

Jugement éthique dans les systèmes multi-agents

N. Cointe^{a,b}
nicolas.cointe@emse.fr

G. Bonnet^b
gregory.bonnet@unicaen.fr

O. Boissier^a
olivier.boissier@emse.fr

^aInstitut Henri Fayol, EMSE, LabHC,
CNRS UMR 5516, F-42000 Saint-Etienne, France

^bNormandie University, GREYC
CNRS UMR 6072, F-14032 Caen, France

Résumé

L'usage croissant des systèmes multi-agents dans divers domaines soulève la nécessité de concevoir des agents capables de prendre des décisions s'appuyant sur des principes éthiques. De plus en plus de travaux proposent de telles approches. Toutefois, ces systèmes considèrent principalement une perspective centrée sur l'agent et mettent de côté le fait que ces agents sont en interaction avec d'autres agents, artificiels ou humains qui utilisent d'autres concepts éthiques. Dans cet article, nous nous intéressons à ce problème en proposant un modèle de jugement éthique qu'un agent peut utiliser pour juger à la fois de son propre comportement et de celui des autres agents au sein de systèmes multi-agents. Ce modèle est basé sur une approche rationaliste et explicite qui distingue théorie du bien et théorie du juste. Une preuve de concept implémentée en Answer Set Programming est fournie pour illustrer ces fonctionnalités.

Mots-clés : Agent (architecture), Modèles de comportement agent, Éthique

Abstract

The increasing use of multi-agent technologies in various areas raises the necessity of designing agents that are able to take decisions based on ethical principles. More and more works propose such approaches. However, those systems consider mainly an agent-centered perspective, letting aside the fact that agents are in interaction with other artificial agents or human beings that can use other ethical concepts. In this article, we address this problem and propose a model of ethical judgment an agent can use in order to judge the ethical dimension of both its own behavior and the other agents' behaviors. This model is based on a rationalist and explicit approach that distinguishes theory of good and theory of right. A proof-of-concept implemented in Answer Set Programming and based on a simple scenario is given to illustrate those functionalities.

Keywords: Agent (architecture), Agent's model of behavior, Ethics

1 Introduction

La présence croissante des agents autonomes dans une grande variété de domaines tels que la santé, la finance ou les transports peut soulever des problèmes si ces agents ne sont pas capables de considérer et suivre certaines règles ainsi qu'adapter leur comportement. Par exemple, la compréhension d'un code de déontologie peut faciliter la coopération entre un agent et un médecin ou un patient, en tenant compte de concepts tels que le secret médical ou le respect de la dignité. Même si plusieurs travaux proposent des implémentations de restriction d'action [27], de simples interdits ou obligations [7], les codes de conduite font appel à des notions plus complexes telles que des valeurs morales ou des principes éthiques. Une implémentation explicite de concepts tels que l'altruisme ou la générosité nécessite des structures et des processus spécifiques dans l'architecture de l'agent. Il est à noter que tous ces travaux considèrent l'éthique d'un point de vue individuel et mono-agent alors que de nombreuses applications réelles mettent en œuvre des agents en interaction. Il est donc nécessaire de prendre en compte un point de vue collectif et multi-agent dans ces travaux.

Un point de vue individuel pourrait être suffisant pour permettre à un agent d'agir de manière éthique dans une organisation. Toutefois, pour évaluer l'éthique des autres (dans l'optique d'une collaboration ou d'une réaction à un comportement inapproprié), les agents doivent pouvoir juger le caractère éthique ou non du comportement des autres. Dans cet article, nous nous intéressons à la question du jugement éthique, c'est à dire la vérification de la conformité ou non du comportement d'un agent vis-à-vis de convictions morales et de principes éthiques. Nous proposons un modèle générique de jugement du comportement pouvant être utilisé par un agent tant pour guider son propre comportement que pour évaluer celui des autres.

Cet article¹ est organisé de la manière suivante : la section 2 introduit quelques concepts de phi-

¹ Cet article est une version retravaillée d'un article plus long publié à AAMAS 2016.

philosophie morale et un court état de l'art sur des approches d'agents autonomes éthiques. Nous détaillons dans la section 3 notre modèle de jugement éthique. La section 4 illustre ensuite son utilisation par un agent dans son interaction avec les autres en fonction des informations disponibles. La section 5 montre une preuve de concept en ASP (Answer Set Programming). Nous comparons notre démarche avec des approches existantes alternatives en section 6 et concluons en section 7 en précisant l'intérêt de ces travaux pour des systèmes multi-agents et en donnant des perspectives pour les prochaines étapes de notre travail.

2 Éthique et agents autonomes

La section 2.1 introduit des concepts de philosophie morale sur laquelle se fonde notre approche, puis nous passons en revue en section 2.2 les architectures d'agent existantes proposant l'implémentation de comportements éthiques. Enfin, la section 2.3 résume les caractéristiques recherchées dans notre approche.

2.1 Concepts de philosophie morale

Des philosophes antiques aux travaux récents de neurologie [9] et sciences cognitives [14], de nombreuses études se sont intéressées à la capacité humaine à définir et distinguer le bien et le juste du mal et de l'injuste. De ces nombreux travaux de philosophie morale sur les concepts tels que la *morale*, l'*éthique*, le *jugement* et les *valeurs*, nous tirons les définitions suivantes :

Définition 1 (Morale) *La morale désigne l'ensemble de règles déterminant la conformité des pensées ou actions d'un individu avec les mœurs, us et coutumes d'une société, d'un groupe (communauté religieuse, etc.) ou d'un individu pour évaluer son propre comportement. Ces règles reposent sur les valeurs normatives de bien et de mal. Elles peuvent être universelles ou relatives, c'est-à-dire liées ou non à une époque, un peuple, un lieu, etc.*

Chacun connaît des règles telles que "il est mal de mentir", "se montrer loyal est une bonne chose" ou "il est mal de tricher". C'est sur ce type de règles que peut se fonder notre raisonnement permettant de distinguer les bonnes et mauvaises actions. La morale se distingue de la loi et du système légal dans le sens où elle ne comporte pas de pénalités explicites ou de règles officiellement établies [13]. Les règles morales sont couramment soutenues et justifiées par des valeurs morales (liberté, bienveillance, sagesse, conformisme, etc.). Psychologues, sociologues et anthropologues admettent pour la plupart que les valeurs morales sont l'élément central dans

l'évaluation de la justesse d'une action, d'une personne ou d'un événement [24].

