Coopération entre agents autonomes fondée sur l'éthique

N. Cointe[a,b] nicolas.cointe@emse.fr
G. Bonnet[b] gregory.bonnet@unicaen.fr
O. Boissier[a] olivier.boissier@emse.fr

[a]Université de Lyon, MINES Saint-Etienne, CNRS, Laboratoire Hubert Curien, UMR 5516

[b]Normandie Université, GREYC CNRS UMR 6072, F-14032 Caen, France

Résumé

Dans le domaine de la décision autonome, la prise en compte de la dimension éthique des décisions est généralement centrée sur l'agent, en laissant de côté sa dimension sociale. Or, l'éthique semble être une notion centrale influençant les interactions sociales entre individus. Dans cet article, nous proposons un modèle permettant à des agents de se construire une image du comportement éthique et moral des autres afin de le prendre en compte dans leurs interactions. Fondé sur une approche rationaliste et explicite, ce modèle distingue l'éthique de la moralité et permet d'aboutir à l'établissement ou non d'une relation de confiance. Nous illustrons ces fonctionnalités dans une preuve de concept dans le domaine de la gestion d'actifs financiers implémentée à l'aide de la plateforme JaCaMo.

Mots-clés : *Architecture d'agent, Modèles de comportement agent, Éthique computationelle*

Abstract

In decision theory, dealing with ethics is mainly considered in an agent-centered perspective, letting aside the social dimension of multi-agent systems. However, ethics seems to be a central notion when considering interactions among agents. In this paper, we propose a model for ethics-based cooperation. Each agent uses an ethical judgment process based on a rationalist and explicit approach to compute images of the other agents' ethical behavior. From these images of the other agents' ethics, the judging agent computes trust used to interact with the judged agents. We illustrate these functionalities in an asset management scenario with a proof-of-concept implemented in the JaCaMo Multi-Agent Platform.

Keywords: *Agent architecture, Agent's model of behavior, Computer ethics*

1 Introduction

L'usage croissant des agents autonomes dans un grand nombre de domaines tels que la santé, les transports ou la finance ajoute aux habituels problèmes de l'optimalité de leurs décisions, la problématique de la prise en compte des dimensions morales et éthiques de leur choix dans leur raisonnement. Par exemple, dans le cas de la gestion éthique d'actifs financiers, un grand nombre de modèles ont été suggérés pour prendre des décisions profitables [2], mais bien peu de solutions permettent à un agent de juger de la conformité de ses investissements avec les valeurs morales et principes éthiques des investisseurs qu'il représente. De plus, l'hétérogénéité de ces éléments soulève de nombreux problèmes lorsque les agents ont besoin de collaborer avec d'autres agents en respectant leur propre éthique.

Par exemple, comment un agent peut-il évaluer la conformité du comportement des autres agents à une éthique qui lui est propre ? Comment un tel agent peut-il décider d'accorder sa confiance à un autre en se fondant sur cette évaluation ? L'objectif de cet article est de répondre à de telles questions en proposant un cadre permettant de construire des mécanismes de coopération entre agents reposant sur la construction d'images du comportement d'autrui et de confiance par agrégation de jugements éthiques.

Pour ce faire, nous incorporons le processus de jugement éthique proposé dans [14] au sein d'une architecture BDI pour permettre à un agent de juger du comportement des autres. La conception de tels agents autonomes suppose qu'un concepteur ou un utilisateur décrive l'éthique et la morale employés. En utilisant ce processus, nous proposons ensuite de construire une image de l'autre en évaluant et en agrégeant ces jugements. Ensuite, nous proposons de permettre à un agent de prendre en compte cette image pour décider de faire confiance à un autre afin d'envisager des actions de coopération. En-

fin, nous instancions ce modèle dans une application de gestion d'actifs financiers et montrons son usage dans une preuve de concept implémentée à l'aide de la plate-forme JaCaMo [7].

Cet article est organisé de la manière suivante. La Sec. 2 introduit et décrit les modèles d'éthique computationnelle et de confiance employés dans ce travail. Ensuite, la Sec. 3 montre comment le jugement éthique peut être utilisé pour se représenter l'éthique du comportement de l'autre. Puis, la Sec. 4 présente la construction et l'utilisation de la confiance. Enfin, nous illustrons l'usage de ce travail dans une preuve de concept en Sec. 5 avant de conclure.

2 Concepts principaux

Nous introduisons dans cette section les concepts nécessaires à la coopération fondée sur l'éthique dans les systèmes multi-agents. La Sec. 2.1 montre comment le concept de confiance peut guider les interactions et la coopération entre agents. La Sec. 2.2 introduit l'éthique et montre comment elle peut mener à la confiance. Enfin, la Sec. 2.3 propose une synthèse des éléments nécessaires à la définition de la coopération fondée sur l'éthique. Cette synthèse constitue l'ossature de la proposition décrite dans ce papier.

2.1 Confiance dans les SMA

Dans les systèmes décentralisés et ouverts, la confiance est un moyen de coexister et d'interagir avec des agents inconnus et à la fiabilité incertaine [11, 16, 31]. La confiance permet aux agents d'évaluer les interactions observées ou effectuées pour décider si collaborer avec un agent est a priori acceptable. Cette notion d'acceptation signifie que le comportement de l'agent observé est considéré comme bon et fiable du point de vue des critères de l'agent observateur.

De nombreuses définitions de la confiance existent mais, comme le fait [11], nous considérons la *confiance* comme *une disposition à coopérer avec un individu de confiance*. Ainsi, elle peut être utilisée comme une condition pour effectuer certaines actions de délégation, de partage de ressources ou d'information, ou toute forme de coopération. Pour construire cette confiance, les agents commencent par se construire une image de l'agent observé [16].

