Rémy Chaput, Jérémy Duval, Olivier Boissier,
Mathieu Guillermin, Salima Hassas

Apprentissage de comportements éthiques multi-valeurs par combinaison d'agents juges symboliques et d'agents apprenants

Volume 4, nº 2 (2023), p. 41-66.

DOI not yet assigned

La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 2967-9672

Revue Ouverte d'Intelligence Artificielle
Volume 4, n° 2, 2023, 41-66

Apprentissage de comportements éthiques multi-valeurs par combinaison d'agents juges symboliques et d'agents apprenants

Rémy Chaput[a], Jérémy Duval[b], Olivier Boissier[c], Mathieu Guillermin[d], Salima Hassas[e]

[a] Univ Lyon, UCBL, CNRS, INSA Lyon, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69622 Villeurbanne, France
E-mail : remy.chaput@univ-lyon1.fr

[b] LIRIS
E-mail : jeremy-duval1@hotmail.fr

[c] Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France
E-mail : Olivier.Boissier@emse.fr

[d] Sciences and Humanities Confluence Research Center, Lyon Catholic University
E-mail : mguillermin@univ-catholyon.fr

[e] Univ Lyon, UCBL, CNRS, INSA Lyon, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69622 Villeurbanne, France
E-mail : salima.hassas@univ-lyon1.fr.

Résumé. — Afin de répondre au besoin d'incorporer des considérations éthiques au sein d'algorithmes d'Intelligence Artificielle, nous proposons une nouvelle méthode hybride, combinant raisonnement et apprentissage, où des agents juges évaluent l'éthique du comportement d'agents apprenants. Cette séparation offre plusieurs avantages : co-construction entre agents et humains ; juges plus accessibles pour des humains non-experts ; récompense plus riche par l'utilisation de multiples valeurs morales. 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.

1. Introduction

Avec l'utilisation croissante d'applications utilisant des modèles d'Intelligence Artificielle (IA), des questions se posent sur les risques de ces applications, et les

moyens d'assurer qu'elles génèrent des comportements en adéquation avec les valeurs humaines, que nous nommerons « comportements éthiques". Ces questions se regroupent sous le problème que Stuart Russell nomme « l'alignement de valeurs » (*Value Alignment*[1]).

Des débats, à la fois sociétaux et académiques, s'organisent pour tenter de répondre à ces questions, par diverses propositions. Plus de 80 documents ont ainsi été produits entre 2016 et 2020 [26], par des acteurs publics (gouvernements, organisations intergouvernementales), privés, ou encore des organisations à but non lucratif. Ces documents incluent notamment des déclarations de multiples « principes » ou « valeurs », ou encore des (ébauches de) réglementations cherchant à limiter l'impact négatif éventuel de telles applications. L'Europe notamment a produit un ensemble de recommandations par le biais du *High-Level Expert Group on AI*[2], ainsi qu'une ébauche de législation pour encadrer les applications d'IA en fonction de leurs risques[3].

Dans le monde académique, différents domaines s'intéressent aux moyens de répondre à ces questions : équité et justice (en particulier vis-à-vis de l'apprentissage automatique), transparence et explicabilité, etc. Nous nous focalisons dans cet article sur le domaine de la *Machine Ethics*, qui a trait à la conception d'« agents à impact éthique » [22], dont les comportements ont des conséquences éthiques sur des vies humaines. En particulier, l'« éthique dans la conception » [18] vise à ce que ces agents soient capables de prendre des décisions selon des considérations éthiques, afin de générer des comportements éthiques. 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, qui seront détaillés en section 2.1.4 ; ainsi, dans cet article, nous présentons une nouvelle approche, hybride, avec un apprentissage de comportements éthiques guidé par des récompenses issues de raisonnements symboliques.

Comme Dignum l'indique : « By definition, hybrid approaches have the potential to exploit the positive aspects of the top-down and bottom-up approaches while avoiding their problems. As such, these may give a suitable way forward. » [18]. Bien qu'un agent artificiel ne puisse compter comme un agent éthique complet, Dignum note que « AI systems will make and are already making decisions that would consider to have an ethical flavour if they were made by people » [18]. Comme évoqué plus haut, l'expression « comportement éthique » désigne les actions exécutées par un agent, lorsqu'elles sont considérées comme morales par un humain, selon un ensemble de valeurs. Par exemple, un agent raisonnant sur la règle « Tuer est immoral » et, ce faisant, décidant de ne pas tuer quelqu'un, exhibe un comportement éthique au sens de notre définition, bien qu'il soit programmé pour raisonner ainsi.

[1] https://www.youtube.com/watch?v=WvmeTaFc_Qw.
[2] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[3] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Cet article est structuré comme suit : nous présentons d'abord la littérature sur laquelle nous appuyons notre approche hybride dans la section 2 ; cette proposition est ensuite détaillée en section 3 ; la section 4 présente des expérimentations sur un cas d'application des *Smart Grids* et les résultats en section 5 démontrent la faisabilité ; finalement, la section 6 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

L'étude de l'éthique dans les systèmes d'IA est un champ relativement nouveau, désigné le plus souvent par *Machine Ethics*. Il concerne la conception de machines avec des principes éthiques guidant la manière dont elles se comportent vis-à-vis de leur environnement, en particulier des humains et des autres machines.

De nombreuses approches ont été développées au fil des deux dernières décennies, et nous proposons un (bref) résumé de certaines de leurs propriétés, afin de déterminer lesquelles sont intéressantes, et lesquelles sont encore manquantes. Ces différentes propriétés sous-tendent les choix de conception que nous détaillons dans la section 3.

2.1.1. *Domaines discrets ou continus*

Parmi les approches proposées dans la littérature, la plupart se focalisent sur l'utilisation de domaines discrets, à la fois pour la représentation des situations traitées, mais aussi des actions des agents ou systèmes d'IA.

L'emblématique dilemme du Tramway [28] propose ainsi deux actions discrètes (« tirer le levier » ou « ne rien faire »). De même, l'expérience *Moral Machine* [5] place l'utilisateur dans une situation de choix binaire (tuer un enfant ou les passagers, un animal ou un piéton, etc.). La défunte base *DilemmaZ*[(4)] répertoriait de nombreux exemples de dilemmes, proposés par la communauté des chercheurs. Ces dilemmes, bien que non définis formellement, semblent, au vu de leur définition, reposer également sur des domaines discrets.

Deux raisons nous semblent expliquer cette utilisation de domaines discrets. D'un point de vue technique, les domaines discrets sont plus faciles à implémenter et manipuler. D'un point de vue philosophique, les dilemmes éthiques peuvent être décrits comme une opposition entre deux (ou plus) actions, chacune impliquant un regret au

[(4)] Accessible via une archive de la *Wayback Machine* :
https://web.archive.org/web/20210323073233/https://imdb.uib.no/dilemmaz/articles/all.

vu d'une théorie éthique (définition simplifiée à partir de celle proposée par Vincent Bonnemains [8]).

Cette utilisation d'actions et situations discrètes dans les dilemmes est intéressante, notamment pour permettre d'évaluer différentes théories éthiques bien connues, *e.g.*, l'Impératif Catégorique de Kant, la Doctrine du Double Effet de Thomas D'Aquin, ou encore pour identifier des failles dans ces théories. Nous voyons aussi, avec l'exemple de GenEth pour les véhicules à conduite automatisée [2], ou pour le domaine médical [3], que cela permet de capturer certains dilemmes concrets, pouvant survenir dans notre monde si de tels algorithmes étaient déployés.