Un ensemble de règles morales ou de valeurs morales établit une *théorie du bien* qui permet à chacun d'évaluer la moralité d'un comportement. Un ensemble de principes forme la *théorie du juste* qui définit des critères pour reconnaître le choix le plus juste ou le plus acceptable (elles sont également appelées *théorie des valeurs* et *théorie de la juste conduite* [25]). Par exemple, bien que le vol soit souvent reconnu comme immoral (au regard d'une théorie du bien), de nombreuses personnes s'accorderont à reconnaître qu'il est acceptable qu'un orphelin affamé dérobe une pomme dans un supermarché (au regard d'une certaine théorie du juste). Les humains acceptent souvent dans certaines situations qu'il soit juste de satisfaire des besoins ou désirs en violation avec certaines règles et valeurs morales. La description de cette conciliation est appelée *éthique* et, en accord avec certains philosophes tels que Paul Ricoeur [22], nous admettons la définition suivante :

Définition 2 (Éthique) *L'éthique est la combinaison de principes éthiques et de règles morales permettant à un processus de décider d'une action satisfaisant au mieux les règles morales, les désirs et les croyances de l'agent, étant donné les capacités de ce dernier.*

Les philosophes ont proposé une grande variété de principes éthiques tels que l'Impératif Catégorique de Kant [16] ou la Doctrine du Double Effet de St Thomas d'Aquin [20], qui sont des ensembles de règles permettant de distinguer une décision éthique parmi un ensemble de choix possibles. Traditionnellement, trois approches majeures se distinguent dans la littérature : l'*éthique des vertus* juge la conformité d'une action à des valeurs telles que la sagesse, le courage ou la justice [15] ; l'*éthique déontologique* juge un comportement par sa conformité avec des obligations et permissions associées à des situations [2] ; l'*éthique conséquentialiste* juge un comportement à la moralité de ses conséquences [26].

Mais parfois un principe éthique est incapable de distinguer la meilleure décision. Ces situations appelées *dilemmes* sont des choix entre deux options, chacune étant supportée par des motivations éthiques, sans qu'il soit possible de réaliser les deux [19]. Chaque option apportera un regret. De nombreux dilemmes tels que le dilemme du trolley [11], sont considérés comme des failles dans la morale ou l'éthique, ou *a minima* comme d'intéressants cas d'études sur la faculté de formuler et expliquer rationnellement un jugement éthique. Dans cet article, nous considérons un dilemme comme un choix pour

lequel un principe éthique donné ne peut distinguer la meilleure option au regard d'une théorie du bien. Face à un dilemme, un agent peut considérer plusieurs principes afin de distinguer la plus juste décision envisageable. C'est pourquoi un agent autonome éthique doit être capable de comprendre un large éventail de principes éthiques et de distinguer ceux qui lui permettent de prendre les décisions les plus justes.

De fait, la faculté de jugement est au cœur de l'éthique et constitue l'étape finale pour prendre une décision éthique en évaluant chaque choix au regard de ses désirs, sa morale, ses capacités et principes éthiques. En accord avec quelques définitions consensuelles [1] et les concepts précédemment évoqués, nous considérons la définition suivante de jugement :

Définition 3 (Jugement) *Le Jugement est la faculté de distinguer l'option la plus satisfaisante d'un choix dans une situation donnée, au regard d'un ensemble de principes, pour soi-même ou autrui.*

2.2 Éthique et agents autonomes

En prenant en considération toutes ces notions, de nombreux cadres ont été définis pour concevoir des agents autonomes comprenant une éthique individuelle. Nous les regroupons au sein de quatre approches que sont l'*éthique par conception*, l'*éthique par étude de cas*, l'*éthique par raisonnement logique* et l'*architecture cognitive éthique*.

L'*éthique par conception* consiste en la création d'un agent en prenant en compte une analyse de chaque situation pouvant être rencontrée lors de son fonctionnement et l'implémentation de la conduite éthique à suivre. Cette approche peut être une implémentation directe et rigide de règles (par exemple les règles militaires d'engagement pour un drone armé [4]). Son inconvénient principal est l'absence de représentation générique de concepts éthiques (théories du bien et du mal). De plus, il est impossible de comparer deux éthiques par conception en raison de l'absence de représentations explicites. Concevoir ainsi des agents hétérogènes coopératifs dotés de divers désirs, principes ou règles morales sans représentation explicite devient difficilement envisageable et se limite bien souvent à l'obéissance stricte à une déontologie dont les règles sont directement implémentées sous formes d'entraves et de contraintes dans le processus de décision de l'agent.

L'*éthique par étude de cas* cherche premièrement à inférer des règles éthiques à partir d'un vaste ensemble de jugements exprimés par des experts, puis à les appliquer pour produire un comportement éthique [3]. Même si cette approche a l'avantage de proposer une solution

générique à l'ensemble des champs applicatifs, l'expertise humaine dans chaque domaine est nécessaire pour envisager un grand ensemble de situations. De plus, le comportement éthique de l'agent n'est pas garanti (notamment dans les cas classiques de sous- ou sur-apprentissage). L'agent n'a pas de description explicite de son éthique et son raisonnement éthique est basé sur des reconnaissances de similarités, non sur de la déduction. Par conséquent, la coopération entre agents hétérogènes se heurte aux mêmes difficultés que l'éthique par conception.

L'*éthique par raisonnement logique* est une implémentation de principes éthiques formalisés (tels l'Impératif Catégorique de Kant ou la Doctrine du Double Effet de Saint Thomas d'Aquin) en programmation logique [12, 23]. Le principal avantage de cette méthode réside dans l'apport d'une représentation explicite de la théorie du juste, même si la théorie du bien n'est souvent qu'un ensemble de paramètres donné. Cette approche permet de juger une décision en prenant en compte un principe éthique.