Si [16] définissent une *image* comme *une croyance qualifiant le sujet de bon ou mauvais selon son comportement* dans le cadre des systèmes de confiance, nous préférons la définir comme *une croyance qualifiant le sujet de conforme ou non selon l'adéquation de son comportement à un ensemble de règles* afin de lever toute ambiguïté sur les termes *bien* et *mal* au sens moral. Dans la littérature, les images sont agrégées à partir de l'expérience, c'est-à-dire l'observation des comportements et de leurs conséquences. Nous pouvons distinguer deux types d'approches : (1) les images statistiques [1, 9, 17, 25, 35] où l'image est une agrégation quantitative d'appréciations d'interactions passées. Cette agrégation estime la tendance d'un agent à agir conformément à des critères. Cela peut être représenté par des réseaux bayésiens, des lois de probabilité bêta, des ensembles flous, des fonctions de Dempster-Shafer et d'autres formalismes quantitatifs ; (2) les images logiques [10, 11, 27, 34] où l'image est un état mental lié à toute action de coopération produite par interaction. Une image persistante permet d'inférer des croyances sur la confiance pouvant être utilisées comme préconditions pour des actions de coopération.

Un agent peut manquer d'observations pour construire une image correcte de l'agent jugé. Une manière de traiter ce problème est d'utiliser la réputation [23, 30]. Cela consiste en l'utilisation de l'image qu'un tiers a de l'agent jugé afin d'obtenir un point de vue collectif sur le jugé. Le choix de cet agent tiers peut lui-même dépendre de l'image que l'agent jugé a des autres Ainsi, les images individuelles et la réputation peuvent être utilisées conjointement pour décider d'établir une confiance [31]. De manière générale, la confiance est dynamique car elle change en fonction de l'évolution des images et des réputations.

2.2 Comportements éthiques

Dans cet article, nous nous intéressons à la construction d'images de l'éthique du comportement des autres agents. En raison de l'absence de définitions formelles dans la littérature, nous admettons la définition suivante [33] : les connaissances de l'agent sont réparties en deux composantes, la *théorie du bien* (ou morale) et la *théorie du juste* (ou éthique). Bien que cette distinction soit discutable en raison de la grande diversité de théories contradictoires en sciences humaines, nous estimons que ces définitions fournissent un cadre intéressant pour représenter la morale et l'éthique.

Une *théorie du bien* est un ensemble de règles et valeurs morales qui permettent d'évaluer la moralité (le caractère bon ou mauvais) d'un com-

portement. Les règles morales attribuent des valuations morales à des comportements (par exemple "Mentir est mal" ou "Être honnête est bien"), et les valeurs permettent de qualifier des actions de manière plus abstraites (par exemple "Il est honnête de dire ce que l'on pense").

Une *théorie du juste* utilise un ensemble de principes éthiques pour reconnaître un choix juste, ou au moins acceptable. Les philosophes ont proposé un ensemble varié de principes éthiques, tels que l'impératif catégorique de Kant[22] ou la doctrine du double effet de Saint Thomas D'Aquin[26]. Par exemple même s'il est immoral de voler, (au regard des commandements divins), plusieurs philosophes admettent qu'il est acceptable pour des gens affamés de voler de la nourriture (au regard de la doctrine du double effet).

Comme la moralité d'un comportement repose sur une théorie du bien, son caractère éthique repose dans la conciliation des désirs, de la morale et des capacités de l'agent au regard d'une théorie du juste [28]. Ainsi, être moral ou éthique caractérise un comportement dans un contexte donné, tout comme être fiable caractérise un comportement dans un système de confiance. Par conséquent, il peut être intéressant de définir une notion de confiance dans la moralité ou le caractère éthique d'un comportement qui pourrait venir renforcer une coopération.

2.3 Une coopération fondée sur l'éthique

Plusieurs travaux prenant en compte la dimension éthique du comportement d'agents autonomes se focalisent sur la modélisation d'un raisonnement moral [6, 19, 20, 32] comme une traduction directe de théories bien connues, ou la modélisation du comportement moral en général [5, 24]. Toutefois, ces travaux ne permettent ni de représenter des valeurs morales, ni d'employer plusieurs principes éthiques dans un raisonnement. D'autres travaux traitent de l'architecture des agents éthiques. Parmi celles-ci, les *architectures éthiques implicites* [3, 4] proposent soit de concevoir des agents en implémentant dans tout état possible des moyens d'empêcher des actions non éthiques, soit un apprentissage supervisé de l'éthique. D'un autre côté, les *architectures éthiques cognitives* [12, 13, 14, 15] consistent en une représentation totalement explicite de chaque composant de l'agent, des croyances classiques (informations sur l'environnement et les autres agents), désirs (objectifs de l'agent) et intentions (actions choisies) à des concepts tels que des heuristiques ou des simulations émotionnelles. Toutefois, ces approches ne tiennent pas compte de la dimension collective de ces systèmes, excepté [29] qui considère la morale comme un élément des sociétés d'agents.

Plus précisément, l'architecture donnée en [14] propose une séparation claire entre théorie du bien et théorie du juste, et propose des croyances portant sur des composants (principes éthiques, valeurs et règles morales, etc.). De plus, l'architecture proposée par [29] permet – sans en proposer une version opérationnelle – de voir des faits moraux (des jugements sur les autres ou des blâmes par exemple) comme des croyances pouvant être employées dans les décisions des agents.

Afin de construire une coopération fondée sur l'éthique, nous avons besoin d'un modèle opérationnel de jugement éthique tel que celui proposé en [14]. Inspiré de [29], nous réutilisons et étendons ce modèle avec des croyances sur les images morales et éthiques des autres agents. Ces images sont ensuite utilisées pour construire des relations de confiance permettant d'influencer la coopération entre agents.

3 Processus de jugement

Nous présentons en Sec. 3.1 le processus de jugement décrit en [14]. Nous montrons ensuite comment un agent peut l'employer pour construire sa propre représentation qualitative de l'éthique (voir Sec. 3.2) et de la morale (voir Sec. 3.3) des autres agents, au regard des croyances de l'agent juge sur la *connaissance du bien* et la *connaissance du juste*.

3.1 Jugement des autres agents

Nous considérons le processus de jugement introduit dans [14] répondant à nos besoins énoncés en Sec. 2. Ce processus fournit une évaluation des actions connues de l'agent en matière de conformité à un ensemble de connaissances données.

