Jugement éthique dans le processus de décision d'un agent BDI

Nicolas Cointe [1,2], **Grégory Bonnet** [2], **Olivier Boissier** [1]

1. Institut Henri Fayol, Laboratoire Hubert Curien
UMR CNRS 5516, Mines Saint-Étienne, Saint-Étienne, 42023 France
nicolas.cointe@mines-stetienne.fr, olivier.boissier@mines-stetienne.fr

2. Équipe Modèle Agent Décision
GREYC, département Intelligence Artificielle et Algorithmique
CNRS UMR 6072 F-14032, Normandie Université, Caen, France
gregory.bonnet@unicaen.fr

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 des décisions sur son propre comportement et du comportement 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 au sein de la plateforme de programmation orientée multi-agent JaCaMo illustre ces fonctionnalités.*

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 the multi-agent oriented programming platform JaCaMo and based on a simple scenario is given to illustrate those functionalities.*

MOTS-CLÉS : *agent (architecture), systèmes multi-agents éthiques.*

KEYWORDS: *agent (architecture), multi-agent ethics.*

DOI:10.3166/RIA.31.471-499

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 (Wiegel, Berg, 2009), de simples interdits ou obligations (Boella *et al.*, 2009), les codes de conduite font appel à des notions plus complexes telles que des valeurs morales ou des principes éthiques. Un traitement explicite de concepts tels que l'altruisme ou la générosité ainsi que des principes s'y attachant nécessite des structures et des fonctions spécifiques dans l'architecture d'un agent autonome. De plus, 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 mettent en œuvre des agents en interaction. Il est donc nécessaire de prendre aussi en compte un point de vue collectif et multi-agent.

En effet, un point de vue purement individuel pourrait être suffisant pour permettre à un agent d'agir de manière éthique. 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 ou de l'organisation à laquelle ils appartiennent. Ainsi par exemple, un agent pourrait évaluer quels sont les autres agents du système dont le comportement est en accord avec sa propre éthique afin de choisir ceux avec lesquels collaborer sans avoir à enfreindre ses principes. À une échelle collective, cela donne la possibilité d'évaluer une organisation par le jugement du comportement de ses membres afin de décider d'y prendre part. 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 règles et valeurs morales et de principes éthiques. Nous proposons un modèle générique de jugement du comportement d'un agent autonome. Ce modèle peut être utilisé par un agent tant pour guider son propre comportement que pour évaluer celui des autres.

Cet article, qui est une version retravaillée et étendue de travaux déjà publiés (Cointe *et al.*, 2016a ; 2016b), est organisé de la manière suivante : la section 2 introduit quelques concepts de philosophie morale et un court état de l'art sur des approches d'agents autonomes éthiques. Il s'agit principalement de fixer le vocabulaire et le cadre sur lequel s'appuie notre proposition et de la positionner dans les approches actuelles dans le domaine des systèmes multi-agents. Nous détaillons dans la section 3 le modèle de jugement éthique que nous proposons. La section 4 décrit ensuite son utilisation par un agent dans son interaction avec les autres en fonction des informations disponibles et permet de définir différents types de jugement de l'éthique des autres. La section 5 présente une preuve de concept sur un système multi-agent de gestion d'actifs financiers. Nous concluons en section 6 en précisant l'intérêt de ces travaux pour des systèmes multi-agents et en donnant des perspectives pour des travaux futurs.

2. Éthique et agents autonomes

La section 2.1 introduit des concepts de philosophie morale sur laquelle se fonde notre proposition, puis nous passons en revue en section 2.2 les architectures d'agent existantes proposant l'implémentation de comportements éthiques. La section 2.3 identifie les spécificités de notre approche que nous comparons aux approches semblables dans la section 2.4.

2.1. Concepts de philosophie morale

Des philosophes antiques aux travaux récents de neurologie (Damasio, 2008) et sciences cognitives (Greene, Haidt, 2002), 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 connait 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 un raisonnement permettant de distinguer les bonnes et mauvaises actions. Notons que la morale se distingue de la loi et du système légal. En effet, la morale ne comporte pas de pénalités explicites ou de règles officiellement établies (Gert, 2015). 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 évènement. Certaines valeurs morales semblent universelles et ces valeurs dites fondamentales existent en nombre fini, seule l'importance qui leur est accordée varie selon les individus (Rokeach, 1973 ; S. H. Schwartz, Bilsky, 1990 ; S. H. Schwartz, 2006 ; S. Schwartz, 2012). En philosophie, la théorie des valeurs, au sens d'axiologie (Schroeder, 2016), vise à comprendre comment les valeurs morales interviennent dans le jugement. Toutefois, elles sont structurées hiérarchiquement au sein d'un *système de valeurs* – propre à un individu ou un groupe – selon un ordre d'importance relative au regard d'un contexte particulier (van Marrewijk, Werre, 2003 ; Wiener, 1988).