Toutefois, d'autres problèmes semblent ne pas se satisfaire d'une description discrète. Par exemple, concernant la répartition d'énergie au sein d'un groupe d'agents, et supposant qu'il n'y a pas assez d'énergie pour satisfaire chacun, la question n'est pas « Faut-il consommer de l'énergie, oui ou non ? », mais plutôt « Quelle quantité dois-je consommer ? ». Cette action peut encore s'exprimer de manière discrète, en considérant un ensemble suffisamment grand, *e.g.*, {Consommer 0 W, Consommer 1 W, . . . , Consommer 1 000 W} (un ensemble de N actions). Toutefois, lorsque l'on considère de multiples dimensions dans une même action, telle que « Quelle quantité dois-je consommer *et* quelle quantité dois-je acheter ? », cette représentation atteint ses limites. L'ensemble d'actions discrètes serait alors {(Consommer 0 W, Acheter 0 W), (Consommer 1 W, Acheter 0 W), . . . , (Consommer 1 000 W, Acheter 1 000 W)} (un ensemble de taille $N * N$). L'explosion combinatoire rend cette représentation discrète peu pratique.

De même, certaines situations s'expriment avec des caractéristiques continues, plutôt que discrètes. Par exemple, plutôt que de décrire une situation comme étant « équitable », ou « non équitable », on peut se demander « à quel point est-elle équitable ? ».

2.1.2. *Mono- ou multi-agent*

De nombreux travaux considèrent un unique agent isolé dans son environnement [32], tel que GenEth [3]. D'autres, tels que Ethicaa [15], utilisent de multiples agents dans un environnement commun.

Nous arguons que ce deuxième cas est important. En effet, il s'agit tout d'abord d'une situation plus réaliste : les « agents éthiques » sont voués à être inclus dans notre société, et non à vivre en parfaite isolation. La question de l'impact d'interactions avec d'autres agents, qu'ils soient humains ou artificiels, est donc primordiale. D'autre part, de telles interactions soulèvent de nouvelles questions.

Reprenons par exemple le problème de la répartition d'énergie. Dans le cas où un agent serait seul à consommer à partir d'une source donnée, il y a très certainement des questions éthiques, comme celle de la provenance de l'énergie (est-elle fortement carbonée ?) et de l'écologie. Toutefois, rajouter d'autres agents, et donc des interactions, peut apporter de nouvelles questions, non initialement présentes. Considérant la source d'énergie comme étant finie, il est possible que tous les agents ne puissent consommer

autant qu'ils voudraient en même temps. Cela rajoute la question de l'équité : nous voulons que les agents s'assurent un juste accès à la quantité d'énergie qu'ils peuvent consommer.

Plus généralement, les systèmes multi-agents permettent de considérer la confrontation de plusieurs éthiques. Que se passe-t-il quand des agents sont guidés par des systèmes de valeurs différents ? Ou quand un agent est amoral, voire purement immoral ? Comment les autres agents réagissent-ils, sont-ils capables de compenser, de s'adapter ?

2.1.3. *Une ou plusieurs valeurs morales*

Tandis que certaines approches considèrent un unique objectif, souvent formulé comme « maximiser le bien-être », ou similaire, de nombreux travaux considèrent de multiples valeurs morales, ou enjeux éthiques, ou plus généralement objectifs à satisfaire. Dans l'exemple précédent sur la répartition d'énergie, nous mentionnions à la fois le respect de l'écologie, mais aussi de l'équité.

Dennis et Fisher notent ainsi que, si dans de nombreuses situations il n'y a que peu d'enjeux éthiques, dans les situations qui nous intéressent, nous avons souvent à faire un choix entre plusieurs objectifs [17]. Ainsi, considérer de multiples valeurs morales permet d'expliciter le compromis à faire entre celles-ci, en particulier dans les situations dans lesquelles elles s'opposent.

2.1.4. *Approche ascendante, descendante, ou hybride*

Probablement la propriété la plus discutée, le type d'approche caractérise le principe éthique et la façon dont il est implémenté dans un système donné. À l'instar de la classification traditionnellement faite en IA, les travaux en *Machine Ethics* sont divisés en trois catégories [1] : approches descendantes (*Top-Down*), ascendantes (*Bottom-Up*) et hybrides.

Descendantes. — Les approches descendantes s'intéressent à la formalisation de principes éthiques existants, tel que l'Impératif Catégorique de Kant. Cette formalisation passe par une représentation sous forme symbolique, le plus souvent par des approches logiques, à base de règles, ou encore des ontologies. Le raisonnement sur ces représentations peut ainsi s'appuyer sur des connaissances expertes, qui sont injectées à priori dans ces représentations. Elles peuvent également offrir une meilleure lisibilité, à la fois des connaissances, mais aussi du raisonnement produit.

Par exemple, l'*Ethical Governor* [4] vérifie l'adéquation des actions avec des règles pré-établies comme les Règles d'Engagement ou le Droit de la guerre. Cette approche en a inspirée d'autres, telle que celle de Bremner *et al.* [13], proposant d'utiliser un module de planification pour générer des plans, *i.e.*, des séquences d'actions, qui soient compatibles avec les règles éthiques. Ces approches sont particulièrement adaptées à la vérification formelle de l'adéquation du comportement : il est possible de garantir

que l'agent ne sortira pas du cadre défini. En revanche, il peut être difficile de générer sous une forme symbolique des plans, quand les règles sont en conflit. Par exemple, si deux valeurs morales recommandent chacune une action différente, mutuellement exclusives, aucun plan ne peut être entièrement satisfaisant.

Dans *Ethicaa* [15], des agents *Beliefs Desires Intentions* (BDI) raisonnent sur plusieurs principes éthiques pour décider de leur comportement et juger les actions des autres agents. Les agents disposent également d'une relation de préférence sur les différents principes, ce qui leur permet de trancher les situations de conflit.

Ces approches permettent de représenter une diversité de principes éthiques : selon celui considéré, elles peuvent ainsi adopter une position conséquentialiste, *i.e.*, axée sur les conséquences, ou bien déontologique, *i.e.*, axée sur le respect de règles.

Un des principaux désavantages de telles approches est que, du fait de leur corpus de connaissances explicite, mais figé, elles ne peuvent s'adapter à des situations non prévues, ou à une évolution de l'éthique. Nous arguons que cette capacité à s'adapter est particulièrement importante ; elle rejoint ce que Nallur appelle l'*apprentissage continu* : « Any autonomous system that is long-lived must adapt itself to the humans it interacts with. All social mores are subject to change, and what is considered ethical behaviour may itself change. » [23]. Nous remarquons par ailleurs que, dans son panorama de la littérature, 1 seule des 10 approches considérées possède cette capacité [23, voir Table 2].

Ascendantes. — 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* [3] apprend, à partir de décisions d'éthicisites dans de multiples contextes, quelle action devrait être prise pour chacun des contextes. Cet apprentissage, par *Inductive Logic Programming* (ILP), conduit à l'élaboration d'une formule logique, permettant de déterminer quelle action doit être prise, quelle que soit la situation. L'ILP permet de formuler une règle qui soit suffisamment générique pour s'appliquer à d'autres situations, non rencontrées dans le jeu de données.

Une autre approche propose d'utiliser de l'apprentissage par renforcement (RL) [31]. L'apprentissage par renforcement repose sur l'utilisation de récompenses pour renforcer, ou au contraire atténuer, un comportement donné. Traditionnellement, cette récompense est calculée en fonction de la tâche à réaliser. Le travail de Wu *et al.* [31] propose d'y ajouter une composante éthique, sous forme de différence entre le comportement de l'agent, et celui d'un humain moyen, obtenu à partir d'un jeu de données de comportements, supposé exhiber des considérations éthiques. La récompense finale est calculée comme la somme de la récompense « par tâche » et de la récompense « éthique », l'agent apprend ainsi à réaliser sa tâche tout en exhibant les considérations éthiques qui sont encodées dans les exemples humains.