Comme le montre la Figure 1, le processus de jugement est organisé en trois parties : (*i*) le processus de reconnaissance de situation et d'évaluation, (*ii*) le processus moral et (*iii*) le processus éthique. Comme ce processus de jugement peut raisonner de manière interchangeable sur les données d'autres agents, nous indiçons dans la suite la totalité des connaissances par l'identifiant de l'agent dont elles proviennent $a_i \in \mathbb{A}$ (par exemple $\mathcal{A}_{r_{a_i}}$) avec $\mathbb{A}$ l'ensemble des

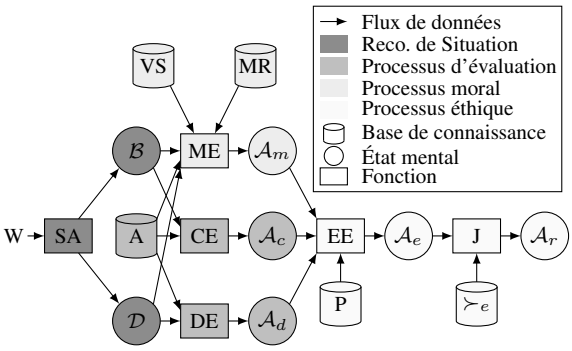

FIGURE 1 – Processus de jugement éthique [14]

agents. Il est ainsi possible pour un agent d'utiliser ses propres connaissances pour juger de son propre comportement ou de celui d'un autre, ou bien d'utiliser tout ou partie des connaissances d'un autre pour le juger.

Processus de reconnaissance et d'évaluation. Dans ce modèle, les actions sont décrites sous forme d'actions et de conséquences exprimées en termes de changements de désirs et des croyances. Le *processus d'évaluation* évalue alors l'ensemble des actions $\mathcal{A}_{a_i}$ qu'il considère comme désirable ($\mathcal{A}_{d_{a_i}}$) et réalisable ($\mathcal{A}_{c_{a_i}}$) du point de vue des connaissances de a_i conformément aux désirs $\mathcal{D}_{a_i}$ et croyances $\mathcal{B}_{a_i}$. $\mathcal{B}_{a_i}$ et $\mathcal{D}_{a_i}$ sont produits par le processus de *reconnaissance de situation* SA à partir de la perception de l'état courant. Ici, DE et CE sont respectivement les fonctions d'*évaluation de la désirabilité* et d'*évaluation de la faisabilité*. Par la suite, nous appelons *connaissance contextuelle* de a_i l'union de $\mathcal{B}_{a_i}$ et $\mathcal{D}_{a_i}$ que nous notons CK_{a_i}.

Processus moral. Le *processus moral* produit l'ensemble des actions moralement évaluées $\mathcal{A}_{m_{a_i}}$ au regard des données contextuelles CK_{a_i} de l'agent a_i, des actions A_{a_i}, des supports de valeurs VS_{a_i} et des règles morales MR_{a_i}. Ces actions sont celles qui, dans la situation décrite par CK_{a_i}, promeuvent ou trahissent les valeurs morales de VS_{a_i} et se trouvent évaluées par les règles morales de MR_{a_i}. Un *support de valeur* est un couple $\langle s, v\rangle \in VS_{a_i}$ où $v \in \mathcal{O}_v$ est une valeur morale et $s = \langle \alpha, w\rangle$ est le support de cette valeur morale où $\alpha \in A_{a_i}$, $w \subset \mathcal{B}ai \cup \mathcal{D}_{a_i}$. $\mathcal{O}_v$ est l'ensemble des valeurs morales utilisées dans le système [1]. Une *règle morale* est un tuple $\langle w, o, m\rangle \in MR_{a_i}$. La situation $w \in 2^{CK_{a_i}}$ est une conjonction de croyances et désirs. o est l'objet de la règle avec $o = \langle \alpha, v\rangle$ où α est une action ($\alpha \in A_{a_i}$) et v est une valeur morale ($v \in \mathcal{O}_v$). Enfin, m est la valuation morale ($m \in \mathcal{O}_m$). Par exemple, $\mathcal{O}_m = \{\texttt{moral}, \texttt{amoral}, \texttt{immoral}\}$ permet d'associer une parmi trois valuations morales à o lorsque w est vrai. Notons qu'un ordre total doit être défini sur $\mathcal{O}_m$ (par exemple `moral` est une valuation supérieure à `amoral`, qui est supérieure à `immoral`). Par la suite, les connaissances sur les règles morales MR_{a_i}, supports de valeurs VS_{a_i} et valeurs $\mathcal{O}_v$, utilisées dans le processus moral de l'agent sont appelés *connaissances morales* et sont notées GK_{a_i}.

Processus éthique. Enfin, le processus éthique évalue l'action juste $\mathcal{A}_{r_{a_i}}$ à partir de l'ensemble des action possibles $\mathcal{A}_{c_{a_i}}$, désirables $\mathcal{A}_{d_{a_i}}$ et morales $\mathcal{A}_{m_{a_i}}$ par rapport à un ensemble de *principes éthiques* P_{a_i} pour concilier ces ensembles d'actions conformement à un ensemble de relations de préférences éthiques $\succ_{e_{a_i}} \subseteq P_{a_i} \times P_{a_i}$. Un *principe éthique* $p \in P_{a_i}$ est une fonction qui évalue s'il est juste ou non d'exécuter une action dans une situation donnée au regard d'une théorie philosophique. Cette évaluation est exprimée au travers d'évaluations des actions de $\mathcal{A}_{c_{a_i}}$, $\mathcal{A}_{d_{a_i}}$ et $\mathcal{A}_{m_{a_i}}$ dans une situation décrite par CK_{a_i}. Un principe est défini par $p : 2^{\mathcal{A}_{a_i}} \times 2^{\mathcal{B}_{a_i}} \times 2^{\mathcal{D}_{a_i}} \times 2^{MR_{a_i}} \times 2^{V_{a_i}} \rightarrow \{\top, \bot\}$. Étant donné un ensemble d'actions évaluées issues de la fonction d'évaluation éthique EE utilisant les principes éthiques, le jugement J est la dernière étape pour choisir l'action juste à effectuer, en considérant l'ensemble des préférences éthiques $\succ_{e_{a_i}}$ définissant un ordre total sur les principes éthiques. Dans ce processus de jugement, les actions justes sont celles qui satisfont les principes éthiques préférés selon un critère lexicographique. Par la suite, les principes éthiques P_{a_i} et préférences $\succ_{e_{a_i}}$ sont appelés *connaissance éthique* et sont notés RK_{a_i}.