Un ensemble de règles morales et de valeurs morales établit une *théorie du bien* qui permet à chacun d'évaluer la moralité d'un comportement. Cependant, dans certaines situations, les règles morales peuvent être contradictoires, ce qui rend nécessaire un niveau de raisonnement supplémentaire. Un ensemble de principes éthiques forme

la *théorie du juste* qui définit des critères pour reconnaitre le choix le plus juste ou le plus acceptable. Cette seconde théorie est également appelée *théorie de la juste conduite* (Timmons, 2012). Tandis que la *théorie du bien* indique pourquoi une action est morale ou non dans une situation donnée, la *théorie du juste* ne permet de qualifier une action de juste que par rapport aux autres actions disponibles. Par exemple, bien que le vol soit souvent reconnu comme immoral (au regard d'une théorie du bien), de nombreuses personnes s'accorderont à reconnaitre qu'il est acceptable qu'un orphelin pauvre et affamé dérobe une pomme dans un supermarché (au regard d'une certaine théorie du juste). Cette même action peut devenir injuste si l'ensemble des actions disponibles est modifié (pour reprendre l'exemple précédent : s'il devient possible pour l'orphelin de se procurer la pomme de manière honnête, il n'est plus juste de la voler). 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 (Ricoeur, 1995), nous admettons la définition suivante :

DÉFINITION 2 (Éthique). — *L'éthique est la combinaison de* principes éthiques *et de règles morales permettant à un modèle 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 (Johnson, 2014) ou la Doctrine du Double Effet de St Thomas d'Aquin (McIntyre, 2014), 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 (Hursthouse, 2013) ; l'*éthique déontologique* juge un comportement par sa conformité avec des obligations et permissions associées à des situations (Alexander, Moore, 2015) ; l'*éthique conséquentialiste* juge un comportement à la moralité de ses conséquences (Walter, 2015).

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 (McConnell, 2014). Chaque option apportera un regret. De nombreux dilemmes, tels que le dilemme du trolley (Foot, 1967), 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 (*Ethical Judgment*, 2015) 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 éthiques, pour soi-même ou autrui.*

Enfin, le jugement permet également d'évaluer la justesse de l'action d'autrui en se projetant dans sa situation et en examinant l'action commise. La prise en compte de ce jugement dans les interactions futures avec l'agent jugé permet de construire une relation de confiance et adapter son comportement à celui de l'autre. Le jugement qu'un individu porte sur les autres et celui que les autres portent sur lui-même engendrent la honte ou la reconnaissance, qui sont des mécanismes régulateurs de la vie en société (Scheff, 2003). Il participe également à la notion de perfectionnisme moral tel qu'envisagé en psychologie (Stoeber, Yang, 2016).

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'*éthique gérée au sein d'une architecture cognitive*.

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 donnant lieu à 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 (Arkin, 2009)) dont la définition peut provenir d'une *conception sensible aux valeurs*. Cette approche, aussi appelée *Value Sensitive Design*, se propose d'identifier clairement durant la phase de conception des valeurs sociales ou morales désirées et de s'assurer que les logiciels ou systèmes conçus les respectent bel et bien (Friedman, 1996 ; Friedman *et al*., 2013). L'inconvénient principal de l'éthique par conception est l'absence de représentation générique de concepts éthiques (théories du bien et du mal). De plus, deux agents programmés de cette manière ne peuvent pas comparer leurs éthiques par conception en raison de l'absence de représentations explicites. Cette approche est essentiellement conçue pour faciliter l'intégration dans un logiciel d'une éthique fixée à la conception en apportant des garanties, et éventuellement une preuve, sur la conformité de son comportement à un ensemble de règles. Une telle démarche peut sembler adaptée lorsqu'il n'est pas nécessaire pour les agents de raisonner sur le caractère éthique de leur comportement ou de celui des autres, mais cette rigidité devient problématique dans un système où des agents dotés d'éthiques et de morales diverses son amenés à coopérer. Le comporte-

ment du logiciel conçu n'est conforme à l'éthique du concepteur que dans un nombre fini de situations envisagées à la conception et se limite bien souvent à l'obéissance stricte à une déontologie dont les règles sont directement implémentées de manière sous formes d'entraves et de contraintes dans le modèle de décision de l'agent.