Enfin, les *architectures cognitives éthiques* consistent en une représentation explicite de chaque élément permettant la prise de décision de l'agent, des croyances décrivant la perception de l'environnement et des autres agents, désirs (objectifs de l'agent) et intentions (décisions prises par l'agent) à des concepts tels que des heuristiques ou simulations d'émotions [5, 8]. Bien que ces approches permettent aux agents de manipuler des règles explicites et justifier leurs décisions, la faculté de raisonnement sur une représentation explicite de l'éthique des autres agents n'a pas encore été représentée.

2.3 Prérequis pour le jugement éthique dans les systèmes multi-agents

Les approches présentées dans la section précédente proposent des techniques et modèles intéressants pour représenter un agent autonome éthique. Toutefois dans un système multi-agent, les agents peuvent avoir besoin d'interagir et collaborer pour partager des ressources, échanger des données ou effectuer des actions collectivement. Les approches précédentes considèrent souvent les autres agents du système comme une partie de l'environnement alors qu'une perspective collective de l'éthique nécessiterait sa représentation et sa prise en compte dans le processus décisionnel de l'agent. Nous identifions deux besoins majeurs pour concevoir ce type d'agents éthiques.

Les agents ont besoin d'une *représentation explicite de l'éthique* comme suggéré par la théorie de l'esprit en psychologie : l'éthique des autres ne peut être comprise que par une représentation au sein de l'agent de l'éthique individuelle d'un autre [17]. Afin d'exprimer et

concilier un maximum de théories du bien et du juste, nous proposons de les représenter au sein de composants clairement définis et d'ordonner les principes selon un ordre de préférence. Ce type de représentation pourrait en outre faciliter la configuration des agents par des non-spécialistes de l'intelligence artificielle et simplifier les communications avec d'autres agents, y compris les humains.

Les agents ont besoin d'un *processus de jugement explicite* afin de permettre à la fois des raisonnements individuels et collectifs sur diverses théories du bien et du juste. En accord avec les précédentes définitions, nous considérons le jugement comme une évaluation de la conformité d'un ensemble d'actions au regard d'un ensemble de valeurs et règles morales ainsi que de principes et préférences éthiques. Nous proposons différents types de jugements basés sur la possibilité de substituer la morale d'un agent par celle d'un autre. Ainsi, nous proposons d'utiliser le jugement à la fois comme un processus de décision dans un problème de choix social [18], et comme une capacité à évaluer le caractère éthique du comportement des autres.

Par la suite, nous décrivons le modèle générique que nous proposons pour juger l'éthique du comportement propre à l'agent ou de celui des autres.

3 Processus de jugement éthique

Dans cette section nous introduisons notre modèle générique de jugement. Après une courte présentation globale nous détaillons chaque fonction et expliquons son fonctionnement.

3.1 Présentation globale

Comme expliqué dans la section 2.1, l'éthique consiste en la conciliation des désirs, de la morale et des capacités. Pour prendre ces dimensions en compte, notre processus de jugement éthique (*EJP*) utilise des connaissances sur l'évaluation de situation, la morale et l'éthique. Nous l'avons structuré en processus de *Reconnaissance de situation*, *Évaluation*, *Morale* et *Éthique* (voir Fig. 1). Dans cet article, nous considérons ce processus dans le contexte d'un modèle BDI, utilisant des états internes tels que des croyances et désirs. Pour des raisons de simplicité, nous considérons ici des raisonnements éthiques à court terme, ne portant que sur un comportement qui se résume à des actions immédiates. Ce modèle se fonde uniquement sur des états mentaux et fait abstraction des spécificités de l'architecture.

Définition 4 Un processus de jugement éthique *EJP* est défini comme une composition d'une

reconnaissance de situation (AP), un processus d'évaluation (EP), un processus moral (GP), un processus éthique (RP) et une ontologie O ($O = O_v \cup O_m$) de valeurs morales (O_v) et valuations morales (O_m). Ce processus de jugement éthique produit une évaluation des actions pour l'état courant du monde W en tenant compte de considérations morales et éthiques.

$$EJP = \langle AP, EP, GP, RP, O \rangle$$

Ce modèle doit être considéré comme un modèle générique composé de fonctions abstraites, états mentaux et bases de connaissances. Ces fonctions peuvent être implémentées de diverses manières. Par exemple les valuations morales de O peuvent prendre la forme d'un ensemble d'éléments discrets tel que { bien, mal } ou continu comme un degré de moralité.

3.2 Reconnaissance de situation et évaluation

Dans ces processus, l'agent commence par évaluer l'état du monde sous la forme de production de croyances et désirs au travers du processus de reconnaissance de situation.

Définition 5 Le processus de reconnaissance de situation *AP* génère l'ensemble des croyances qui décrivent l'état courant du monde W et l'ensemble des désirs qui décrivent les buts de l'agent. Il est défini comme :

$$AP = \langle \mathcal{B}, \mathcal{D}, SA \rangle$$

où \mathcal{B} est l'ensemble des croyances que l'agent a sur W , et \mathcal{D} à partir de W :

$$SA : W \rightarrow 2^{\mathcal{B} \cup \mathcal{D}}$$

À partir de ses croyances \mathcal{B} et désirs \mathcal{D} , un agent exécute le processus d'évaluation *EP* pour établir les actions désirables d'une part (i.e. les actions qui permettent de satisfaire un désir) et les actions exécutables (i.e. les actions pouvant être effectuées dans l'état courant du monde).