3.2 Juger de la conformité éthique

Nous étendons maintenant le processus de jugement précédemment présenté pour juger l'éthique et la moralité d'un comportement observé attribué à un agent a_j sur une période de temps de t_0 à t. Inspiré de [29] qui considère les croyances sur des faits moraux, le processus de jugement produit ici des croyances (`ethical_conf`, `moral_conf`) informant de la conformité d'une action à des principes éthiques ou des règles et valeurs morales. Avant de définir ces croyances, nous définissons le comportement d'un agent de la façon suivante :

1. Notons que dans [14] les valeurs morales et valuations morales sont partagées entre les agents du système : les agents se distinguent par les règles morales et les éléments de leur processus éthique.

Définition 1 (Comportement) *Le* comportement $b_{a_j,[t_0,t]}$ *d'un agent* a_j *sur l'intervalle* $[t_0, t]$ *est l'ensemble des actions* α_k *que* a_j *a exécuté entre* t_0 *et* t *tel que* $0 \leqslant t_0 \leqslant t$*.*

$b_{a_j,[t_0,t]} = \{\alpha_k \in A : \exists t' \in [t_0, t] \text{ s.t. } \texttt{done}(a_j, \alpha_k, t')\}$

où $A = \bigcup_{a_i=a_1}^{a_n} \mathcal{A}_{a_i}$ *est l'ensemble des actions disponibles dans un système à* n *agents et* $\texttt{done}(a_j, \alpha_i, t')$ *un prédicat signifiant que l'action* α_i *a été exécutée*[2] *par* a_j *à l'instant* t'*.*

Un agent a_i peut juger la conformité d'une action α_k exécutée par un autre agent a_j conformément à sa connaissance morale et éthique.

Définition 2 (Conformité éthique) *Une action* α_k *est* éthiquement conforme *par rapport aux connaissances contextuelles (*CK_{a_i}*), connaissances morales* GK_{a_i} *et connaissances éthiques* RK_{a_i} *d'un agent* a_i *au temps* t' *si et seulement si* α_k *est dans l'ensemble des actions justes* $\alpha_k \in \mathcal{A}_{r_{a_i}}$ *évaluées par le jugement éthique* J_{a_i} *de l'agent juge* a_i *en se fondant sur* $[CK_{a_i}, GK_{a_i}, RK_{a_i}]$ *à l'instant* t'*. Une telle action produit une croyance notée :*

$$\texttt{ethical\_conf}(\alpha_k, [CK_{a_i}, GK_{a_i}, RK_{a_i}], t')$$

Notons que la conformité éthique d'une action peut être appliquée aux actions de l'agent juge ou celles exécutées par un autre agent et observées par l'agent juge. Cette conformité éthique peut être évaluée par rapport aux connaissances contextuelles, morales et éthiques de l'agent juge, ou bien celles d'un autre agent si l'agent juge peut disposer d'une représentation de celles-ci. Finalement, la conformité éthique est employée pour produire l'ensemble EC^+ des actions éthiquement conformes (resp. l'ensemble EC^- de celles qui ne sont pas éthiquement conformes) du comportement observé $b_{a_j,[t_0,t]}$ de l'agent jugé a_j entre t_0 et t (voir Fig. 2).

Ces deux ensembles fournissent une information sur le comportement de l'agent jugé et sa conformité avec les connaissances employées pour le juger. Néanmoins, il est n'est pas possible en l'état de savoir pour quelle raison un comportement ne serait pas éthique. De fait, les raisons peuvent être une différence de connaissance contextuelle, morale ou éthique. Par la suite, nous notons $EC_{a_j,[t_0,t]} = EC^+_{a_j,[t_0,t]} \cup EC^-_{a_j,[t_0,t]}$.

2. Un comportement peut tenir compte de la concurrence : plusieurs actions peuvent être exécutées à un même instant.

3.3 Juger de la conformité morale

Dans le modèle de jugement, si le jugement éthique indique une conformité ou non avec un principe, les règles morales indiquent une conformité par rapport à un ensemble de valuations morales. Ainsi, l'évaluation de la conformité morale d'une action à une règle morale donnée se fait en comparant la valuation morale associée à l'action par la règle à une valuation seuil $mt \in MV$.

Définition 3 (Conformité morale) *Une action* α_k *est* moralement conforme *par rapport aux connaissances contextuelles (*CK_{a_i}*), connaissances morales* GK_{a_i} *et connaissances éthiques* RK_{a_i} *d'un agent* a_i *au temps* t' *au regard de la règle* $mr \in MR_{a_i}$ *et un seuil* $mt \in MV_{a_i}$ *si et seulement si* α_k *appartient à l'ensemble des actions morales* $A_{m_{a_i}}$ *et se trouve affectée d'une valuation morale supérieure ou égale à* mt*, au regard de l'action morale* mr *et des connaissances* CK_{a_i}*,* GK_{a_i} *et* RK_{a_i} *à l'instant* t'*. Une telle action produit une croyance notée :*

$$\texttt{moral\_conf}(\alpha_k, [CK_{a_i}, GK_{a_i}, RK_{a_i}], mr, mt, t')$$

De même que pour la conformité éthique, nous utilisons la conformité morale pour générer l'ensemble MC^+ (resp. MC^-) des actions moralement conformes (resp. moralement non conformes) du comportement observé $b_{a_j,[t_0,t]}$ de l'agent jugé a_j entre t_0 et t au regard de la règle mr et d'un seuil moral mt (voir Fig. 2).

Nous généralisons cette équation de la conformité morale au regard d'une règle en la transposant à un ensemble de règles, en supposant la définition préalable d'ensembles de règles morales $R \subseteq MR_{a_i}$. Un ensemble R peut, par exemple représenter l'ensemble des règles en rapport avec l'expression d'une valeur, une situation spécifique, etc. Par la suite, nous notons $MC_{a_j,mr,mt,[t_0,t]} = MC^+_{a_j,mr,mt,[t_0,t]} \cup MC^-_{a_j,mr,mt,[t_0,t]}$.