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. En

effet, si les normes courantes de la société vis-à-vis de l'éthique changent, il faudra que le comportement de ces agents change également. Dans ces deux approches, cela nécessitera très probablement de devoir créer un nouveau jeu de données, et de réapprendre les agents depuis 0 à partir de ces nouvelles données. De plus, les approches ascendantes sont plus difficiles à interpréter que les approches descendantes. Par exemple, un régulateur humain souhaitant comprendre le comportement attendu devra s'intéresser au jeu de données, potentiellement difficile à appréhender, de par sa structure, mais aussi la quantité de données.

Hybrides. — 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.

Par exemple, Honarvar *et al.* [20] proposent de combiner des agents BDI avec du Raisonnement à partir de Cas et un réseau de neurones artificiel. Face à une situation donnée, l'agent propose une action à effectuer, puis utilise sa base de cas déjà connus pour voir si, dans un cas similaire, une action similaire a été considérée comme éthique. Si c'est le cas, l'action est effectuée. Si l'agent ne dispose pas de cas suffisamment proche, l'agent effectue l'action, et utilise le réseau de neurones pour évaluer si cette action était effectivement en accord avec les considérations éthiques. Le résultat de cette évaluation est mémorisé dans la base de cas, pour réutilisation éventuelle lors de la prochaine décision.

Cette approche combine effectivement des capacités de raisonnement et d'apprentissage ; toutefois, elle peut se révéler difficile à appliquer. En effet, l'utilisation de raisonnement à partir de cas permet de regrouper des situations et actions proches, mais implique de spécifier comment doit se faire ce regroupement, et comment adapter une évaluation quand l'action diffère. Par exemple, supposons que dans une situation s donnée, une action ait été effectuée ; l'action correspond à un paramètre de valeur X et a été évaluée comme éthique. Dans une situation s' reconnue comme similaire à s, l'agent propose une action $X + 1$. Cette action est-elle éthique ? Comment traduire la différence entre X et $X + 1$ en terme d'impact éthique ? Doit-on dire que l'action $X + 1$ est autant, plus, ou moins éthique que l'action X ?

Pour plus de détails, le lecteur peut se référer à [1], ou au panorama des différentes implémentations proposé par Tolmeijer *et al.* [29].

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 [10]. 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 [24] ou BDI-RL [12], 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 [4, 13], et sont souvent qualifiés d'hybride. Cependant, nous arguons que l'agent n'apprend pas vraiment à partir du « Ethical Layer » mais est plutôt contraint par celui-ci. Ainsi, l'approche complète ne bénéficie pas des avantages de l'apprentissage, telle que la capacité de généraliser.

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 [15]. À notre connaissance, l'intégration d'un jugement symbolique pour donner une récompense numérique aux agents apprenants n'a pas 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. Ainsi, nous proposons une approche dans laquelle les situations et actions sont décrites de manière continue. Cette approche, hybride, combine les avantages de l'apprentissage, telle que la capacité à généraliser sur des situations, et les avantages du raisonnement, telles qu'une meilleure intelligibilité du comportement attendu et la possibilité d'introduire des connaissances expertes.

Dans les sous-sections suivantes, nous décrivons cette architecture à un niveau abstrait, puis nous formalisons le modèle comme un jeu Markovien, détaillons le fonctionnement des agents apprenants et juges, et enfin expliquons les interactions entre juges et apprenants, qui constituent la fonction de récompense.

3.1. Architecture Abstraite

Considérons un système multi-agent comprenant des humains et des agents artificiels, représenté dans la figure 3.1. 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.

Le jugement, par raisonnement symbolique, et l'apprentissage sont séparés en des agents différents, qui peuvent évoluer indépendamment. Par exemple, les règles des agents juges peuvent être mises à jour par des concepteurs humains tandis que les agents apprenants adaptent leur comportement pour se conformer à ces nouvelles règles. Il

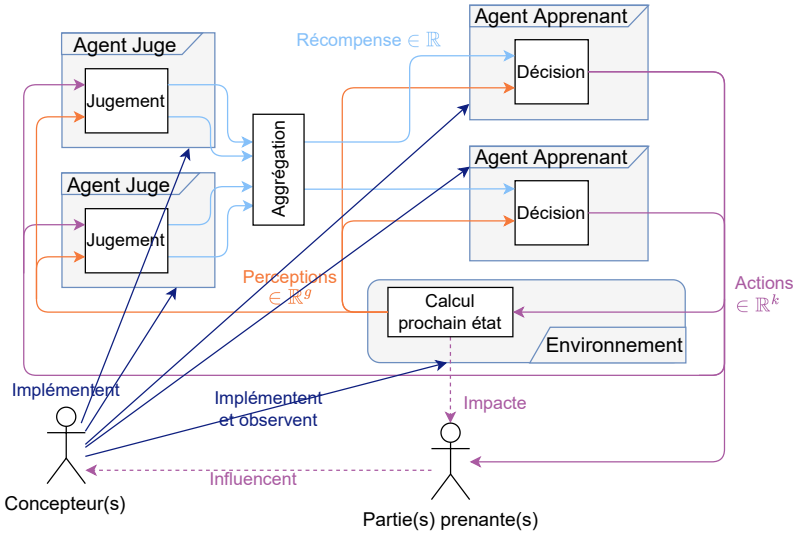

FIGURE 3.1. Architecture de notre approche, comprenant des concepteurs humains implémentant des règles pour juger les agents apprenants. Les actions des apprenants modifient un environnement partagé, ce qui impacte les humains. Les parties prenantes influencent les concepteurs pour implémenter des règles de jugement qui rendent cet impact bénéfique.

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[(5)].

Remarquons que cela mène à une boucle de co-construction : il peut être difficile de spécifier exactement ce que l'agent doit faire, dans chaque situation, en particulier lorsque les valeurs morales s'opposent. En revanche, il est plus facile de juger une situation, de spécifier les éléments qui ne sont pas acceptables et doivent être améliorés, et de récompenser les bons comportements. Ainsi, les agents apprenants vont, à travers cette boucle exécution-jugement-apprentissage, encoder les règles symboliques, *i.e.*, la « connaissance éthique » fournie par les concepteurs humains, sous forme d'un comportement.

De plus, afin de faciliter l'utilisation de multiples valeurs morales, nous proposons de distinguer chaque valeur comme un agent juge séparé. Cela offre plusieurs avantages : le modèle en est clarifié ; il est possible de modifier les règles de jugement par l'addition et la suppression d'agents juges ; cela ouvre également la voie à des interactions complexes entre les agents juges, tels que des processus d'argumentation. La combinaison de multiples points de vue par ces différents juges permet d'obtenir une récompense plus riche pour les agents apprenants, notamment par la confrontation des valeurs morales.

[(5)] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Ces agents juges sont basés sur les agents Ethicaa [15], 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.

Afin d'évaluer l'impact de plusieurs agents, et leurs interactions, dans un environnement partagé, notre approche introduit de multiples agents apprenants. Ainsi, notre approche proposée est un système multi-agent, à plusieurs niveaux : séparation des agents entre juges et apprenants, de multiples juges, et de multiples apprenants.

Lorsque le comportement produit ne convient pas, les règles de jugement peuvent être modifiées afin de déclencher une adaptation de la part des agents apprenants. 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. Cette capacité d'adaptation est nécessaire également lorsque le comportement attendu évolue au cours du temps, du fait des dynamiques de notre société.

