffage dans une ville par rapport à une autre. Le jeu de données contient la charge horaire, i.e. la quantité d'énergie consommée par un bâtiment pour chaque heure sur une année ; tandis que le travail précédent utilisait un profil moyenné sur tous les jours de l'année (profil journalier), nous avons retenu le profil annuel complet. Les courbes de confort et le besoin en énergie par heure utilisés dans nos expérimentations sont visibles dans la Figure 5.

2. https://openei.org/datasets/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states

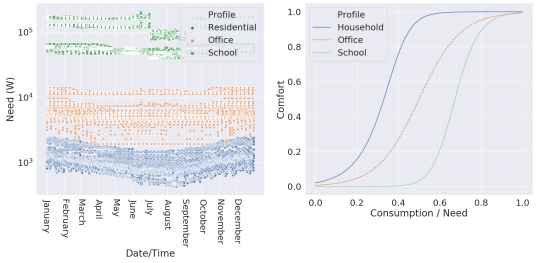

FIGURE 5 – Le besoin et le confort pour chaque profil d'agent.

Actions. À chaque pas de temps, chaque agent apprenant effectue une action, vecteur de paramètres suivants : la quantité d'énergie consommée depuis la micro-grille, la quantité consommée depuis la batterie personnelle, l'énergie stockée dans sa batterie depuis la grille et inversement la quantité donnée de la batterie vers la micro-grille. Si la grille ne dispose pas d'assez d'énergie, l'excès est automatiquement acheté depuis la grille nationale pour éviter une coupure, mais dans ce cas l'excès est considéré comme sur-consommé. L'agent peut également interagir avec la grille nationale en achetant ou vendant de l'énergie dans sa batterie personnelle.

Perceptions. Afin de prendre une décision, l'agent obtient l'état de l'environnement représenté par un vecteur de nombres réels. Ces perceptions incluent des données communes partagées par tous les agents : l'heure, l'énergie disponible, l'équité des conforts (calculée comme une dispersion statistique par l'index de Hoover), la quantité d'énergie non-utilisée et donc perdue, l'autonomie (l'absence de transactions avec le réseau national), le bien-être (médiane des conforts), l'exclusion (la proportion d'agents dont le confort est inférieur à 50% du bien-être), la sur-consommation. Les agents perçoivent en plus des détails sur eux-mêmes, auxquels les autres agents n'ont pas accès : la quantité d'énergie disponible dans la batterie personnelle, le confort au pas de temps précédent, et le bénéfice obtenu en vendant et achetant de l'énergie.

Récompenses. Comme décrit dans notre modèle, les récompenses sont calculées à partir des jugements des agents juges. Ceux-ci sont implémentés en langage Jason [8] sur la plateforme JaCaMo [5], qui est implémentée en Java. Une API REST permet d'assurer la communication entre les agents apprenants, implémentés en Python, et les agents juges sur JaCaMo.

Nous avons implémenté quatre agents juges, un pour chacune des valeurs morales proposées, contenant des règles dans un langage pseudo-Prolog, par exemple `supporte(donne_energie(X)) :- X > 0`, qui signifie que l'action de donner une quantité X d'énergie supporte la valeur associée (dans ce cas, Viabilité Environnementale) si la quantité est positive. De manière similaire, des règles "trahit" déterminent si l'action trahit la valeur.

5 Résultats

Nous avons mené plusieurs simulations, en considérant différents paramètres. En variant le nombre d'agents apprenants, les simulations "Petit" (20 Habitations, 5 Bureaux, 1 École) et "Moyen" (80 Habitations, 19 Bureaux, 1 École) permettent d'évaluer le passage à l'échelle de notre approche. Les simulations "Journalier" utilisent le profil de consommation moyenné sur une seule journée, tandis que les simulations "Annuel" utilisent le jeu de données complet, introduisant ainsi des variations saisonnières et donc plus de difficultés pour l'apprentissage. Enfin, nous proposons sept scénarios pour la configuration des agents juges, dont quatre qui incluent un seul agent juge (nommés "mono-valeur"), un dans lequel les juges sont activés un par un à des pas de temps différents ("Incrémental"), un dans lequel les juges sont initialement tous actifs et désactivés un par un à des pas de temps différents ("Décremental") et un scénario "Défaut" dans lequel les quatre agents sont activés en permanence. Cette variété de scénarios permet de comparer la présence et l'absence de chaque règle morale, l'impact sur les autres règles, et la capacité des agents apprenants à s'adapter quand les règles évoluent au fil du temps, soit en les ajoutant soit en les enlevant. Chaque ensemble de simulations a été lancé 20 fois sur 10 000 pas de temps.