L'*éthique par étude de cas* cherche premièrement à inférer des règles éthiques statistiques à partir d'un vaste ensemble de jugements exprimés par des experts, puis à les appliquer pour produire un comportement éthique (Anderson, Anderson, 2014). 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 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, bien que ces techniques offrent l'avantage de pouvoir agréger de manière générique le résultats de jugements, la coopération entre agents hétérogènes en éthique 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 (Ganascia, 2007a ; 2007b ; Berreby *et al*., 2015 ; Saptawijaya, L. Moniz Pereira, 2014). 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'implicitement exprimée. Cette approche permet de juger une décision en prenant en compte un principe éthique.

Enfin, les *modèles de représentation de l'éthique au sein d'architectures cognitives* 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 (Arkoudas *et al*., 2005 ; Coelho, Rocha Costa, 2009 ; Coelho *et al*., 2010 ; Rocha-Costa, 2016). En particulier l'une de ces approches (Rocha-Costa, 2016) propose la construction de connaissances sur la morale d'un autre agent à partir de l'observation de son comportement et définit une notion de faits moraux (jugements, blâmes) comme des croyances pouvant être utilisées explicitement par les agents. Bien que ces approches permettent aux agents de manipuler des règles explicites et justifier leurs décisions, un modèle opérationnel n'a pas encore été implémenté.

2.3. Vers un modèle de jugement éthique pour agents autonomes

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 modèle 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 leur éthique individuelle (Kim, Lipson, 2009). 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 : des *valeurs morales* caractérisant des états du monde ou des actions, des *règles morales* caractérisant en bien ou mal des états du monde ou des actions, des *principes éthiques* caractérisant en termes de juste ou d'injuste des actions et des *préférences* exprimant une forme de système de valeurs. 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 *modèle 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, que nous appelons *connaissance morale*, ainsi que de principes et préférences éthiques, que nous appelons *connaissance éthique*. Ainsi, des agents capables de se construire une représentation des connaissances morales ou éthiques des autres agents peuvent être en mesure de les juger de manière informée. Enfin, ce jugement peut être utilisé dans les procédures de décisions, qu'elles soient individuelles ou collectives, de coopération et de confiance (Mao, Gratch, 2013).

Enfin, nous excluons la représentation des émotions au sein du raisonnement de l'agent pour fonder son jugement sur un modèle entièrement rationnel. En effet, nous souhaitons pouvoir *a posteriori* vérifier la conformité éthique de ses décisions en examinant son raisonnement et non une simulation émotionnelle. Ainsi, nous ne cherchons pas à imiter le raisonnement humain à la fois rationnel et intuitif (Greene, Haidt, 2002), mais à concevoir des agents raisonnant sur une éthique perçue comme un ensemble de règles logiques utilisant des connaissances spécifiques (valeurs hiérarchisées, règles morales, principes éthiques, etc.).

2.4. Positionnement et travaux similaires

Si nous avons adopté dans ce travail une approche pleinement rationaliste (fondée exclusivement sur du raisonnement sans émotions), d'autres travaux proposent une approche différente (Wiegel, Berg, 2009 ; Battaglino *et al.*, 2013). 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.

En effet, (Battaglino *et al*., 2013) se distinguent par 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.

De son côté, (Wiegel, Berg, 2009) proposent une approche logique modélisant le raisonnement moral par des contraintes déontiques. Ce modèle est une manière d'implémenter une théorie du bien et a fait l'objet de l'implémentation d'un système de vérification automatique de la moralité d'un comportement (Dennis *et al*., 2015). Cependant le raisonnement éthique n'est considéré dans (Wiegel, Berg, 2009) 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.

Dans 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. Modèle 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. Les sections suivantes montreront comment ce modèle de jugement peut être employé dans le modèle décisionnel d'un agent.

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 modèle de jugement éthique (EJP) utilise des connaissances sur l'évaluation de situation, la morale et l'éthique. Nous l'avons structuré en modèle de *Reconnaissance de situation*, modèle d'*Évaluation*, modèle *Moral* et modèle *Éthique* [1] (voir fig. 1). Dans cet article, nous considérons ce modèle dans le contexte d'un modèle d'architecture BDI (Bratman, 1987), 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.