4 Construction de la confiance

Dans cette section nous utilisons les résultats de jugements définis dans la section précédente pour construire l'image des autres agents (cf. Sec. 4.1). Nous montrons ensuite comment ces images sont employées pour établir une relation de confiance (cf. Sec. 4.2). La Sec. 4.3 montre enfin comment cette relation influence le comportement de l'agent.

$$EC^{+}_{a_j,[t_0,t]}=\{\alpha_k{:}\exists t'\in[t_0,t] \text{ s.t. } \texttt{done}(a_j,\alpha_i,t') \wedge \texttt{ethical\_conf}(\alpha_k,[CK_{a_i},GK_{a_i},RK_{a_i}],t')\}$$

$$EC^{-}_{a_j,[t_0,t]}=\{\alpha_k{:}\exists t'\in[t_0,t] \text{ s.t. } \texttt{done}(a_j,\alpha_i,t') \wedge \neg\texttt{ethical\_conf}(\alpha_k,[CK_{a_i},GK_{a_i},RK_{a_i}],t')\}$$

$$MC^{+}_{a_j,mr,mt,[t_0,t]}=\{\alpha_k{:}\exists t'\in[t_0,t] \text{ s.t. } \texttt{done}(a_j,\alpha_i,t') \wedge \texttt{moral\_conf}(\alpha_k,[CK_{a_i},GK_{a_i},RK_{a_i}],mr,mt,t')\}$$

$$MC^{-}_{a_j,mr,mt,[t_0,t]}=\{\alpha_k{:}\exists t'\in[t_0,t] \text{ s.t. } \texttt{done}(a_j,\alpha_i,t') \wedge \neg\texttt{moral\_conf}(\alpha_k,[CK_{a_i},GK_{a_i},RK_{a_i}],mr,mt,t')\}$$

FIGURE 2 – Ensembles des actions moralement ou éthiquement évaluées

4.1 Images éthique et morale des agents

Comme mentionné en Sec. 2.1, les images *éthiques* et *morales* d'un agent sont des croyances décrivant la conformité du comportement d'un agent au regard d'une connaissance du juste (RK) et du bien (GK).

Définition 4 (Image éthique (resp. morale)) *Une* image éthique *(resp.* image morale*) d'un agent a_j est le jugement du comportement $b_{a_j,[t_0,t]}$ de cet agent conformément à une éthique (resp. à un ensemble de règles morales R et d'un seuil moral mt) au regard des connaissances du contexte CK, de la moralité GK et de l'éthique RK d'un agent a_i. Cette image associe au comportement un élément $cv \in CV$ où CV est un ensemble ordonné de* niveaux de conformité [3]. *Ces images sont notées* $\texttt{ethical\_image}(a_j,a_i,cv,t_0,t)$ *et* $\texttt{morality\_image}(a_j,a_i,cv,R,mt,t_0,t)$

Remarquons tout d'abord que le premier paramètre fait référence à l'agent dont le comportement est évalué tandis que le second paramètre fait référence à l'agent dont la connaissance est employée pour construire l'image.

De plus, comme un jugement éthique indique une conformité ou non avec des principes éthiques, l'image éthique indique si un agent est conforme ou non dans son comportement. Ainsi, un agent ne peut avoir qu'une seule image éthique du comportement d'un autre. Dans le cas de l'image morale, comme les règles morales sont associées à divers niveaux de moralité, une image morale est associée à un seuil d'exigence pour indiquer une conformité morale ou non. Ainsi, un agent peut avoir plusieurs images morales en s'appuyant sur divers ensembles de règles R et leur seuil mt.

Afin de construire ces images, un agent a_i utilise deux fonctions d'agrégation $ethicAggregation$ et $moralAggregation$ appliquées respectivement aux actions éthiquement évaluées $EC_{a_j,[t_0,t]}$ et moralement évaluées $MC_{a_j,[t_0,t]}$. Ces deux fonctions d'agrégation calculent respectivement la proportion pondérée d'actions positivement évaluées au regard de l'éthique et de la morale. Le poids de chaque action dépend d'un critère (tel que le temps passé depuis son évaluation, les conséquences de l'action, etc.) que nous laissons volontairement abstrait dans cet article.

Définition 5 (Fonction d'agrégation éthique) $ethicAggregation : 2^{\mathcal{A}} \rightarrow \mathbb{R}$ *est une fonction d'agrégation éthique où :*

$$ethicAggregation(EC_{a_j,[t_0,t]}) = \sum_{\alpha_i \in EC^{+}_{a_j,[t_0,t]}} weight(\alpha_i) / \sum_{\alpha_i \in EC_{a_j,[t_0,t]}} weight(\alpha_i)$$

Définition 6 (Fonction d'agrégation morale) $moralAggregation : 2^{\mathcal{A}} \rightarrow \mathbb{R}$ *est une fonction d'agrégation morale où :*

$$moralAggregation(MC_{a_j,[t_0,t]}) = \sum_{\alpha_i \in MC^{+}_{a_j,[t_0,t]}} weight(\alpha_i) / \sum_{\alpha_i \in MC_{a_j,[t_0,t]}} weight(\alpha_i)$$

Afin de transformer l'évaluation quantitative en une évaluation qualitative, chaque niveau de conformité est associé à un intervalle dans l'ensemble des valeurs que peuvent prendre les fonctions d'agrégation éthiques et morales. Une fois le niveau de conformité obtenu, les états mentaux associés $\texttt{moral\_image}(a_j,a_i,cv,R,mt,t_0,t)$ ou $\texttt{ethical\_image}(a_j,a_i,cv,t_0,t)$ sont produits. Par exemple, si le niveau de conformité `congruent` correspond à l'intervalle $[0.75;1]$, le comportement de l'agent est considéré comme éthique si $ethicAggregation \geq 0.75$.

Une fois construites, ces images peuvent être employées pour influencer les interactions en construisant des relations de confiance, ou pour décrire la moralité d'interaction dépendantes du comportement des autres.

3. De manière analogue aux valuations morales, les niveaux de conformité peuvent être { `improper`, `neutral`, `congruent` }.