La comparaison entre les expérimentations "Petit" et "Moyen" n'a pas montré de réelle différence entre les moyennes (T-Test, p-value=0.83), ce qui indique que notre approche passe à l'échelle, bien que le temps d'exécution soit naturellement bien plus long.

Nous nous concentrons sur les scénarios "Défaut" et "Incrémental" car ils sont les plus pertinents ; les "mono-valeurs" sont utiles en tant que scénarios de contrôle pour comparer les effets d'une valeur morale sur le comportement des agents quand la valeur est isolée ou agrégée avec d'autres. Le scénario "Decrémental" montre la capacité de supprimer des règles mais

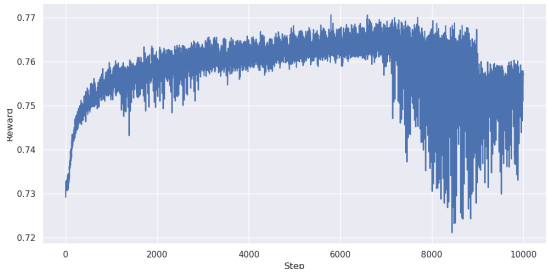

FIGURE 6 – Récompense moyenne pour tous les agents apprenants, sur chaque pas de temps, dans les simulations “Petit - Annuel - Défaut”.

n’est pas aussi intéressant que la capacité d’en ajouter (le scénario “Incrémental”). Nous nous concentrons également sur le profil “Annuel” comme il contient des variations de saisons et donc plus de difficultés.

La figure 6 montre que la récompense moyenne augmente au fur et à mesure de la simulation, bien que l’augmentation pourrait être plus importante ; la méthode proposée est donc efficace et les agents apprenants sont capables de se conformer aux règles données. Toutefois, les récompenses chutent vers la fin de la simulation, en grande partie à cause de la récompense d’“Inclusion”, qui semble plus difficile à apprendre. Il n’est pas clair s’il s’agit d’un problème lié à l’algorithme d’apprentissage ou à l’implémentation proposée des règles morales. Nous remarquons que la valeur d’“Inclusion” est celle avec le plus grand nombre de règles implémentées ; peut-être que cela est lié à son apparente difficulté d’apprentissage.

De plus, la Figure 7 montre que les agents ont été capables d’apprendre la valeur de “Viabilité Environnementale”, et en particulier quand ils ne disposaient que de la récompense agrégée, bien que la variation n’était pas aussi importante que quand ils disposaient spécifiquement de cette valeur comme récompense. Il est intéressant de noter que la comparaison entre “Défaut” et “Incrémental” montre que l’addition un par un des Juges semble mitiger l’impact négatif d’“Inclusion”. Les agents sont encore capables d’apprendre la “Viabilité Environnementale” et performent légèrement mieux sur l’“Inclusion”.

Selon ces figures, les agents *Écoles* ont eu les plus grandes variations dans les récompenses, tandis que les *Habitations* et *Bureaux* avaient une augmentation plus stable. Ce n’est pas surprenant, car les agents *École* ont la plus grande capacité d’action et ont donc un impact plus important sur l’environnement.

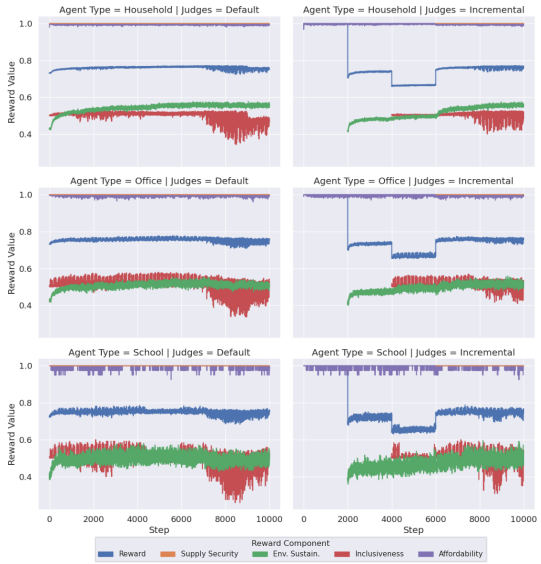

FIGURE 7 – Comparaison entre les récompenses individuelles de chaque agent apprenant, au fil des pas de temps, moyennées sur les simulations “Petit - Annuel - Défaut” et “Petit - Annuel - Incrémental”, et sur les agents de même profil. La courbe “Récompense” est la moyenne des quatre composantes.