Le rapport du HLEG AI reconnaît également l'intelligibilité du comportement attendu comme étant de plus en plus important dans les applications d'IA. Pour ce faire, le processus de jugement utilise des valeurs et règles morales explicites, symboliques, dans les agents juges. Nous arguons que l'intelligibilité est encore plus importante dans le contexte des *Machine Ethics*. Les humains, qu'ils soient concepteurs, ou parties prenantes, notamment utilisateurs du système et régulateurs externes, devraient être capables de comprendre le jugement produit, afin de vérifier la compatibilité avec leur propre système de principes éthiques. Bien que ce point ne sera pas évalué dans les expérimentations, il sert en partie de motivation à notre approche.

3.2. Modèle Formel

Dans la suite, nous prenons quelques conventions afin de simplifier la lecture : les ensembles sont notés de manière calligraphiée, *e.g.*, $\mathcal{E}$; les vecteurs sont notés en gras, *e.g.*, $\mathbf{v}$; les fonctions sont en police à chasse fixe, *e.g.*, $\mathtt{F}$.

Considérons un environnement composé de deux ensembles d'agents artificiels. Dans l'ensemble $\mathcal{J}$ des agents *juges*, chaque agent $j \in \mathcal{J}$ est associé à une seule valeur morale, et un ensemble de règles morales permettant de décider si une action, dans une situation donnée, *supporte* ou *trahit* cette valeur morale (*e.g.*, la Justice, l'Inclusivité). Dans le second ensemble $\mathcal{L}$ des agents *apprenants*, chaque agent $l \in \mathcal{L}$ apprend un comportement et effectue des actions dans l'environnement. Ils reçoivent une récompense, basée sur l'agrégation des jugements $\mathtt{Jugement}_j(l)$ par les $j \in \mathcal{J}$ sur leur comportement.

Conceptuellement, cette description se rapproche d'un jeu Markovien (ou *Stochastic Game*), une extension du Processus de Décision Markovien (MDP) à plusieurs agents.

Définition 3.1 (Jeu Markovien). — *Formellement, nous définissons un jeu Markovien comme un n-uplet* $\langle \mathcal{S}, \mathcal{L}, \mathcal{A}_0, \ldots, \mathcal{A}_n, \mathtt{T}, \mathtt{R}_0, \ldots, \mathtt{R}_n \rangle$ *:*

- $\mathcal{S}$ *est l'ensemble de tous les états possibles, tel que* $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^g$ *est un vecteur de nombres réels de dimension g, i.e., un état multi-dimensionnel et continu.*
- $\mathcal{L}$ *est l'ensemble des agents, de taille n ; les agents juges n'agissant pas dans l'environnement, nous ne considérons ici que les apprenants* $\mathcal{L}$*.*
- $\mathcal{A}_l$ *est l'ensemble des actions possibles pour l'agent l, tel que* $\mathbf{a}_l \in \mathcal{A}_l \subseteq \mathbb{R}^k$ *est un vecteur de nombres réels de dimension k, i.e., une action paramétrée. De plus, on définit l'ensemble des actions-jointes* $\mathcal{A} = \mathcal{A}_0 \times \cdots \times \mathcal{A}_n$ *comme l'ensemble regroupant les combinaisons des actions possibles par les différents agents apprenants.*
- $\mathtt{T}$ *est la fonction de probabilité de transition, définie par* $\mathtt{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$*, i.e.,* $\mathtt{T}(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ *est la probabilité de passer de l'état* $\mathbf{s}$ *à* $\mathbf{s}'$ *en effectuant l'action-jointe* $\mathbf{a}$*.*
- $\mathtt{R}_l$ *est la fonction de récompense de l'agent* $l \in \mathcal{L}$*, définie par* $\mathtt{R}_l : \mathcal{S} \times \mathcal{A}_l \times \mathcal{S} \to \mathbb{R}$*, i.e.,* $\mathtt{R}_l(\mathbf{s}, \mathbf{a}_l, \mathbf{s}')$ *est la récompense de l'agent l pour avoir effectué l'action* $\mathbf{a}_l$ *dans l'état* $\mathbf{s}$*, résultant en l'état* $\mathbf{s}'$*.*

Les dimensions des différents espaces, g pour les états et k pour les actions, ne sont pas fixées par notre approche. Elles dépendent du cas d'application considéré, ce qui permet d'expérimenter avec des environnements complexes, *e.g.*, avec des états décrits par de nombreuses perceptions, ou des actions contenant de nombreux paramètres.

Les MDPs et jeux Markovien peuvent être résolus avec de l'apprentissage par renforcement [27], une méthode pour apprendre des comportements, représentés par une politique $\pi_l : \mathcal{S} \times \mathcal{A}_l \to [0, 1]$, qui affecte à chaque paire état-action une probabilité, telle que $\pi_l(\mathbf{s}_t, \mathbf{a}_i)$ est la probabilité de choisir l'action $\mathbf{a}_i$, au pas de temps t dans l'état $\mathbf{s}_t$. 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_l^*, en mettant de côté les détails de $\mathtt{R}_l$ que nous décrirons ensuite.

3.3. Agents Apprenants

Les agents apprenants doivent apprendre à exhiber un « comportement éthique », *i.e.*, réaliser leur tâche tout en respectant les considérations éthiques des concepteurs. Ces considérations étant encodées dans les règles de jugement des agents juges, qui sont ensuite transformées en récompenses, cela revient à apprendre comment sélectionner une action dans chaque état donné, afin de maximiser l'espérance des récompenses reçues 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 [14]. Cet algorithme utilise deux *Dynamic Self-Organizing Map* (DSOM) [25], inspiré des Cartes Auto-Organisatrices de Kohonen [21], 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 [30], 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 [14] pour une description détaillée de l'algorithme. Un exemple est présenté dans la figure 3.2.

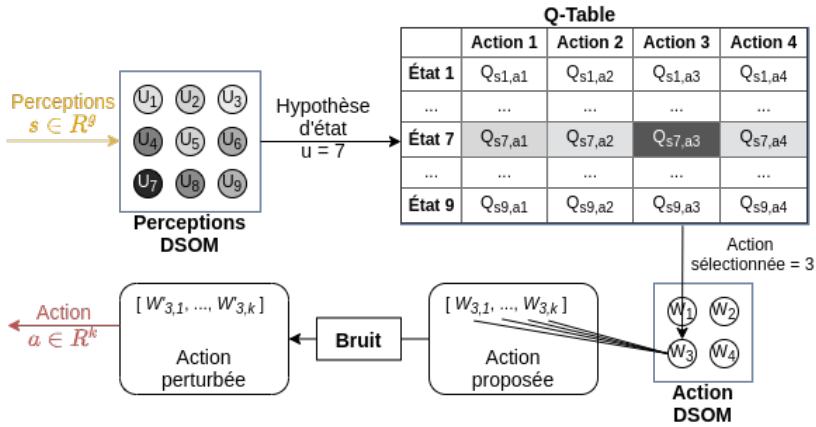

FIGURE 3.2. Exemple de décision : l'agent reçoit un état, vecteur de réels, comparé à la *State-DSOM*. Le 7e neurone, qui a le vecteur prototype le plus proche, est choisi comme hypothèse d'état. À partir de la Q-Table et de ce 7e état, la 3e action est choisie. L'action obtenue est le résultat de la perturbation par un bruit aléatoire du vecteur associé au 3e neurone de l'*Action-DSOM*.