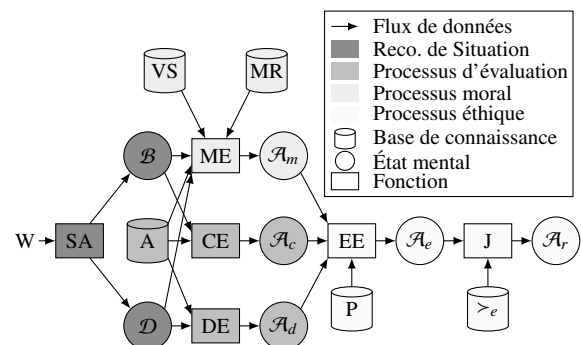


FIGURE 1 – Processus de jugement éthique

Définition 6 Le processus d'évaluation EP produit les ensembles d'actions désirables et d'actions exécutables à partir des ensembles de désirs et croyances. Nous le définissons comme :

$$EP = \langle A, \mathcal{A}_d, \mathcal{A}_c, DE, CE \rangle$$

où A est l'ensemble des actions (chacune étant décrite comme une paire de conditions et conséquences portant sur les croyances et les désirs), $\mathcal{A}_d \subseteq A$ et $\mathcal{A}_c \subseteq A$ sont respectivement l'ensemble des actions désirables et exécutables, DE et l'évaluation de capacités CE sont des fonctions telles que :

$$DE : 2^{\mathcal{D}} \times 2^A \rightarrow 2^{\mathcal{A}_d}$$

$$CE : 2^{\mathcal{B}} \times 2^A \rightarrow 2^{\mathcal{A}_c}$$

L'évaluation de désirabilité est la capacité à déduire les actions pertinentes à effectuer au regard des désirs et des connaissances sur les conditions et conséquences des actions.

Maintenant que nous avons défini les processus de reconnaissance de situation et d'évaluation, nous pouvons aborder les processus au cœur du modèle de jugement qui emploient les règles morales (respectivement les principes éthiques) pour définir le processus moral (respectivement le processus éthique).

3.3 Processus moral

Comme nous l'avons vu dans l'état de l'art, un agent éthique doit évaluer la moralité des actions dans la situation reconnue. À cette fin, nous définissons le processus moral ci-dessous.

Définition 7 Le processus moral GP identifie les actions morales à partir des croyances, désirs et connaissances sur les actions de l'agent ainsi que ses valeurs et règles morales. Nous le définissons formellement comme :

$$GP = \langle VS, MR, \mathcal{A}_m, ME \rangle$$

où VS est la base de connaissance du support de valeurs, MR est la base de connaissance de règles morales, $\mathcal{A}_m \subseteq A$ est l'ensemble des actions morales². La fonction d'évaluation morale ME est :

$$ME : 2^{\mathcal{D}} \times 2^{\mathcal{B}} \times 2^A \times 2^{VS} \times 2^{MR} \rightarrow 2^{\mathcal{A}_m}$$

Afin de réaliser ce processus moral, un agent doit d'abord pouvoir associer un ensemble fini de valeurs morales à des combinaisons d'actions et de situations. L'exécution d'une action dans la situation associée promet une valeur correspondante. On peut envisager diverses combinaisons pour une même valeur morale : par

2. $A_m \not\subseteq A_d \cup A_c$ car une action peut être morale en soi-même si elle n'est pas désirée ou réalisable (sauver le monde).

exemple l'honnêteté peut être définie comme "Ne pas dire quelque chose d'incompatible avec mes croyances" (car c'est mentir sciemment) ou comme "Dire ce que je crois lorsque je crois qu'un autre agent croit l'inverse" (pour éviter les mensonges par omission).

Définition 8 Un support de valeur est un couple $\langle s, v \rangle \in VS$ où $v \in \mathcal{O}_v$ est une valeur morale et $s = \langle a, w \rangle$ est le support de cette valeur morale où $a \subseteq A$, $w \subset \mathcal{B} \cup \mathcal{D}$.

La description précise d'une valeur morale repose sur la description des croyances, désirs et actions. Par exemple, les définitions de la générosité comme "donner à tout agent pauvre" et de l'honnêteté comme "Ne pas dire quelque chose d'incompatible avec mes croyances" peuvent être représentées :

$$\langle \langle give(\alpha), \{belief(poor(\alpha))\} \rangle, generosity \rangle$$

$$\langle \langle tell(\alpha, \phi), \{belief(\phi)\} \rangle, honesty \rangle$$

où α représente un agent et $poor(\alpha)$ (resp. ϕ) est une croyance représentant le contexte d'exécution de l'action $give(\alpha)$ (resp. $tell(\alpha, \phi)$) supportant la valeur *generosity* (resp. *honesty*).

En plus des valeurs morales, un agent peut manipuler la représentation de règles morales. Une règle morale décrit l'association d'une valuation morale (par exemple parmi un ensemble tel que {moral, amoral, immoral}) à des actions ou valeurs morales dans une situation.

Définition 9 Une règle morale est un n -uplet $\langle w, o, m \rangle \in MR$ où w est une situation de l'état courant du monde décrit par $w \subset \mathcal{B} \cup \mathcal{D}$ interprété comme une conjonction de croyances et désirs, $o = \langle a, v \rangle$ où $a \in A$ et $v \in V$, et $m \in \mathcal{O}_m$ est une valeur morale décrite dans \mathcal{O}_m qui qualifie o quand w est l'état courant.

Certaines règles morales classiques telles que "tuer un humain est immoral" ou "être honnête avec un menteur est plutôt moral" peuvent être représentées comme :

$$\langle \{human(\alpha)\}, \langle kill(\alpha), _ \rangle, immoral \rangle$$

$$\langle \{liar(\alpha)\}, \langle _, honesty \rangle, quite\ good \rangle$$

Une règle peut être plus ou moins spécifique à une situation w ou un objet o . Par exemple "La justice est morale" est plus générale (s'applique à un plus grand nombre de valeurs de w et o) que "Juger un meurtrier en prenant compte de sa religion, sa couleur de peau, son origine ethnique ou ses opinions politiques est immoral". De manière classique, les théories morales peuvent être représentées selon trois approches (voir section 2.1) : une approche *vertueuse* utilise des règles générales s'exprimant sur des

valeurs morales (e.g. “Il est moral d’être généreux”); une approche *déontologique* est généralement décrite par des règles les plus précises possible (e.g. “Les journalistes doivent refuser toute faveur aux publicitaires, donateurs ou groupes d’intérêt et résister aux pressions internes ou externes qui tenteraient de les influencer”³); une approche *consequentialiste* utilise à la fois des règles générales et spécifiques concernant les états et les conséquences (e.g. “Tout médecin doit s’abstenir, même en dehors de l’exercice de sa profession, de tout acte de nature à déconsidérer celle-ci.”⁴).