4.2 Construction de la confiance

Grâce aux images morales et éthiques, un agent peut décider d'accorder sa confiance à un autre ou non. La confiance peut être absolue (une confiance dans la conformité à une éthique du comportement de l'autre) ou relative à un ensemble de règles morales (confiance dans la prudence de l'autre, sa responsabilité, son obéissance à un ensemble de règles de conduite, etc.). Nous définissons deux actions épistémiques internes permettant d'évaluer la possibilité d'établir ces deux types de confiance.

Définition 7 (Fonction de confiance) *La* fonction de confiance éthique $TB^e_{a_i}$ *(resp.* fonction de confiance morale $TB^m_{a_i}$*) est définie comme :* $TB^e_{a_i} : \mathbb{A} \rightarrow \{\top, \bot\}$ *(resp.* $TB^m_{a_i} : \mathbb{A} \times 2^{\mathcal{MR}_{a_i}} \times MV_{a_i} \rightarrow \{\top, \bot\}$*)*

Ici, ces fonctions de confiance sont abstraites et doivent être instanciées. Par exemple, lorsqu'un agent a_i évalue la conformité du comportement d'un autre agent a_j au regard de CK_{a_i}, GK_{a_i} et RK_{a_i} (i.e. l'image éthique), la fonction de confiance éthique produit une croyance $\texttt{ethical\_trust}(a_j, a_i)$. De même, lorsque l'agent a_i évalue la conformité du comportement de a_j au regard de R (i.e. vérifie que la conformité morale de l'image de son comportement par rapport à R est au moins égale à mt), la fonction de confiance morale produit une croyance $\texttt{moral\_trust}(a_j, a_i, R, mt)$.

4.3 Éthique de la confiance

Les croyances sur l'image et la confiance peuvent devenir des éléments de contexte permettant d'exprimer la moralité ou l'éthique d'une action. Autrement dit, la moralité d'une action à l'égard d'un agent peut être conditionnée à la confiance ou l'image que l'agent juge a de l'autre.

Premièrement, la confiance éthique et morale peuvent enrichir la description des règles et valeurs morales. Par exemple, la valeur de *responsabilité* pourrait être supportée lorsque les actions de délégation ne sont confiées qu'à des agents de confiance. Ici, la responsabilité est définie comme la capacité à déléguer des actions sensibles uniquement à des agents appropriés.

Deuxièmement, des croyances spécifiques de confiance morale peuvent être employées comme des éléments de règle morale. Par exemple, étant donnée une valeur d'honnêteté et ses supports de valeur, un agent peut être doté d'une règle exprimant "Il est immoral de ne pas agir honnêtement à l'encontre de tout agent honnête.". Ici, "tout agent honnête" peut être modélisé par l'existence d'une croyance `moral_trust` associant à un agent une confiance morale dans la conformité de son comportement à l'ensemble R des règles définissant la moralité d'un comportement honnête.

Enfin, puisque évaluer et juger les autre constituent des actions, il est également possible d'exprimer et évaluer leur caractère moral ou éthique. Ainsi, la valeur morale de *tolérance* peut être supportée par la construction d'une image des autres avec un seuil peu élevé tant que les ensembles $EC_{a_j,[t_0,t]}$ ou $MC_{a_j,[t_0,t]}$ ne sont pas assez significatifs. Le choix du seuil, des pondérations et la conversion de l'agrégation en niveau de conformité peuvent également permettre de représenter diverses formes de confiance. Une valeur telle que l'*indulgence* peut être supportée par le fait d'accorder toujours une pondération plus faibles aux actions les moins récentes. Il est ainsi possible de décrire une morale de la confiance par l'emploi de règles comme "Il est immoral de construire la confiance sans tolérance ni indulgence" [21].

5 Preuve de concept

Cette section illustre la manière dont les éléments présentés dans la section précédente ont été implémentés dans un système multi-agent. Cette preuve de concept est implémentée à l'aide de la plate-forme JaCaMo [7] avec des agents BDI décrits en langage Jason partageant un environnement comportant des artefacts conformes aux standards de Cartago. Le code source complet est téléchargeable sur notre site [4]. L'environnement simule un marché financier sur lequel des actifs sont cotés et échangés par des agents. La Sec. 5.1 introduit le domaine de la gestion éthique d'actifs et les caractéristiques de notre application. Les morales et éthiques employés sont définies en Sec. 5.2. La construction des images et de la confiance est présentée en Sec. 5.3.

5.1 Modèle de marché financier

La gestion d'actifs financiers soulève bon nombre de problématiques éthiques et pratiques [5]. Ces décisions sont déléguées à des

4. https://cointe.users.greyc.fr/projects/ethical_market_simulator
5. http://sevenpillarsinstitute.org/

agents autonomes auxquels des utilisateurs humains délèguent les décisions d'achat et de vente, pouvant avoir des conséquences sur l'économie réelle [18]. Comme montré par [8], certains fonds d'investissement proposent une offre de gestion éthique et responsable de placements, et leur nombre ainsi que leur proportion sur les marchés tend à croître de manière significative ces dernières années. Toutefois, si les performances de ces fonds en matière de gain est objectivement mesurable, l'appréciation de la dimension éthique de leur comportement semble plus difficile et subjective car dépendante des convictions de l'observateur.

Dans cette preuve de concept, nous considérons un marché sur lequel les agents autonomes gestionnaires d'actifs peuvent échanger de la monnaie et des parts de capitaux. Chaque agent peut déposer des ordres d'achat et de vente qui seront exécutés lorsque le marché aura trouvé un autre ordre correspondant en terme de prix et de quantité. L'agent peut également annuler un ordre qui n'aurait pas encore été exécuté. Dans notre implémentation, le marché emploie une structure de *Central Limit Order Book* (CLOB) [2] pour conserver les ordres et faire correspondre l'offre et la demande.

En observant le marché, les agents peuvent percevoir les ordres en attente d'exécution et un ensemble d'indicateurs (prix et volume des derniers échanges effectués, moyenne et écart-type des prix au cours du temps, etc.).

Les agents disposent également d'informations sur le contenu de leur propre portefeuille d'actions. En raisonnant sur ces croyances comme une connaissance contextuelle CK, l'agent peut déduire ce qu'il peut vendre ou acheter à l'instant présent pour produire $\mathcal{A}_p$. En y ajoutant les informations dont il dispose sur l'évolution du marché, il est également capable de produire $\mathcal{A}_d$. Pour cela nous avons implémenté une stratégie simple basée sur des comparaisons de moyennes mobiles.