1. Notons que la dénomination de ces deux modèles est un raccourci par rapport à leur fonction exacte qui est dédiée respectivement à l'évaluation de la moralité (modèle moral) et à l'évaluation du respect de l'éthique (modèle éthique) des actions considérées. Il ne s'agit nullement de prétendre que l'un est un modèle fonctionnant de manière morale et l'autre de manière éthique.

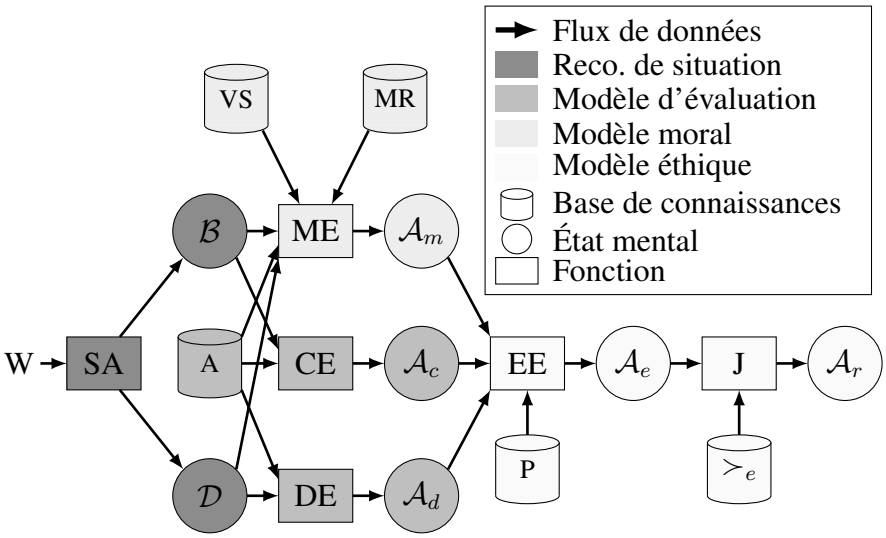

Figure 1. Modèle de jugement éthique

DÉFINITION 4 (Modèle de jugement éthique). — *Un modèle de jugement éthique (ou Ethical Judgment Process) EJP est défini comme une composition d'une reconnaissance de situation (AP), un modèle d'évaluation (EP), un modèle moral (GP), un modèle éthique (RP) et une ontologie $\mathcal{O}$ ($\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_m$) de valeurs morales ($\mathcal{O}_v$) et valuations morales ($\mathcal{O}_m$). Ce modèle de jugement éthique produit une évaluation des actions pour l'état courant du monde W en tenant compte de considérations morales et éthiques.*

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

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 $\mathcal{O}$ peuvent prendre la forme d'un ensemble d'éléments discrets tel que { bien, mal } ou continu comme un degré de moralité.

Comme nous allons le montrer dans la suite de cet article, le modèle de jugement peut être intégré comme nouveau composant du mécanisme de décision d'un agent BDI pour qu'un agent puisse décider de son comportement mais aussi comme composant permettant de juger du comportement des autres agents. Afin de mettre en place cette double utilisation, nous indiçons chacun des ensembles de données entrant dans le modèle par l'agent a_i, source de la représentation de ces données : ainsi X_{a_i} signifie que X regroupe un ensemble de données issues de l'agent a_i. Notons que l'ontologie $\mathcal{O}$ de valeurs et de valuations morales n'est pas indicée par a_i car nous considérons

qu'elle est commune à l'ensemble des agents du système. Cela permet aux agents d'employer les mêmes noms de valeurs et d'exprimer la moralité de leurs actions sur une même échelle de valuations, à la différence de MR et VS qui décrivent une morale pouvant être propre à chacun des agents(dépendant de contextes socio-culturels par exemple).

3.2. Reconnaissance de situation et évaluation

Dans ces modèle, l'agent commence par évaluer l'état du monde, c'est-à-dire à produire des croyances et désirs en appliquant un modèle de reconnaissance de situation à partir de l'état de l'environnement dans lequel l'agent est situé (l'environnement inclut également les autres agents du système).