3.4 Processus éthique

À partir de l’ensemble des actions possibles, désirables et morales, nous pouvons introduire le processus éthique qui a pour but de déterminer l’action juste. Comme vu en section 2, un agent éthique peut utiliser plusieurs *principes éthiques* pour concilier ces ensembles d’actions.

Définition 10 *Un processus éthique RP produit les actions justes selon une représentation donnée de l’éthique. Nous définissons ce processus comme :*

$$RP = \langle P, >_e, \mathcal{A}_r, EE, J \rangle$$

où P est une base de connaissance sur les principes éthiques, $>_e \subseteq P \times P$ un ensemble de relations de préférences représentant un ordre partiel sur ces principes, $\mathcal{A}_r \subseteq A$ l’ensemble des actions justes et deux fonctions EE (évaluation éthique) et J (jugement) permettant d’en déduire l’action éthique :

$$EE : 2^{\mathcal{A}_d} \times 2^{\mathcal{A}_p} \times 2^{\mathcal{A}_m} \times 2^P \rightarrow 2^{\mathcal{E}}$$

où $\mathcal{E} = A \times P \times \{\perp, \top\}$

$$J : 2^{\mathcal{E}} \times 2^{>_e} \rightarrow 2^{\mathcal{A}_r}$$

Un principe éthique est une fonction qui présente une théorie philosophique permettant d’estimer s’il est juste ou non d’effectuer une action dans une situation donnée au regard de cette théorie.

Définition 11 *Un principe éthique $p \in P$ est une fonction décrivant la justesse d’une action évaluée en termes de capacités, désirs et moralité dans une situation donnée. Nous la définissons comme :*

$$p : 2^A \times 2^B \times 2^D \times 2^{MR} \times 2^V \rightarrow \{\top, \perp\}$$

La fonction d’évaluation éthique EE renvoie l’évaluation de toutes les actions désirables (\mathcal{A}_d), réalisables (\mathcal{A}_p) ou morales (\mathcal{A}_m) étant donné l’ensemble P des principes éthiques connus.

3. Extrait de [21], section “Act Independently”.

4. Code de déontologie médicale, article 31.

4 Jugement éthique des autres

Nous avons montré comment le processus de jugement éthique décrit dans la section précédente peut être utilisé pour qu’un agent juge de la meilleure action à effectuer. Toutefois ce modèle peut aussi être employé pour juger le comportement d’un autre agent de manière plus ou moins informé en se projetant à la place de l’agent jugé, de manière plus ou moins complète.

Dans un processus de jugement éthique EJP tel que défini dans la section précédente, les états mentaux \mathcal{B} , \mathcal{D} , \mathcal{A}_d , \mathcal{A}_p , \mathcal{E} , \mathcal{A}_m et les connaissances sur les actions (A), connaissances morales – *théorie du bien* – (MR , VS) et connaissances éthiques – *théorie du juste* – (P , $>_e$) peuvent être échangées entre agents. L’ontologie \mathcal{O} est considérée comme une connaissance commune, même si nous pourrions envisager dans des travaux futurs la coexistence de plusieurs ontologies. La manière dont ces connaissances sont échangées peut prendre de nombreuses formes comme du partage ou de l’inférence, et n’est pas abordée dans cet article. Dans tous les cas, nous distinguons trois catégories de jugement éthique : (1) le *jugement éthique aveugle* dans lequel l’agent juge n’a pas d’autre information sur l’agent jugé que son comportement observé ; (2) le *jugement éthique partiellement informé* lorsque le juge dispose d’informations partielles sur les connaissances de l’agent jugé ; (3) le *jugement éthique parfaitement informé* lorsque l’agent juge dispose de la totalité des informations existantes sur l’agent jugé.

Dans tous ces types de jugement, l’agent juge raisonne sur ses propres états mentaux à défaut de disposer de ceux de l’agent jugé. Ce type de jugement peut être comparé chez l’Homme à la théorie de l’esprit [17] (la faculté pour un humain à se représenter les états mentaux d’un autre). Ainsi, l’agent juge peut utiliser son propre processus de jugement éthique EJP en substituant autant que possible les états mentaux de l’autre agent aux siens afin de comparer \mathcal{A}_r et \mathcal{A}_m au comportement observé chez l’autre agent. Si l’action effectuée se trouve dans \mathcal{A}_r , l’agent juge peut supposer que l’agent jugé agit conformément à son éthique et, respectivement, si elle se trouve dans \mathcal{A}_m , elle est conforme à sa morale. Chaque évaluation est à considérer dans la situation de l’agent jugé, sa théorie du bien et sa théorie du juste utilisées pour le jugement. Nous considérons que tout jugement éthique est relatif aux connaissances, ontologies et états mentaux employés.

4.1 Jugement éthique aveugle

Ce premier type de jugement peut être effectué sans aucune information sur la morale ou l’éthique de l’agent jugé (par exemple dans

le cas d'une impossibilité de communiquer). L'agent juge a_j utilise alors sa propre évaluation de la situation (\mathcal{B}_{a_j} and \mathcal{D}_{a_j})⁵, sa propre théorie du bien $\langle MR_{a_j}, VS_{a_j} \rangle$ et théorie du juste $\langle P_{a_j}, >_{e,a_j} \rangle$ afin d'évaluer le comportement de l'agent jugé a_t . C'est un jugement *a priori* et a_t est jugé comme ayant effectué une action injuste ou immorale si $\alpha_{a_t} \notin \mathcal{A}_{r,a_j}$ ou $\alpha_{a_t} \notin \mathcal{A}_{m,a_j}$.

4.2 Jugement éthique partiellement informé

Le second type de jugement peut tenir compte d'une information partielle sur l'agent jugé s'il est capable de l'acquérir (par perception ou communication). Trois types de jugement éthique partiel sont considérés, en disposant respectivement (i) de la situation (i.e. $\mathcal{B}_{a_t}, \mathcal{D}_{a_t}, A_{a_t}$), (ii) de la théorie du bien (i.e. $\langle VS_{a_t}, MR_{a_t} \rangle$) et A_{a_t} ⁶ ou (iii) de la théorie du juste (i.e. $\langle P_{a_t}, >_{e,a_t} \rangle$) de l'agent jugé.