Trois types d'agents sont présents dans l'expérience : (1) des *agents aléatoires* assignés à des actifs cotés, passant des ordres au hasard en termes de prix et de volume afin de générer de l'activité et simuler le "bruit" d'un marché réel ; (2) des *agents sans éthique*, uniquement dotés d'une fonction d'évaluation de la désirabilité en guise de stratégie leur permettant de spéculer. La même fonction sera employée chez les agents éthiques pour générer $\mathcal{A}_d$; (3) des *agents éthiques* implémentant le jugement éthique comme processus décisionnel afin de se comporter conformément à leur éthique.

5.2 Paramétrage éthique

Afin d'informer leurs jugements et permettre de définir des contextes de règles morales, les agents éthiques disposent de connaissances sur les actifs. Par exemple, les certifications d'entreprise attestant des engagements pris envers l'environnement, ou encore leur secteur d'activité telle que la production d'énergie d'origine nucléaire.

De plus, nous avons ajouté à ces informations sur la situation une description de convictions morales directement inspirées de la littérature [6]. Les agents éthiques sont ainsi dotés d'un ensemble de valeurs et sous-valeurs hiérarchisées : par exemple *environmental reporting* est considéré comme une sous-valeur (au sens d'une spécificité plus forte) de la valeur *environment*.

Ces valeurs sont concrètement décrites par un ensemble de supports tes que "échanger des actifs de producteur d'énergie nucléaire n'est pas conforme à la sous-valeur *promotion of renewable energy*", "échanger des actifs d'une société labelisée FSC est conforme à la sous-valeur *environmental reporting*" ou "échanger des actifs de producteur d'énergie nucléaire est conforme à la sous-valeur *fight climate change*".

Les agents sont également dotés de règles morales pouvant employer des valeurs morales pour définir la moralité d'un comportement. Par exemple "Il est moral d'agir conformément à la valeur *environment*". Ces règles morales sont regroupées au sein d'ensembles sur lesquels pourront être agrégées les images.

À ce stade, un agent éthique est capable d'inférer par exemple que, au regard de ses connaissances sur le contexte CK et sur la morale GK, échanger des actifs d'une entreprise labellisée FSC est moral tandis qu'échanger des actifs d'une compagnie produisant de l'énergie nucléaire est à la fois moral et immoral. Pour déterminer s'il est juste d'échanger le second actif, l'agent aura besoin de sa théorie du juste. Nous avons ainsi doté ces agents de principes éthiques élaborés tels que l'éthique d'Aristote (inspirée de [20]) et de plus simples tels que "Un acte est juste s'il est possible, moral et désirable". Chaque agent peut avoir ainsi de nombreux principes éthiques et l'action juste sera celle qui satisfait le mieux les principes dans un ordre lexicographique.

6. http://www.ethicalconsumer.org/

5.3 Construction d'images et de confiance

À chaque action exécutée sur le marché, les agents reçoivent un message et réévaluent leurs images des agents impliqués dans la transaction. Comme mentionné dans la précédente section, évaluer la conformité de comportements, construire l'image et la confiance sont des actions. Elles sont donc décrites comme des plans Jason [7]. Dans la suite de cette section nous détaillons la construction de la confiance morale. La confiance éthique se construit de manière analogue.

Premièrement, un plan évalue la conformité d'une action avec chaque règle morale de l'ensemble R sur lequel porte l'image. Le nombre d'action morales ou immorales se trouve incrémenté à l'issue de chaque évaluation. Dans l'implémentation actuelle, nous utilisons une agrégation linéaire (c'est-à-dire associant la même pondération à chaque action). Ensuite le niveau de conformité est attribué en fonction de la proportion d'actions conformes afin de construire l'image. Dans cette expérimentation nous n'utilisons que trois niveaux de conformité (arbitrairement `neutral` pour un résultat compris dans $[0.4, 0.6[$, `improper` pour les résultats inférieurs et `congruent` pour les résultats supérieurs). Enfin, lorsque le niveau de conformité franchi le seuil de confiance, un plan met à jour la confiance dans l'agent jugé au regard des règles concernées.

5.4 Resultats

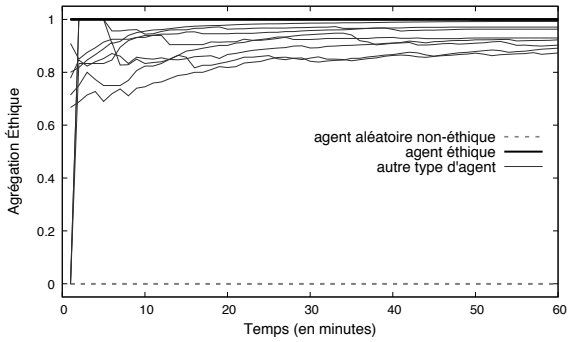

FIGURE 3 – Évolution des images des agents en sortie de la fonction d'agrégation éthique

La figure 3 montre l'évolution de l'agrégation des images éthiques en sortie de la fonction d'agrégation d'un agent éthique observant les autres agents du système (ici 20 agents aléatoires, 10 agents sans éthique et 3 agents éthiques). Remarquons premièrement qu'un agent peut construire une image d'un agent aléatoire ou d'un agent non-éthique. C'est l'une des propriétés de ce modèle : nous n'évaluons que la conformité d'un comportement observé au regard d'une éthique, sans nécessairement chercher à connaître l'intention des autres agents. Comme attendu, les agents ayant la même éthique restent durant toute l'observation à la valeur maximale de 1.0 (traits épais). À l'inverse, les agents aléatoires affectés à l'animation du cours d'un actif immoral aux yeux de l'agent juge restent à 0.0 (traits pointillés). Tous les autres agents convergent lentement vers une valeur intermédiaire dépendante de leurs investissements. En observant les croyances des agents au cours de l'expérience, il est possible de voir les agents éthiques construire les images des autres et établir une confiance lorsque cette image franchit les seuils de conformité.