À partir des récompenses produites par les jugements, les agents apprennent l'intérêt des actions dans les états, *i.e.*, la *Q-Table*, par application de l'équation de Bellman, avec s_t, a_t l'état observé et l'action effectuée au pas de temps t (identifiants discrétisés par la *State-DSOM* et l'*Action-DSOM*), et r_{t+1} la récompense reçue après avoir effectué a_t dans s_t :

$$Q^{t+1}(s_t, a_t) = (1-\alpha)Q^t(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \arg\max_{a'} Q^t(s_{t+1}, a')\right] \tag{3.1}$$

3.4. Agents Juges

L'architecture BDI des agents juges (*cf.* figure 3.3) s'appuie sur les travaux du projet Ethicaa [15] en simplifiant le mécanisme d'évaluation morale.

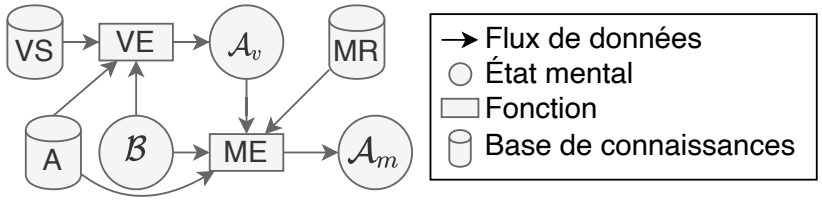

Figure 3.3. Architecture des agents juges, adaptée d'Ethicaa [15].

À chaque pas de temps, les agents juges génèrent des croyances ($\mathcal{B}$) à partir de leurs perceptions de l'environnement et des actions effectuées par ces apprenants. Notons que les travaux d'Ethicaa [15] proposent des jugements partiels ou omniscients ; dans cet article, nous choisissons de nous restreindre à des jugements partiellement informés, *i.e.*, les agents juges reçoivent les mêmes perceptions de l'environnement que les agents apprenants. Il est possible de rendre les agents juges omniscients en leur fournissant des données additionnelles sur la situation, que ne connaissent pas les agents apprenants : cela aurait pour effet de rendre le jugement plus précis, et donc certainement d'améliorer la qualité de l'apprentissage, au détriment de la vie privée, ces données pouvant concerner d'autres agents que l'agent jugé par exemple. À l'inverse, il serait aussi possible d'utiliser des jugements « aveugles », dans lesquels les juges n'auraient pas conscience des perceptions des agents apprenants, mais seulement de leurs propres observations de ces agents.

Rappelons que les actions sont des vecteurs de k nombres réels, que l'on appellera « paramètres ». Les agents juges décomposent une action $\mathbf{a}_l$ et traitent séparément chacun des k paramètres. En d'autres termes, ils considèrent comme des symboles différents $\{\forall i \in [[1,k]] : \mathbf{a}_{l,i}\}$, avec $\mathbf{a}_{l,i} \in \mathbb{R}$ le i-ème composant du vecteur.

Chaque agent juge dispose de son propre ensemble fixé de règles morales associées à sa valeur (VS et MR), qu'il utilise pour déterminer si chaque paramètre de l'action trahit ou supporte la valeur. L'Évaluation Morale (`ME`) utilise les croyances générées $\mathcal{B}$, notamment sur la situation, ainsi que les actions $\mathcal{A}$ pour produire une valuation morale ($\mathcal{A}_m$), c'est-à-dire un symbole v parmi l'ensemble des valuations possibles $\mathcal{V} = \{moral, immoral, neutral\}$.

À chaque pas de temps, le jugement par un agent juge j de l'action $\mathbf{a}_l$ d'un agent apprenant l est l'évaluation morale de chacun des k paramètres de cette action $\mathbf{a}_l$:

$$\texttt{Jugement}_j(l) = \left\{i \in [[1,k]] : \texttt{ME}_j(\mathcal{B}, \mathbf{a}_{l,i})\right\} \tag{3.2}$$

Autrement dit, le jugement d'un agent apprenant par un agent juge retourne une liste de valuations $\in \mathcal{V}^k$.

Chaque agent juge calcule un jugement différent pour chaque agent apprenant, de sorte que la fonction d'évaluation finale $\mathtt{F} : \mathcal{L} \to \left(\mathcal{V}^k\right)^{|\mathcal{J}|}$ retourne une liste de listes de valuations, soit $\forall l \in \mathcal{L} : \mathtt{F}(l) = \left\{\forall j \in \mathcal{J} : \mathtt{Jugement}_j(l)\right\}$.

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 chaque pas de temps de la simulation, les agents apprenants reçoivent des perceptions sur la situation, et choisissent une action à effectuer. Les agents juges perçoivent ensuite plusieurs données, et les stockent sous forme de croyances : liste des agents, les actions de chacun, les propriétés globales représentant l'état de l'environnement et les propriétés individuelles par agent.

Nous avons expliqué que l'évaluation des agents apprenants retourne une liste de liste de valuations, *i.e.*, des symboles. Or, la fonction de récompense $\mathtt{R}_l : \mathcal{S} \times \mathcal{A}_l \times \mathcal{S} \to \mathbb{R}$ doit retourner un nombre réel, souvent compris entre 0 et 1. Afin de transformer ces valuations symboliques en une récompense numérique, plusieurs méthodes sont possibles. Dans cet article, nous proposons une première version, simple, afin de démontrer la faisabilité de l'approche.

Soit $\mathbf{v}_{l,j}$ la liste des valuations $\in \mathcal{V}^k$ retournée par le jugement de l'agent juge j sur l'agent apprenant l, et soient *moral* (resp. *immoral*) le nombre de valuations *moral* (resp. *immoral*) dans $\mathbf{v}_{l,j}$. Nous transformons cette liste de valuations en une récompense numérique « par juge » : $r_{l,j} = \frac{moral}{moral+immoral}$. Cette récompense correspond au jugement d'un seul agent juge j. Ainsi, les actions morales, *i.e.*, celles dont le plus de paramètres reçoivent une valuation *moral*, tendent vers 1, tandis que les actions immorales, *i.e.*, celles dont le plus de paramètres reçoivent une valuation *immoral*, tendent vers 0. Notons que les valuations *neutral* sont ignorées dans cette formule : nous considérons que toutes les actions n'ont pas forcément d'enjeu éthique. En revanche, 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 une valeur par défaut de 0,5, correspondant à la situation où il y a autant de valuations *moral* qu'*immoral*.

Finalement, afin de retourner une récompense unique à chaque agent apprenant, nous prenons la moyenne des récompenses « par juge ». Pour résumer,

$$\begin{aligned}\mathtt{R}_l(\mathbf{s}, \mathbf{a}, \mathbf{s}') &= \mathtt{moyenne}\left(\left\{\forall j \in \mathcal{J} : r_{l,j}\right\}\right) \\ &= \mathtt{moyenne}\left(\left\{\forall j \in \mathcal{J} : \frac{\left|moral \in \mathbf{v}_{l,j}\right|}{\left|moral \in \mathbf{v}_{l,j}\right| + \left|immoral \in \mathbf{v}_{l,j}\right|}\right\}\right)\end{aligned} \tag{3.3}$$

$$\begin{aligned}\mathbf{v}_{l,j} &= \mathtt{Jugement}_j(l) \\ &= \left\{i \in [[1,k]] : \mathtt{ME}_j(\mathcal{B}, \mathbf{a}_{l,i})\right\}\end{aligned} \tag{3.4}$$

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 [14] : 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 4.1) : 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 [6]. Ces valeurs, originellement, prennent le point de vue de décideurs (notamment politiques) ; par exemple, la valeur de l'Abordabilité exprime que ces réseaux doivent être conçus de telle sorte que les utilisateurs n'ont pas besoin de payer trop cher pour les utiliser et satisfaire leurs besoins en énergie. Étant donné que nos agents apprenants agissent en tant que mandataires pour les *prosumers*, nous voulons qu'ils soutiennent les mêmes valeurs morales que ces utilisateurs. Ainsi, au lieu de prendre le point de vue de décideurs politiques, nous avons reformulé ces valeurs afin de prendre le point de vue de citoyens participant à un tel système, et prenant des décisions pour allouer de l'énergie.