Jugement considérant la situation Premièrement, si l'agent juge a_j connaît les croyances \mathcal{B}_{a_t} et désirs \mathcal{D}_{a_t} de l'agent jugé a_t , a_j peut se placer dans la position d' a_t et juger de l'action α effectuée par a_t en vérifiant si elle fait partie de \mathcal{A}_{r,a_t} , en utilisant ses propres théories du bien et du juste. Premièrement, a_j est capable d'évaluer la moralité d' α en générant \mathcal{A}_{m,a_t} à partir d' A_{a_t} et qualifier la moralité du comportement d' a_t (i.e. si α est ou non dans \mathcal{A}_{m,a_t}). L'agent a_j peut aller plus loin en générant \mathcal{A}_{r,a_t} à partir de \mathcal{A}_{m,a_t} pour vérifier si α est conforme à la théorie du juste (i.e. fait partie de \mathcal{A}_{r,a_t}).

Jugement considérant la théorie du bien Deuxièmement, si l'agent juge est capable d'obtenir les règles morales et valeurs de l'agent jugé, il est possible d'évaluer l'action dans une situation (partagée ou non), au regard de ces règles. Dans la simple perspective d'une évaluation de la morale de l'agent jugé, l'agent juge peut comparer leurs théories du bien en vérifiant si les valeurs morales et règles morales de l'agent jugé sont constantes avec sa propre théorie du bien (i.e. s'il a les mêmes définitions que a_j ou au moins qu'il n'y a pas de contradictions). Dans la perspective d'un jugement moral, l'agent juge peut évaluer la moralité d'une action donnée du point de vue de l'agent jugé. Cette forme de jugement prend tout son intérêt par exemple lorsque les agents sont tenus à des devoirs moraux différents (en raison d'un rôle ou d'une responsa-

bilité particulière par exemple) comme un humain peut juger un médecin sur la conformité de son comportement vis-à-vis du code de déontologie médicale sans être lui-même un membre du corps médical.

Jugement considérant la théorie du juste Troisièmement, nous pouvons également considérer le jugement d'un agent juge capable de raisonner sur les principes et préférences éthiques d'un agent jugé en considérant une situation (partagée ou non) et une théorie du bien (partagée ou non)⁷. Cela permet d'évaluer comment l'agent a_t concilie ses désirs et sa morale dans une situation en comparant l'ensemble des actions justes \mathcal{A}_r, a_j et \mathcal{A}_r, a_t respectivement générées en utilisant $P_{a_j}, >_{e,a_j}$ et $P_{a_t}, >_{e,a_t}$. Par exemple, si $\mathcal{A}_r, a_j = \mathcal{A}_r, a_t$ avec une théorie du bien qui n'est pas partagée, cela montre que les deux théories du juste produisent un même jugement dans ce contexte. Ce jugement peut être utile pour un agent afin d'estimer comment un autre agent peut juger de l'éthique d'une action dans une situation avec une morale donnée.

4.3 Jugement pleinement informé

Enfin, l'agent juge peut prendre en considération à la fois la morale et l'éthique de l'agent jugé. Ce type de jugement nécessite la totalité des états mentaux internes et connaissances de l'agent jugé. Ce type de jugement est utile pour vérifier la conformité d'un comportement à une éthique publiquement déclarée.

5 Preuve de concept

Dans cette section nous illustrons le fonctionnement de chaque partie du modèle présenté dans les sections précédentes à travers l'implémentation d'un agent en Answer Set Programming (ASP). Le code source est disponible en ligne⁸. Cet agent est un exemple d'agent éthique dans un système où les agents ont des croyances (sur la richesse, le genre, le statut marital et la noblesse), des désirs et leur propre processus de jugement éthique. Les agents peuvent donner, courtiser, taxer et voler les autres ou attendre. Nous nous concentrerons principalement sur l'éthique de l'agent nommé `robin_hood`.

5.1 Reconnaissance de situation

Dans cet exemple, la fonction de reconnaissance de situation `SA` n'est pas implémentée et les croyances de l'agent sont fournies directement. Le code suivant représente un sous ensemble des croyances de `robin_hood` :

5. Nous utilisons une notation indicée pour désigner l'agent concerné par l'information.

6. Dans ce cas, A_{a_t} est nécessaire car, à l'inverse des principes éthiques, les règles morales peuvent porter directement sur des actions spécifiques.

7. Si la situation et la théorie du bien sont toutes deux partagées, il s'agit d'un jugement pleinement informé (voir 4.3).

8. <https://cointe.users.greyc.fr/download/>

```

agent(paul).
agent(prince_john).
agent(marian).
-poor(robin_hood).
-married(robin_hood).

-man(marian).
rich(prince_john).
man(prince_john).
noble(prince_john).
poor(paul).

```

L'ensemble des désirs \mathcal{D} est celui de robin_hood. Dans cette implémentation, nous considérons deux types de désirs, les désirs d'accomplir des actions (`desirableAction`) et ceux d'atteindre des états (`desirableState`).

```

desirableAction(robin_hood,robin_hood,court,marian).
desirableAction(robin_hood,robin_hood,steal,A):-
  agent(A), rich(A).
desireState(prince_john,rich,prince_john).
-desireState(friar_tuck,rich,friar_tuck).

```

Les deux premiers désirs concernent des actions : robin_hood désire courtiser marian et voler n'importe quel agent riche. Les deux désirs suivants concernent des états : prince_john désire être riche, et friar_tuck désire être pauvre, peu importe l'action.

5.2 Processus d'évaluation

Les connaissances de l'agent à propos des actions A sont décrites sous la forme d'un nom d'action associé à des ensembles (potentiellement vides) de conditions et conséquences. Par exemple, l'action `give` est décrite comme :

```

action(give).
condition(give,A,B):-
  agent(B), agent(A), A!=B, not poor(A).
consequence(give,A,B,rich,B):- agent(A), agent(B).
consequence(give,A,B,poor,A):- agent(A), agent(B).