DÉFINITION 5 (Modèle de reconnaissance de situation). — *Le modèle de reconnaissance de situation (ou Awareness Process) 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 $\mathcal{B}_{a_i}$ est l'ensemble des croyances de l'agent a_i sur W parmi l'ensemble B_{a_i} de ses croyances possibles, et $\mathcal{D}_{a_i}$ ses désirs à partir de W parmi l'ensemble D_{a_i} de ses désirs possibles :

$$SA : W \rightarrow \mathcal{B}_{a_i} \cup \mathcal{D}_{a_i}$$

À partir d'un ensemble de croyances $\mathcal{B}_{a_i}$ et d'un ensemble de désirs $\mathcal{D}_{a_i}$ (a_i comme expliqué ci-dessus peut désigner l'agent effectuant le modèle de jugement – dans ce cas il s'agit de ses propres croyances et désirs – ou un autre agent – dans ce cas il s'agit de la représentation que l'agent effectuant le jugement a sur les croyances et désirs de a_i) un agent exécute le modèle d'évaluation EP pour établir les actions désirables $\mathcal{A}_d$ d'une part (c.-à-d. les actions qui permettent de satisfaire un désir) et les actions exécutables $\mathcal{A}_c$ d'autre part (c.-à-d. les actions pouvant être effectuées dans l'état courant du monde). Ces actions sont déduites par raisonnement sur les conditions et conséquences des actions décrites dans A_{a_i}, c.-à-d. les actions à disposition de l'agent effectuant le jugement si a_i représente cet agent, ou la représentation des actions qu'un autre agent a_i peut réaliser.

DÉFINITION 6 (Modèle d'évaluation). — *Le modèle d'évaluation (ou Evaluation Process) EP produit les ensembles d'actions désirables et d'actions exécutables à partir des ensembles de désirs et croyances. Nous le définissons comme :*

$$EP = \langle A_{a_i}, \mathcal{A}_{d_{a_i}}, \mathcal{A}_{c_{a_i}}, DE, CE \rangle$$

où A_{a_i} est un ensemble d'actions qu'il s'agit de juger (chacune étant décrite comme une paire de conditions et conséquences portant sur les croyances et les désirs),

$\mathcal{A}_{d_{a_i}} \subseteq A_{a_i}$ et $\mathcal{A}_{c_{a_i}} \subseteq A_{a_i}$ sont respectivement l'ensemble des actions désirables et exécutables, DE et l'évaluation de capacités CE sont des fonctions telles que :

$$DE : 2^{D_{a_i}} \times 2^{A_{a_i}} \rightarrow 2^{A_{a_i}}$$

$$CE : 2^{B_{a_i}} \times 2^{A_{a_i}} \rightarrow 2^{A_{a_i}}$$

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. Ainsi, une action α est évaluée comme étant désirable si l'agent désire la réalisation de α ou la réalisation de ses conséquences (et inversement, elle peut être indésirable s'il désire que ces éléments ne se réalisent pas). L'action peut être désirable et indésirable simultanément si sa réalisation ou les conséquences de sa réalisation sont évaluées différemment. Notons ici qu'il est possible d'envisager que les conséquences d'une action α puissent être désirables en raison de connaissances sur une autre action α', désirable, et dont les conditions sont des conséquences de α.

Les conditions d'une action α permettent également de savoir si l'action est exécutable dans le contexte courant décrit par l'ensemble des croyances.

Par la suite, nous désignons par CK_{a_i} l'union des croyances $\mathcal{B}_{a_i}$ et désirs $\mathcal{D}_{a_i}$. Il s'agit des connaissances contextuelles (ou *Contextual Knowledge*) d'un agent a_i sur l'état du monde.

Maintenant que nous avons défini les modèle de reconnaissance de situation et d'évaluation, nous pouvons aborder les modèle au cœur du modèle de jugement : le modèle moral qui emploie les règles morales et valeurs morales, le modèle éthique qui emploie les principes éthiques.