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 : Abordabilité : 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.

Les règles MR1 et MR2 sont donc associées à un point de vue individuel, tandis que MR3 et MR4 considèrent l'ensemble de la société. Nous considérons qu'elles sont toutes des règles « morales », dans le sens où il serait immoral de demander à un utilisateur de ne pas satisfaire son confort, par exemple (en opposition à MR1).

Dans sa thèse, Vincent Bonnemains propose une définition formelle d'un dilemme éthique [8] : une situation est un dilemme éthique s'il y a un choix entre au moins deux actions possibles, tel que chacune des alternatives implique une transgression d'une règle morale, soit par sa nature propre, soit par ses conséquences, par rapport à une référence, telle qu'une norme contextuelle, une doctrine, une valeur morale, un principe ou un désir. Les valeurs morales que nous proposons sont susceptibles d'entraîner de telles situations de dilemme. En effet, supposons la quantité d'énergie disponible comme finie, et non suffisante pour satisfaire tous les agents. Un agent donné aura donc le choix entre consommer pour satisfaire son besoin, ce qui est soutenu par MR1, mais empêche les autres agents de satisfaire leur besoin, ce qui est défendu par MR3. S'il réduit sa consommation, et ne satisfait plus son besoin, il respecte MR3, mais trahit MR1. Il peut également compenser le manque en achetant auprès du réseau national, mais cela est interdit par MR4, et éventuellement MR2. Ainsi, chaque action implique une transgression d'au moins une règle morale, en supposant qu'il n'y ait pas assez d'énergie disponible. Dans notre simulateur, nous avons orienté la quantité d'énergie disponible de sorte que la plupart des pas de temps sont en conflit ; dans quelques pas de temps, il y a, à l'inverse, plus d'énergie que nécessaire, afin de permettre aux agents apprenants de faire des réserves, et de déployer des comportements plus complexes.

Notre simulateur permet donc de générer des dilemmes éthiques, au sens de Bonnemains : toute action implique un regret. Plus précisément, les agents devant choisir les paramètres (continus) des actions, plutôt que des actions discrètes, leur but est 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 4.1 ; nous résumons ses composants ci-après.

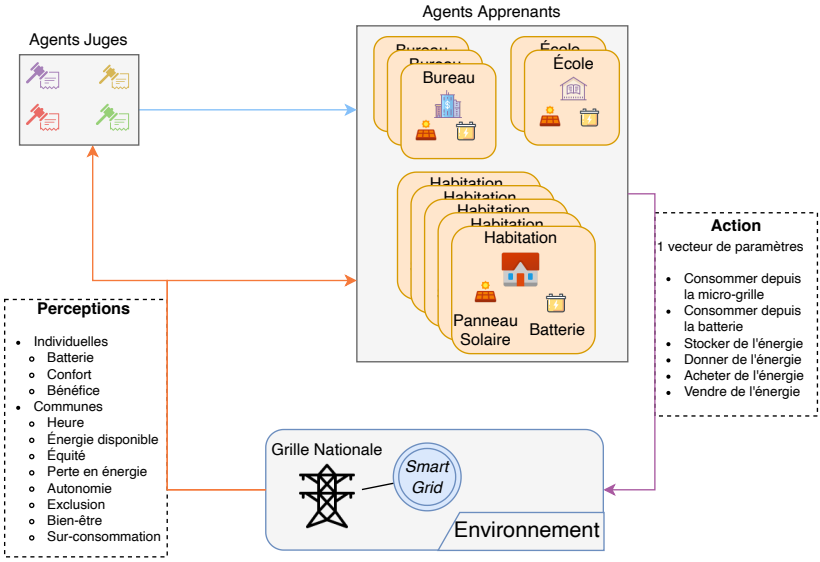

FIGURE 4.1. Schéma du simulateur de *Smart Grid*. Une micro-grille, liée à la grille nationale, contient plusieurs agents représentant des bâtiments.

4.2.1. *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.

Exemple 4.1. — Considérons un agent Habitation, et un agent École ; ils ont un besoin de respectivement 1 007 W et 122 784 W, à une certaine heure. L'Habitation peut consommer jusqu'à 2 500 W à chaque action (donc pas de temps), tandis que l'École peut consommer jusqu'à 205 000 W. Ces capacités d'action ont été choisies en fonction de la quantité d'énergie maximale dont peuvent avoir besoin les agents. Ainsi, ils peuvent choisir de consommer plus que ce dont ils n'ont besoin (mais il n'y aura certainement pas autant d'énergie la plupart du temps). Supposons que l'agent Habitation décide de consommer 806 W, tandis que l'agent École décide de consommer 98 227 W. Cela correspond, pour chacun d'eux, à environ 80 % de leur besoin. La courbe *flexible* du confort de l'Habitation fait que son confort sera évalué à 0,99. En revanche, l'École utilise une courbe *stricte*, qui, pour le même ratio consommation-besoin, évalue son confort à 0,89.

Nous utilisons un jeu de données publiques de consommation d'énergie[6] 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. L'utilisation de l'année entière introduit des difficultés pour les agents apprenants, car les différentes saisons impliquent des variations. Par exemple, nous consommons plus d'énergie en Hiver qu'en Été ; ils doivent donc apprendre à traiter ces variations. Les courbes de confort et le besoin en énergie par heure utilisés dans nos expérimentations sont visibles dans la figure 4.2.

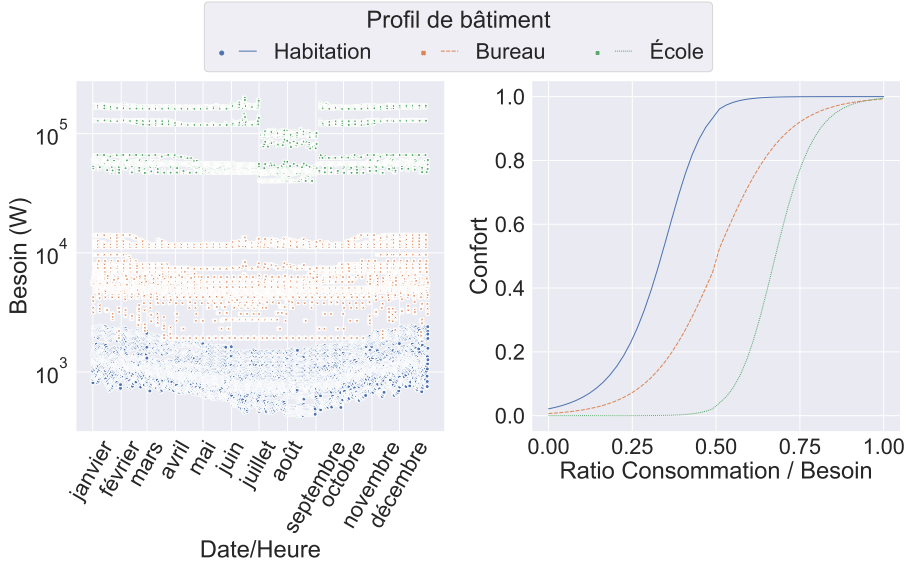

FIGURE 4.2. Besoins et confort pour chaque profil d'agent.

4.2.2. *Actions*

À chaque pas de temps, chaque agent apprenant effectue une action, vecteur de taille $k = 6$ contenant les 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.