6 Discussion

Pour rappel, notre contribution est une nouvelle méthode pour apprendre des “comportements éthiques”, i.e. des comportements qui exhibent une ou plusieurs valeurs morales et qui seraient considérés comme éthiques d’un point de vue humain, en utilisant des jugements symboliques comme sources de récompenses pour les agents apprenants dans un système multi-agent. Nous avons évalué cette approche sur un problème de répartition d’énergie, dans un contexte de *Smart Grid* simulé. Les expérimentations menées et en cours servent de preuve de concept pour montrer l’intérêt de notre approche.

Par rapport à la littérature existante, cette approche offre plusieurs avantages. Tout d’abord, il est important de noter que l’acceptation courante de la société sur l’éthique peut évoluer au fil du temps ; ainsi, les approches qui visent l’“éthique par conception” doivent considérer la capacité du système à s’adapter à des règles changeantes. Cet aspect n’a pas été extensivement étudié, y compris parmi les travaux se focalisant sur l’apprentissage ; dans cet article, nous avons montré grâce aux scénarios “Incrémental” et “Decré-

mental" la capacité de nos agents à s'adapter à l'ajout ou la suppression de règles. Cela est particulièrement visible en se comparant au travail de *reward-shaping* discuté précédemment [25] : si le comportement exemple n'est plus en accord avec le comportement attendu, il leur serait probablement nécessaire de re-créer un nouveau jeu de données et d'entraîner l'agent depuis zéro. Dans notre cas, nous pouvons simplement ajouter ou supprimer les règles. Toutefois, un avantage de leur approche est qu'ils supposent une récompense éthique qui n'est pas spécifique à la tâche, tandis que nos règles morales sont spécifiques aux domaines. Il serait peut-être possible d'implémenter des règles morales plus génériques, toutefois cela requiert l'existence de telles règles ; une possible source d'inspiration peut être les nombreux principes directeurs proposés sur *Ethical AI* ou *Responsible AI* [21].

Les précédents cas d'applications pour des "agents éthiques" étaient limités à des actions discrètes (e.g., dilemmes tel que le Dilemme du Tramway [23], robot accompagnant, [2], soldats robots [3], gestion autonome d'actifs en bourse [14]). Ces travaux sont importants mais il existe de nombreuses situations requérant de plus fines actions ; il est ainsi importer de proposer et d'expérimenter sur des environnements avec actions continues tel que le simulateur utilisé ici.

Certains travaux proposent d'utiliser des méthodes de vérification formelle pour garantir la conformité aux règles morales dans n'importe quelle situation identifiée [11, 7]. Dans notre cas, l'introduction d'agents apprenants nuit à cette possibilité ; il existe toutefois des travaux qui tentent d'appliquer de la vérification formelle à des algorithmes d'apprentissage par renforcement [16].

Ce travail cible les considérations "par conception", mais il y a également d'autres implications à l'éventuelle intégration d'un tel système dans la société. En effet, nous pouvons noter au moins un impact positif et un négatif : d'un côté, l'utilisation de règles symboliques est supposée plus facile à comprendre qu'une fonction mathématique. Toutefois, l'intelligibilité n'était pas l'objectif principal de ce travail et n'était pas évaluée par nos expérimentations. Nous pensons que l'intelligibilité du processus de récompense est cruciale, en particulier pour de la supervision humaine, que ce soit par les concepteurs du système ou des régulateurs externes. Il s'agit ainsi d'un point important à considérer et améliorer dans de futurs travaux.

D'un autre côté, les jugements nécessitent de nombreuses données sur les agents apprenants, e.g. leurs actions, leurs perceptions, ce qui entrave leur vie privée. Il pourrait être possible de limiter les données échangées en offrant des jugements limités, ou d'anonymiser les données pour que les juges ne puissent pas identifier les agents. Dans cet article, nous avons simplement considéré que les données étaient librement accessibles.

Notre approche a toutefois quelques limites. Premièrement, les règles morales utilisées servent de preuve de concept pour montrer l'intérêt de notre approche hybride, mais il serait intéressant d'étendre les agents juges et leurs règles afin de juger des situations plus complexes. Deuxièmement, la méthode utilisée pour transformer les jugements symboliques en récompense numérique par association des symboles à des nombres pour prendre la moyenne permet de facilement résoudre les conflits entre règles, mais d'autres méthodes sont possibles, en particulier un mécanisme d'argumentation entre juges de sorte à établir une priorité entre les règles selon le contexte. Par exemple, imaginons le cas où un hôpital est en manque crucial d'énergie, la règle interdisant l'achat d'énergie selon la valeur de Viabilité Environnementale pourrait être mise de côté dans ce cas précis.