3.3. Modèle 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 modèle moral ci-dessous.

DÉFINITION 7 (Modèle moral). — *Le modèle moral (ou Goodness Process) GP identifie les actions morales à partir des croyances, désirs et connaissances sur les actions d'un agent a_i ainsi que de ses valeurs et règles morales. Nous le définissons formellement comme :*

$$GP = \langle VS_{a_i}, MR_{a_i}, \mathcal{A}_{m_{a_i}}, ME \rangle$$

où VS_{a_i} est la base de connaissances du support de valeurs de l'agent a_i, MR_{a_i} est sa base de connaissances de règles morales, $\mathcal{A}_{m_{a_i}} \subseteq A_{a_i}$ est l'ensemble de ses actions morales[2]*. La fonction d'évaluation morale ME est:*

$$ME : 2^{D_{a_i}} \times 2^{B_{a_i}} \times 2^{A_{a_i}} \times 2^{VS_{a_i}} \times 2^{MR_{a_i}} \rightarrow 2^{A_{a_i}}$$

Afin de réaliser ce modèle 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 promeut une valeur correspondante. On peut envisager diverses combinaisons pour une même valeur morale : par 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 le contraire » (pour éviter les mensonges par omission).

DÉFINITION 8 (Support de valeur). — *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 avec $\alpha \in A_{a_i}$, $w \subset B_{a_i} \cup D_{a_i}$.*

EXEMPLE 9. — Les supports 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 représentés par :

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

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

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

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. Ici « amoral » est un élément de référence de cet ensemble permettant de préciser que la morale est indifférente à une action (ce qui est différent de l'inexistence de connaissances sur la moralité d'une action).

DÉFINITION 10 (Règle morale). — *Une règle morale est un n-uplet $\langle w, o, m\rangle \in MR_{a_i}$ où w est un état du monde décrit par $w \subset CK_{a_i}$ interprété comme une conjonction de croyances et désirs, $o = \langle a, v\rangle$ où $a \in A_{a_i}$ et $v \in \mathcal{O}_v$, et $m \in \mathcal{O}_m$ est une valeur morale décrite dans $\mathcal{O}_m$ qui qualifie o quand w est l'état courant.*

EXEMPLE 11. — Certaines règles morales classiques telles que « tuer un humain est immoral » ou « être honnête avec un menteur est plutôt moral » peuvent être représentées comme :

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

2. Notons que $A_{m_{a_i}} \nsubseteq A_{d_{a_i}} \cup A_{c_{a_i}}$ car une action peut être morale en soi, même si elle n'est pas désirée ou réalisable (ex : sauver le monde).

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

□

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 spécifiques décrivant des devoirs ou des interdits (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]). La définition 10 laissant la possibilité d'exprimer la moralité d'une action en fonction d'un contexte, de supports de valeurs ou de conséquences, ces trois approches sont compatibles avec le modèle proposé ici.

Par la suite, nous ferons référence à ces différentes connaissances (règles morales MR_{a_i}, support de valeurs VS_{a_i} et valeurs $\mathcal{O}_v$) utilisées dans le modèle moral de l'agent a_i sous l'appellation de connaissance du bien, notée GK_{a_i}

3.4. Modèle éthique

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

DÉFINITION 12 (Modèle éthique). — *Un modèle éthique (ou Rightness Process) RP produit les actions justes selon une représentation donnée de l'éthique. Nous définissons ce modèle comme :*

$$RP = \langle P_{a_i}, \succ_{e_{a_i}}, \mathcal{A}_{r_{a_i}}, EE, J \rangle$$

où P_{a_i} est la base de connaissances sur les principes éthiques de l'agent a_i, $\succ_{e_{a_i}} \subseteq P_{a_i} \times P_{a_i}$ est un ensemble de relations de préférences représentant un ordre total sur ces principes. Les deux fonctions sont EE (évaluation éthique) et J (jugement) permettent respectivement de construire l'ensemble $\mathcal{A}_{e_{a_i}}$ des actions éthiques, c'est-

3. Extrait de (Professional Journalists, 2014), section « Act Independently ».
4. Code de déontologie médicale, article 31.

à-dire conformes aux principes éthiques et l'ensemble $\mathcal{A}_{r_{a_i}} \subseteq A_{a_i}$ *des actions justes à partir de* $\mathcal{A}_{e_{a_i}}$ *et des préférences :*

$$EE : 2^{\mathcal{A}_{d_{a_i}}} \times 2^{\mathcal{A}_{c_{a_i}}} \times 2^{\mathcal{A}_{m_{a_i}}} \times 2^{P_{a_i}} \rightarrow 2^{\mathcal{A}_{a_i}}$$

où $\mathcal{A}_{e_{a_i}} = A_{a_i} \times P_{a_i} \times \{\bot, \top\}$

$$J : 2^{\mathcal{A}_{e_{a_i}}} \times 2^{\succ_{e_{a_i}}} \rightarrow 2^{\mathcal{A}_{r_{a_i}}}$$

Nous représentons chaque principe éthique par une fonction – inspirée d'une théorie philosophique – qui estime s'il est juste ou non d'effectuer une action dans une situation donnée au regard de cette théorie.