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

4.2.3. *Perceptions*

Afin de prendre une décision, l'agent obtient l'état de l'environnement représenté par un vecteur de nombres réels de taille $g = 11$. 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.

4.2.4. *Récompenses*

Comme décrit dans notre modèle, les récompenses sont calculées à partir des jugements des agents juges. Ceux-ci sont implémentés en langage Jason [11] sur la plateforme JaCaMo [7], 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. 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 première propriété que nous cherchons à valider est le passage à l'échelle de notre approche, en terme de nombres d'agents apprenants. La comparaison entre les

expérimentations « Petit » et « Moyen » n'a pas montré de réelle différence entre les moyennes des scores des simulations (T-Test, p-value = 0,83), ce qui indique que les scores sont comparables, sur les tailles d'environnement considérées, et notre approche passe à l'échelle, bien que le temps d'exécution soit naturellement bien plus long. Le temps d'exécution varie à peu près de manière linéaire selon le nombre d'agents.

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 n'est pas aussi intéressant que la capacité d'en ajouter (le scénario « Incrémental »).

La deuxième propriété que nous cherchons à valider est l'apprentissage de multiples valeurs morales, en même temps. La figure 5.1 montre que, dans le scénario « Défaut », la récompense moyenne augmente au fur et à mesure de la simulation, bien que l'augmentation pourrait être plus importante ; l'approche hybride proposée est donc efficace et les agents apprenants sont capables de se conformer aux règles données par les agents juges. 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.

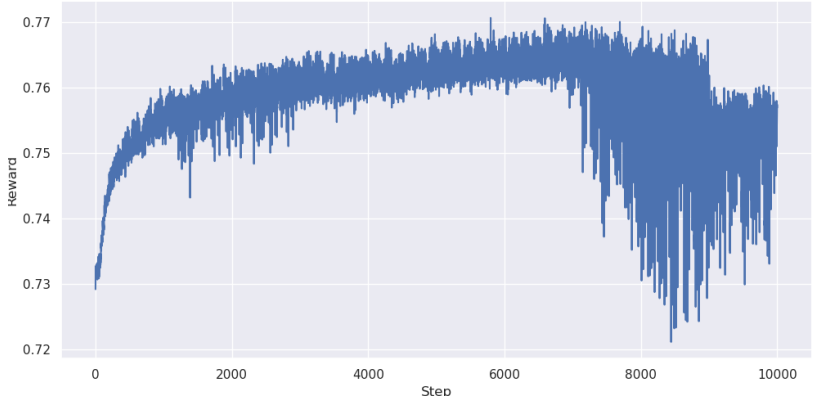

FIGURE 5.1. Récompense moyenne pour tous les agents apprenants, sur chaque pas de temps, dans les simulations « Petit - Annuel - Défaut ».

La troisième propriété que nous cherchons à valider est la capacité d'adaptation face au changement, et en particulier à l'addition de nouvelles valeurs morales. La figure 5.2 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.

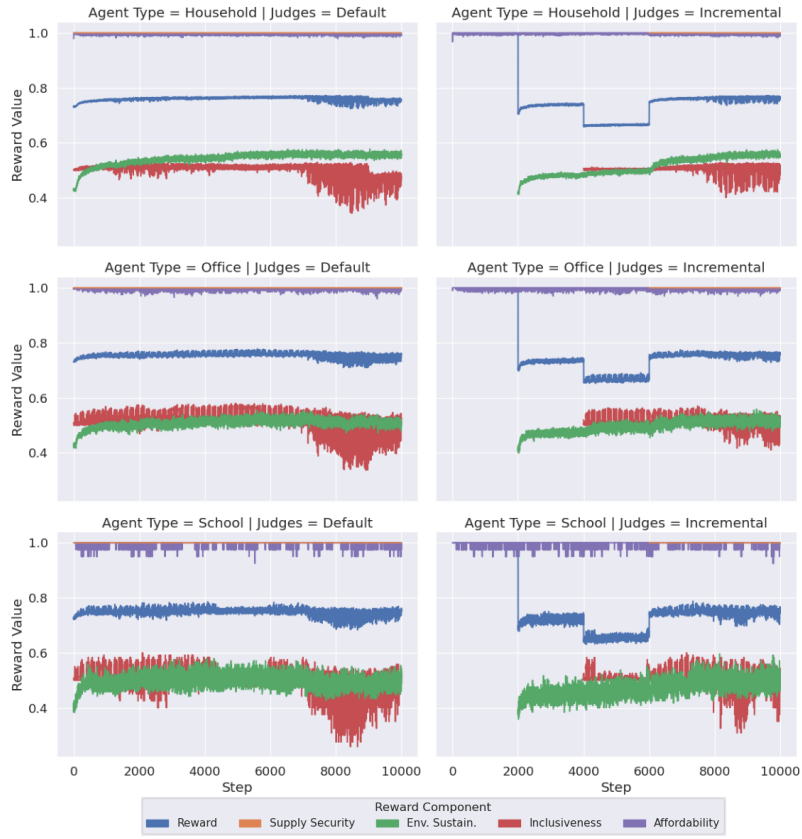

FIGURE 5.2. Comparaison entre les récompenses individuelles de chaque agent apprenant, au fil des pas de temps, moyennées sur les simulations « Petit - Annuel - Défaut » et « Petit - Annuel - Incrémental », et sur les agents de même profil. La courbe « Récompense » est la moyenne des quatre composantes.

6. Discussion

Pour rappel, notre contribution est une nouvelle approche 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. Cette approche se base sur une architecture hybride, utilisant des agents juges symboliques comme sources de récompenses pour des 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 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 « Decrémental » 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 [31] : 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* [26].

Les précédents cas d'applications pour des « agents éthiques » étaient limités à des actions discrètes (*e.g.*, dilemmes tel que le Dilemme du Tramway [28], robot accompagnant [3], soldats robots [4], gestion autonome d'actifs en bourse [16]). Ces travaux sont importants, mais il existe de nombreuses situations requérant de plus fines actions ; il est ainsi important de proposer et d'expérimenter sur des environnements avec actions continues, tel que le simulateur utilisé ici.

L'utilisation d'apprentissage par renforcement dans notre approche permet aux agents apprenants d'automatiquement apprendre s'il est intéressant de faire une action avec une récompense légèrement négative, si cela permet d'obtenir plus tard une récompense extrêmement positive. Par exemple, stocker ou acheter de l'énergie quand elle est disponible, afin de pouvoir l'utiliser en temps de pénurie sans empêcher les autres voisins de consommer. Les agents peuvent donc avoir un raisonnement sur le long-terme (à plus d'un pas de temps), tandis que les concepteurs du système n'ont besoin de spécifier que le jugement sur le pas de temps actuel, ce qui simplifie leur travail.

Certains travaux proposent d'utiliser des méthodes de vérification formelle pour garantir la conformité aux règles morales dans n'importe quelle situation identifiée [13, 9]. 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 [19]. Une perspective intéressante est donc la vérification formelle de comportements appris par notre méthode hybride. Les agents juges symboliques, par le biais des récompenses qu'ils fournissent, guident et orientent les agents apprenants. Ils peuvent donc être mis à profit par de telles techniques de vérification.

Bien que nous ne l'ayons pas mentionné dans le modèle, par souci de clarté, notre approche permet d'agréger à la fois des jugements symboliques et des récompenses numériques. Ceci est particulièrement utile pour un objectif plus « technique », facilement exprimable par une fonction mathématique, *e.g.*, « Le ratio production/consommation doit s'approcher de 1 ». Toutefois, n'ayant pas évalué cet aspect dans nos expérimentations, d'autres travaux seront nécessaires pour approfondir cet aspect.