Remerciements

Ce travail a été financé par la Région Auvergne Rhônes-Alpes (AURA), au sein du projet Ethics.AI (Pack Ambition Recherche). Les auteurs remercient leurs partenaires dans ce projet.

Références

[1] Colin Allen, Iva Smit, and Wendell Wallach. Artificial Morality : Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7(3) :149–155, September 2005.

[2] Michael Anderson, Susan Leigh Anderson, and Vincent Berenz. A value-driven eldercare robot : Virtual and physical instantiations of a case-supported principle-based behavior paradigm. *Proc. IEEE*, 107(3) :526–540, 2019.

[3] Ronald C Arkin, Patrick D Ulam, and Brittany Duncan. An ethical governor for constraining lethal action in an autonomous system. Technical report, Georgia Institute of Technology, 2009.

[4] Anne Boijmans. The Acceptability of Decentralized Energy Systems. Master's thesis, Delft University of Technology, July 2019.

[5] Olivier Boissier, Rafael Bordini, Fred Hübner, Jomi, and Alessandro Ricci. *Multi-Agent Oriented Programming : Programming Multi-Agent Systems Using JaCaMo*. The MIT Press, 2020.

[6] Vincent Bonnemains. *Formal ethical reasoning and dilemma identification in a human-artificial agent system.* PhD thesis, Institut supérieur de l'aéronautique et de l'espace, Toulouse, France, 2019.

[7] Grégory Bonnet, Bruno Mermet, and Gaële Simon. Vérification formelle et éthique dans les sma. In *JFSMA*, pages 139–148, 2016.

[8] Rafael H Bordini, Jomi Fred Hübner, and Michael Wooldridge. *Programming multi-agent systems in AgentSpeak using Jason*, volume 8. John Wiley & Sons, 2007.

[9] Rafael H Bordini, Amal El Fallah Seghrouchni, Koen Hindriks, Brian Logan, and Alessandro Ricci. Agent programming in the cognitive era. *Autonomous Agents and Multi-Agent Systems*, 34, 2020.

[10] Michael Bosello and Alessandro Ricci. From programming agents to educating agents–a jason-based framework for integrating learning in the development of cognitive agents. In *International Workshop on Engineering Multi-Agent Systems*, pages 175–194. Springer, 2019.

[11] Paul Bremner, Louise A Dennis, Michael Fisher, and Alan F Winfield. On proactive, transparent, and verifiable ethical reasoning for robots. *Proceedings of the IEEE*, 107(3) :541–561, 2019.

[12] Rémy Chaput, Olivier Boissier, Mathieu Guillermin, and Salima Hassas. Apprentissage adaptatif de comportements éthiques. In *28e Journées Francophones sur les Systèmes Multi-Agents (JFSMA'2020)*. Cépaduès, 2020.

[13] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Jugement éthique dans les systèmes multi-agents. In *JFSMA*, pages 149–158, 2016.

[14] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Multi-agent based ethical asset management. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 52–57, 2016.

[15] Virginia Dignum. *Responsible Artificial Intelligence : How to Develop and Use AI in a Responsible Way*. Springer Nature, 2019.

[16] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods : Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Teuvo Kohonen. Essentials of the self-organizing map. *Neural Networks*, 37 :52–65, 2013.

[18] James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4) :18–21, 2006.

[19] Shelley Nason and John E Laird. Soar-rl : Integrating reinforcement learning with soar. *Cognitive Systems Research*, 6(1) :51–59, 2005.

[20] Nicolas P. Rougier and Yann Boniface. Dynamic self-organising map. *Neurocomputing*, 74(11) :1840–1847, 2011.

[21] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas. What's next for ai ethics, policy, and governance ? a global overview. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 153–158, 2020.

[22] Richard S. Sutton and Andrew G. Barto. Reinforcement learning : An introduction. *IEEE Trans. Neural Networks*, 9(5) :1054–1054, 1998.

[23] Judith Jarvis Thomson. Killing, letting die, and the trolley problem. *The Monist*, 59(2) :204–217, 1976.

[24] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3) :279–292, May 1992.

[25] Yueh-Hua Wu and Shou-De Lin. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[26] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 5527–5533, Stockholm, Sweden, July 2018. AAAI Press.