```

Une condition est constituée d'une conjonction de croyances. La conséquence d'une action est une clause composée de nouvelles croyances générées par l'action et l'agent concerné par ces conséquences. L'évaluation de la désirabilité DE (voir 6) déduit l'ensemble \mathcal{A}_d des actions désirées. Une action est dans \mathcal{A}_d si elle est directement désirée (dans \mathcal{D}) ou si ses conséquences sont des états désirés :

```

desirableAction(A, B, X, C):-
  desireState(A,S,D), consequence(X,B,C,S,D).

```

L'évaluation des capacités CE (voir 6) évalue les actions réalisables \mathcal{A}_c à partir des croyances et connaissances de l'agent. Une action est réalisable si ses conditions sont satisfaites.

```

possibleAction(A,X,B):- condition(X,A,B).

```

5.3 Processus moral

Dans le processus moral, les supports de valeurs VS sont implémentés de la manière suivante :

```

generous(A,give,B) :- A != B, agent(A), agent(B).
-generous(A,steal,B) :- A != B, agent(A), agent(B).

```

Nous pouvons alors exprimer les règles morales de l'agent en fonction de l'approche éthique (voir les sections 2.1 et 3.3). Un exemple de règle morale pour l'approche vertueuse est :

```

moral(robin_hood,A,X,B):-
  generous(A,X,B), poor(B), action(X).

```

L'évaluation de moralité ME donne l'ensemble des actions morales \mathcal{A}_m (voir section 3.3) :

```

moralAction(A,X,B):- moral(A,A,X,B).
-moralAction(A,X,B):- -moral(A,A,X,B).

```

et produit le résultat :

```

moralAction(robin_hood,give,paul)
-moralAction(robin_hood,tax,paul)

```

Nous ne présentons ici qu'une approche vertueuse. Toutefois, des implémentations d'approches déontologiques et conséquentialistes sont également disponibles en ligne.

5.4 Processus éthique

Afin d'évaluer chaque action, nous définissons plusieurs principes éthiques naïfs illustrant des priorités entre morale et désirs. Par exemple, l'exemple ci-dessous nommé `perfAct` illustre les critères d'une action parfaite (i.e. parfaitement morale, désirable et possible) :

```

ethPrinciple(perfAct,A,X,B):-
  possibleAction(A,X,B),
  desirableAction(A,A,X,B),
  not -desirableAction(A,A,X,B),
  moralAction(A,X,B),
  not -moralAction(A,X,B).

```

Principe \ Intention	perfAct	dutNR	desNR	dutFst	nR	desFst
give,paul	⊥	⊤	⊥	⊤	⊤	⊥
give,little_john	⊥	⊥	⊥	⊥	⊤	⊥
give,marian	⊥	⊥	⊥	⊥	⊤	⊥
give,prince_john	⊥	⊥	⊥	⊥	⊤	⊥
give,peter	⊥	⊥	⊥	⊥	⊤	⊥
steal,little_john	⊥	⊥	⊥	⊥	⊤	⊥
steal,marian	⊥	⊥	⊥	⊥	⊤	⊥
steal,prince_john	⊥	⊥	⊤	⊥	⊤	⊤
steal,peter	⊥	⊥	⊤	⊥	⊤	⊤
court,marian	⊥	⊥	⊤	⊥	⊤	⊤
wait,robin_hood	⊥	⊥	⊥	⊥	⊤	⊥

FIGURE 2 – Évaluation éthique \mathcal{E} des actions

Si paul est le seul agent pauvre, marian n'est pas mariée et robin_hood n'est pas pauvre, robin_hood obtient l'évaluation présentée en Figure 2. Tous les principes sont ordonnés conformément aux préférences :


```

prefEthics(robin_hood,perfAct,dutNR).
prefEthics(robin_hood,dutNR,desNR).
prefEthics(robin_hood,desNR,dutFst).
prefEthics(robin_hood,dutFst,nR).
prefEthics(robin_hood,nR,desFst).
prefEthics(A,X,Z):- prefEthics(A,X,Y), prefEthics(A,Y,Z).

```

Ces lignes décrivent l'ordre des principes éthiques et la transitivité de la relation de préférence (ici $\text{perfAct} >_e \text{dutNR} >_e \text{desNR} >_e \text{dutFst} >_e \text{nR} >_e \text{desFst}$). Enfin, le jugement J est représenté par :

```

existBetter(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  prefEthics(A,PE2,PE1),
  ethPrinciple(PE2,A,Y,C).
ethicalJudgment(PE1,A,X,B):-
  ethPrinciple(PE1,A,X,B),
  not existBetter(PE1,A,X,B).

```

Ainsi, l'action juste a_r pour `robin_hood` est `give`, `paul` qui respecte `dutNR`.

5.5 Jugement éthique multi-agent

Afin de permettre un jugement aveugle, nous introduisons une nouvelle croyance sur le comportement d'un autre agent, puis `robin_hood` compare ses actions justes et cette croyance pour juger `little_john` :

```

done(little_john,give,peter).

blindJudgment(A,ethical,B):-
  ethicalJudgment(_,A,X,C), done(B,X,C), A!=B.
blindJudgment(A,unethical,B):-
  not blindJudgment(A,ethical,B),
  agent(A), agent(B), done(B,_,_), A!=B.

```

Dans cet exemple, l'action `give` envers `peter` n'est pas dans \mathcal{A}_r . Alors `little_john` est jugé non-éthique par `robin_hood`.

Pour passer à un jugement partiel, nous remplaçons une partie des connaissances et états de `robin_hood` par ceux de `little_john`. Avec les croyances de `little_john` (qui pense que `peter` est pauvre et que `paul` est riche), `robin_hood` le juge éthique.

Enfin, pour un jugement pleinement informé, nous remplaçons les croyances, désirs et connaissances de `robin_hood` par ceux de `little_john`. Alors, `robin_hood` est capable de reproduire la totalité du jugement éthique de `little_john` et de comparer les deux jugements pour une même action.