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, mais seulement par les agents juges.

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. Les résultats soulignent également des problèmes de calibration, notamment au niveau de ces règles morales, telles que les règles correspondant à la valeur d'Inclusivité, qui ont été plus difficiles à apprendre que les autres. Cette calibration doit être améliorée, notamment par la formalisation des règles elle-même, mais également par la transformation des jugements en récompense numérique.

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 une gestion plus complexe est sans doute nécessaire. Par exemple, on peut vouloir hiérarchiser

les valeurs morales, et ainsi établir une priorité entre les jugements selon le contexte. Imaginons le cas où un hôpital est en manque crucial d'énergie, la règle jugeant l'achat d'énergie comme immorale selon la valeur morale de Viabilité Environnementale peut avoir une priorité plus faible qu'une autre valeur morale, telle que l'Assurance de Confort, selon laquelle l'achat d'énergie est une action morale. Une moyenne pondérée pourrait être utilisée pour atteindre ce résultat ; cependant, cette méthode ne permet pas (facilement) des priorités dépendantes de la situation. Il peut en effet exister d'autres situations où, à l'inverse, nous souhaitons que la Viabilité Environnementale soit prioritaire sur l'Assurance de Confort. Afin de gérer cette approche contextuelle de l'éthique, un niveau d'interaction entre les agents juges pourrait être introduit, tel qu'un mécanisme d'argumentation, ou encore un méta-juge chargé de déterminer la récompense à partir des jugements.

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.

Bibliographie

[1] C. Allen, I. Smit & W. Wallach, « Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches », *Ethics and Information Technology* **7** (2005), nº 3, p. 149-155.

[2] M. Anderson & S. L. Anderson, « Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm », in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[3] M. Anderson, S. L. Anderson & V. Berenz, « A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm », *Proc. IEEE* **107** (2019), nº 3, p. 526-540.

[4] R. C. Arkin, P. D. Ulam & B. Duncan, « An ethical governor for constraining lethal action in an autonomous system », Tech. report, Georgia Institute of Technology, 2009.

[5] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon & I. Rahwan, « The moral machine experiment », *Nature* **563** (2018), nº 7729, p. 59-64.

[6] A. Boijmans, « The Acceptability of Decentralized Energy Systems », master thesis, Delft University of Technology, 2019.

[7] O. Boissier, R. Bordini, F. Hübner, Jomi & A. Ricci, *Multi-Agent Oriented Programming: Programming Multi-Agent Systems Using JaCaMo*, The MIT Press, 2020.

[8] V. Bonnemains, « Formal ethical reasoning and dilemma identification in a human-artificial agent system », Thèse, Institut supérieur de l'aéronautique et de l'espace, Toulouse, France, 2019.

[9] G. Bonnet, B. Mermet & G. Simon, « Vérification formelle et éthique dans les SMA », in *Systèmes Multi-Agents et simulation – Vingt-quatrièmes journées francophones sur les systèmes multi-agents, JFSMA 16, Saint-Martin-du-Vivier (Rouen), France, Octobre 5-7, 2016* (F. Michel & J. Saunier, éds.), Cépaduès Éditions, 2016, p. 139-148.

[10] R. H. Bordini, A. El Fallah Seghrouchni, K. Hindriks, B. Logan & A. Ricci, « Agent programming in the cognitive era », *Autonomous Agents and Multi-Agent Systems* **34** (2020), nº 2, p. 1-31.

[11] R. H. Bordini, J. F. Hübner & M. Wooldridge, *Programming multi-agent systems in AgentSpeak using Jason*, vol. 8, John Wiley & Sons, 2007.

[12] M. Bosello & A. 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*, Springer, 2019, p. 175-194.

[13] P. Bremner, L. A. Dennis, M. Fisher & A. F. Winfield, « On proactive, transparent, and verifiable ethical reasoning for robots », *Proceedings of the IEEE* **107** (2019), n° 3, p. 541-561.

[14] R. Chaput, O. Boissier, M. Guillermin & S. Hassas, « Apprentissage adaptatif de comportements éthiques », in *Architectures multi-agents pour la simulation de systeèmes complexes - Vingt-huitième journées francophones sur les systèmes multi-agents, JFSMA 2020, Angers, France, June 29 - July 3, 2020* (N. Sabouret, éd.), Cépaduès, 2020.

[15] N. Cointe, G. Bonnet & O. Boissier, « Jugement éthique dans les systèmes multi-agents », in *Systèmes Multi-Agents et simulation – Vingt-quatrièmes journées francophones sur les systèmes multi-agents, JFSMA 16, Saint-Martin-du-Vivier (Rouen), France, Octobre 5-7, 2016* (F. Michel & J. Saunier, éds.), Cépaduès Éditions, 2016, p. 149-158.

[16] ———, « Multi-agent based ethical asset management », in *1st Workshop on Ethics in the Design of Intelligent Agents*, 2016, p. 52-57.

[17] L. A. Dennis & M. Fisher, « Practical Challenges in Explicit Ethical Machine Reasoning », in *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2018, Fort Lauderdale, Florida, USA, January 3-5, 2018*, 2018.

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

[19] N. Fulton & A. Platzer, « Safe reinforcement learning via formal methods : Toward safe control through proof and learning », in *Proceedings of the AAAI Conference on Artificial Intelligence, volume 1*, vol. 32, 2018.

[20] A. R. Honarvar & N. Ghasem-Aghaee, « An artificial neural network approach for creating an ethical artificial agent », in *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA)*, IEEE, 2009, p. 290-295.

[21] T. Kohonen, « Essentials of the self-organizing map », *Neural Networks* **37** (2013), p. 52-65.

[22] J. H. Moor, « The nature, importance, and difficulty of machine ethics », *IEEE intelligent systems* **21** (2006), n° 4, p. 18-21.

[23] V. Nallur, « Landscape of machine implemented ethics », *Science and engineering ethics* **26** (2020), n° 5, p. 2381-2399.

[24] S. Nason & J. E. Laird, « Soar-RL: Integrating reinforcement learning with Soar », *Cognitive Systems Research* **6** (2005), n° 1, p. 51-59.

[25] N. P. Rougier & Y. Boniface, « Dynamic self-organising map », *Neurocomputing* **74** (2011), n° 11, p. 1840-1847.

[26] D. Schiff, J. Biddle, J. Borenstein & K. 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*, 2020, p. 153-158.

[27] R. S. Sutton & A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[28] J. J. Thomson, « Killing, letting die, and the trolley problem », *The Monist* **59** (1976), n° 2, p. 204-217.

[29] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen & A. Bernstein, « Implementations in Machine Ethics : A Survey », *ACM Comput. Surv.* **53** (2021), n° 6, article no. 132 (38 pages).

[30] C. J. C. H. Watkins & P. Dayan, « Q-Learning », *Machine Learning* **8** (1992), n° 3, p. 279-292.

[31] Y.-H. Wu & S.-D. Lin, « A low-cost ethics shaping approach for designing reinforcement learning agents », *Proceedings of the AAAI Conference on Artificial Intelligence* **32** (2018), n° 1, p. 1687-1694.

[32] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser & Q. Yang, « Building Ethics into Artificial Intelligence », in *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden), IJCAI'18, AAAI Press, 2018, p. 5527-5533.

Abstract. — To answer the need to imbue Artificial Intelligence algorithms with ethical considerations, this article propose a method combining reasoning and learning, where judging agents evaluate the ethics of learning agents' behavior. This separation offers several advantages: co-construction between agents and humans; judges more accessible for non-experts humans; richer feedback by using multiple judgments. 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.

Manuscrit reçu le 18 mars 2022, accepté le 23 novembre 2022.