6 Conclusion

Dans cet article nous avons montré comment un processus de jugement éthique peut être employé dans une architecture d'agent BDI et nous avons défini des mécanismes permettant de construire des images caractérisant la conformité d'un comportement du point de vue d'une éthique ou une morale. Nous avons ensuite montré la manière dont ces images peuvent être employées pour décider d'établir ou non une relation de confiance afin de coopérer. Une preuve de concept montre comment un tel modèle peut être implémenté dans une plate-forme BDI et utilisé dans le cadre de la gestion d'actifs financiers. D'un point de vue de la modélisation, ce travail répond au problème de l'évaluation de la proximité entre une éthique et un comportement, problème d'autant plus difficile dans les cas où l'éthique repose sur des convictions personnelles d'agents hétérogènes. Avec ce modèle, les agents peuvent savoir quel ensemble de règles en particulier ou quelle éthique en général sont concernées par cette proximité. Grâce à l'expressivité de ce modèle en matière d'éthique et de morale, nous envisageons de représenter des agents dotés d'un plus vaste ensemble de principes éthiques et de valeurs et règles morales prenant en compte dans leur description de l'image des autres, et décrire des notions telles que l'indulgence ou l'intransigeance.

Remerciements

Ce travail a été réalisé dans le cadre du projet EthicAa [7] (référence ANR-13-CORD-0006).

7. `http://ethicaa.org/`

Références

[1] A. Abdul-Rahman and S. Hailes. Supporting trust in virtual communities. In *33th IEEE International Conference on Systems Sciences*, pages 1–9, 2000.

[2] I. Aldridge. *High-frequency trading : a practical guide to algorithmic strategies and trading systems*, volume 459. John Wiley and Sons, 2009.

[3] M. Anderson and S.L. Anderson. Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm. *Industrial Robot*, 42(4) :324–331, 2014.

[4] R. Arkin. *Governing lethal behavior in autonomous robots*. CRC Press, 2009.

[5] K. Arkoudas, S. Bringsjord, and P. Bello. Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pages 17–23, 2005.

[6] F. Berreby, G. Bourgne, and J.-G. Ganascia. Modelling moral reasoning and ethical responsibility with logic programming. In *20th International Conference on Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548, 2015.

[7] Olivier Boissier, Rafael H Bordini, Jomi F Hübner, Alessandro Ricci, and Andrea Santi. Multi-agent oriented programming with jacamo. *Science of Computer Programming*, 78(6) :747–761, 2013.

[8] S. Bono, G. Bresin, F. Pezzolato, S. Ramelli, and F. Benseddik. Green, social and ethical funds in europe. Technical report, Vigeo, 2013.

[9] J. Carbo, J. Molina, and J. Davila. Comparing predictions of SPORAS vs. a fuzzy reputation agent system. In *3rd International Joint Conference on Fuzzy Sets and Fuzzy Systems*, pages 147–153, 2002.

[10] J. Carter, E. Bitting, and A. Ghorbani. Reputation formalization for an information-sharing multi-agent system. *Computational Intelligence*, 18(2) :515–534, 2002.

[11] C. Castelfranchi and R. Falcone. *Trust theory : A socio-cognitive and computational model*, volume 18. John Wiley & Sons, 2010.

[12] H. Coelho and A.C. da Rocha Costa. On the intelligence of moral agency. *Encontro Português de Inteligência Artificial*, pages 12–15, October 2009.

[13] H. Coelho, P. Trigo, and A.C. da Rocha Costa. On the operationality of moral-sense decision making. In *2nd Brazilian Workshop on Social Simulation*, pages 15–20, 2010.

[14] N. Cointe, G. Bonnet, and O. Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *15th International Conference on Autonomous Agents & Multiagent Systems*, pages 1106–1114, 2016.

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

[16] R. Conte and M. Paolucci. *Reputation in artificial societies : Social beliefs for social order*, volume 6. Springer Science & Business Media, 2002.

[17] B. Esfandiari and S. Chandrasekharan. On how agents make friends : Mechanisms for trust acquisition. In *4th Workshop on Deception, Fraud, and Trust in Agent Societies*, pages 27–34, 2001.

[18] Directorate-General for Economic and Financial Affairs. Impact of the current economic and financial crisis on potential output. Occasional Papers 49, European Commission, June 2009.

[19] J.-G. Ganascia. Ethical system formalization using non-monotonic logics. In *29th Annual Conference of the Cognitive Science Society*, pages 1013–1018, 2007.

[20] J.-G. Ganascia. Modelling ethical rules of lying with Answer Set Programming. *Ethics and Information Technology*, 9(1) :39–47, 2007.

[21] H.J.N Horsburgh. The ethics of trust. *The Philosophical Quartely*, 10(41) :343–354, 1960.

[22] R. Johnson. Kant's moral philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer edition, 2014.

[23] A. Josang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, 43(2) :618–644, 2007.

[24] E. Lorini. On the logical foundations of moral agency. In *11th International Conference on Deontic Logic in Computer Science*, pages 108–122, 2012.

[25] S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Stirling, 1994.

[26] A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter edition, 2014.

[27] G. Muller, L. Vercouter, and O. Boissier. Towards a general definition of trust and its application to openness in MAS. In *6th Workshop on Deception, Fraud and Trust in Agent Societies*, pages 49–56, 2003.

[28] P. Ricoeur. *Oneself as another*. University of Chicago Press, 1995.

[29] A. Rocha-Costa. Moral systems of agent societies : Some elements for their analysis and design. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 32–37, 2016.

[30] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence*, 24(1) :33–60, 2005.

[31] Jordi Sabater-Mir and Laurent Vercouter. Trust and reputation in multiagent systems. *Multiagent Systems*, page 381, 2013.

[32] A. Saptawijaya and L. Moniz Pereira. Towards modeling morality computationally with logic programming. In *Practical Aspects of Declarative Languages*, pages 104–119. 2014.

[33] M. Timmons. *Moral theory : an introduction*. Rowman & Littlefield Publishers, 2012.

[34] L. Vercouter and G. Muller. L.I.A.R. : Achieving social control in open and decentralized multiagent systems. *Applied Artificial Intelligence*, 24(8) :723–768, 2010.

[35] B. Yu and M.P. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4) :535–549, 2002.