6 Travaux similaires

Nous avons adopté dans ce travail une approche pleinement rationaliste (fondée exclusivement sur du raisonnement, sans émotions). Cependant d'autres travaux ont proposé des démarches

similaires [27, 6]. La principale spécificité de notre travail est d'éviter toute intervention des émotions afin de pouvoir justifier la conformité d'un comportement avec une éthique donnée par un ensemble de déductions rationnelles fondé sur un ensemble de valeurs et règles morales et de principes éthiques explicitement définis.

D'un côté, [6] suggère une approche intuitionniste qui évalue les plans en fonction des émotions provoquées. Les valeurs ne sont alors vues que comme des sources d'émotions et la construction des plans est influencée par leur anticipation. De notre point de vue, les valeurs et désirs doivent être séparés car nous souhaitons des agents capables de faire cette distinction et expliquer comment ils les concilient.

D'un autre côté, [27] est une approche logique modélisant le raisonnement moral par des contraintes déontiques. Ce modèle est une manière d'implémenter une théorie du bien et à fait l'objet de l'implémentation d'un système de vérification automatique de la moralité d'un comportement [10]. Cependant le raisonnement éthique n'est considéré dans [27] que comme un méta-raisonnement et envisagé comme un choix de relaxation de contraintes sur le comportement. Nous préférons permettre l'implémentation de principes éthiques pour répondre à ces situations qui sont justement les plus complexes.

7 Conclusion

Afin d'agir collectivement en conformité avec une éthique et une morale donnée, un agent autonome à besoin d'être capable d'évaluer la justesse et la moralité de son comportement et de celui des autres. En se basant sur des concepts de philosophie morale, nous proposons dans cet article une faculté de jugement générique pour les agents autonomes. Ce processus utilise des représentations explicites d'éléments tels que des principes éthiques et des valeurs et règles morales. Nous illustrons la manière dont ce modèle compare les éthiques de différents agents. De plus, ce modèle de jugement éthique a été conçu afin de pouvoir être incorporé à une architecture existante pour fournir une composante éthique à un processus de décision. Comme ce processus de jugement peut être employé sur des informations partagées par un collectif d'agent, ce travail définit une ligne directrice vers une étude de la notion d'éthique collective.

Bien que cet article présente un cadre pour représenter une éthique et l'utiliser pour effectuer un jugement, le modèle reste fondé sur une approche qualitative. Et bien que l'on puisse définir diverses valuations morales, nous n'avons pas expérimenté l'introduction d'incertitudes sur les conséquences ou la représentation d'un degré d'éthique. De plus, la notion de principe

éthique reste à approfondir pour refléter l'immense variété des théories philosophiques.

La suite de ce travail pourrait consister en la représentation de divers codes de déontologies dans des domaines applicatifs (ex : en éthique médicale ou financière) afin de montrer la généralité de notre approche. De plus, ce modèle devrait être adapté à des évaluations quantitatives pour évaluer plus finement la proximité entre un comportement et l'attitude exemplaire dans une situation. Enfin il serait intéressant de définir une notion de degré de similarité entre des éthiques pour permettre à un agent de se positionner plus facilement face à leur diversité.

Remerciements

Les auteurs remercient l'Agence Nationale de la Recherche (ANR) pour sa contribution financière sous la référence ANR-13-CORD-0006.

Références

- [1] Ethical judgment. Free Online Psychology Dictionary, August 2015.
- [2] L. Alexander and M. Moore. Deontological ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring edition, 2015.
- [3] M. Anderson and S.L. Anderson. Toward ensuring ethical behavior from autonomous systems : a case-supported principle-based paradigm. *Industrial Robot : An International Journal*, 42(4) :324–331, 2015.
- [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] C. Battaglino, R. Damiano, and L. Lesmo. Emotional range in value-sensitive deliberation. In *AAMAS*, pages 769–776, 2013.
- [7] G. Boella, G. Pigozzi, and L. van der Torre. Normative systems in computer science - Ten guidelines for normative multiagent systems. In *NorMAS*, Dagstuhl Seminar, 2009.
- [8] 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*, 2010.
- [9] A. Damasio. *Descartes' error : Emotion, reason and the human brain*. Random House, 2008.
- [10] L.A. Dennis, M. Fisher, and A.F.T. Winfield. Towards verifiably ethical robot behaviour. In *1st International Workshop on AI and Ethics*, 2015.
- [11] P. Foot. The problem of abortion and the doctrine of the double effect. *Oxford Review*, pages 5–15, 1967.
- [12] J.-G. Ganascia. Ethical system formalization using non-monotonic logics. In *29th Annual Conference of the Cognitive Science Society*, pages 1013–1018, 2007.
- [13] B. Gert. The definition of morality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015.
- [14] J. Greene and J. Haidt. How (and where) does moral judgment work? *Trends in cognitive sciences*, 6(12) :517–523, 2002.
- [15] R. Hursthouse. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2013.
- [16] R. Johnson. Kant's moral philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014.
- [17] K.-J. Kim and H. Lipson. Towards a theory of mind in simulated robots. In *11th ACC-GECC*, pages 2071–2076, 2009.
- [18] W. Mao and J. Gratch. Modeling social causality and responsibility judgment in multi-agent interactions. In *IJCAI*, pages 3166–3170, 2013.
- [19] T. McConnell. Moral dilemmas. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014.
- [20] A. McIntyre. Doctrine of double effect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2014.
- [21] Society of Professional Journalists. Code of ethics, September 2014.
- [22] P. Ricoeur. *Oneself as another*. University of Chicago Press, 1995.
- [23] A. Saptawijaya and L. Moniz Pereira. Towards modeling morality computationally with logic programming. In *Practical Aspects of Declarative Languages*, pages 104–119. 2014.
- [24] S.H. Schwartz. Basic human values : Theory, measurement, and applications. *Revue française de sociologie*, 47(4) :249–288, 2006.
- [25] M. Timmons. *Moral theory : an introduction*. Rowman & Littlefield, 2012.
- [26] S.A. Walter. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2015.
- [27] V. Wiegel and J. van den Berg. Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, 1(3) :233–242, 2009.