La fonction d'évaluation éthique EE renvoie l'évaluation de toutes les actions désirables ($\mathcal{A}_{d_{a_i}}$), réalisables ($\mathcal{A}_{c_{a_i}}$) ou morales ($\mathcal{A}_{m_{a_i}}$) étant donné l'ensemble P_{a_i} des principes éthiques connus.

DÉFINITION 13 (Principe éthique). — *Un principe éthique (ou Ethical Principle)* $p \in P_{a_i}$ *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 :*

$$p : 2^{A_{a_i}} \times 2^{\mathcal{B}_{a_i}} \times 2^{\mathcal{D}_{a_i}} \times 2^{MR_{a_i}} \times 2^{V_{a_i}} \rightarrow \{\top, \bot\}$$

EXEMPLE 14. — Par exemple, considérons trois agents dans la situation suivante inspirée de la critique de Benjamin Constant de l'Impératif Catégorique de Kant (1967).

Un agent A est caché chez un agent B pour échapper à un agent C, et C vient demander à B où se trouve A pour le tuer. Les règles morales de B sont mr_1 : « mettre autrui en danger est immoral » et mr_2 : « mentir est immoral ». B sait qu'il sera tué à la place de A s'il refuse de répondre. B désire éviter tout problème avec C. B connait la vérité et doit choisir l'une des trois actions suivantes : dire la vérité à C (satisfaisant ainsi mr_2 et son désir), mentir (satisfaisant mr_1 et son désir) ou refuser de répondre (satisfaisant les deux règles morales mais pas son désir). B connait deux principes éthiques (implémentés en P comme fonctions) : P1 pour lequel une action est juste si elle est possible, motivée par au moins une règle morale ou un désir et P2 pour lequel une action est juste si elle est possible et n'enfreint aucune règle morale.

L'évaluation de l'éthique de B renvoie les n-uplets donnés par le tableau 1 où chaque ligne représente une action et chaque colonne représente un principe éthique.

Étant donné un ensemble d'actions issues de l'évaluation éthique $\mathcal{E}$, le jugement J est la dernière étape qui sélectionne l'action juste à effectuer, au regard d'un ensemble de préférences éthiques (définissant un ordre total sur les principes éthiques).

Pour poursuivre notre exemple, supposons que les préférences éthiques de l'agent B sont P1 $\succ_e$ P2 et que J utilise une règle de bris d'égalité basée sur l'ordre lexicographique. Ici le principe préféré, P1 considère que chacune des actions est éthique. Cependant, « refuser de répondre » est l'action juste car elle satisfait également P2 à

Tableau 1. Évaluation éthique des actions de B

Action \ Principe	P1	P2
dire la vérité	$\top$	$\bot$
mentir	$\top$	$\bot$
refuser	$\top$	$\top$

l'inverse de « mentir » ou « dire la vérité ». Notons que ce jugement pourrait faire apparaître un dilemme entre « dire la vérité » et « refuser de répondre » en l'absence de règle de bris d'égalité (c'est-à-dire la seule prise en considération du principe préféré). □

Par la suite, nous ferons référence à ces différentes connaissances (c'est-à-dire principes éthiques P_{a_i} et préférences éthiques $\succ_{e_{a_i}}$) utilisées dans le modèle éthique de l'agent a_i sous l'appellation de connaissance du juste, notée RK_{a_i}.

4. Usage du modèle de jugement éthique

Nous avons évoqué au travers des exemples employés pour illustrer le processus de jugement éthique décrit dans la section précédente comment ce modèle peut permettre à un agent a_i de juger de l'action la plus éthique à effectuer au regard de ses propres connaissances CK_{a_i}, GK_{a_i} et RK_{a_i}. Toutefois, ce modèle peut aussi être employé pour juger le comportement d'un autre agent de manière plus ou moins informée en se projetant à la place de l'agent jugé a_j. Dans un processus de jugement éthique EJP tel que défini dans la section précédente, les états mentaux des éléments de CK_{a_i}, GK_{a_i} et RK_{a_i} peuvent être échangés entre agents. Comme discuté en début de section précédente, l'ontologie $\mathcal{O}$ est considérée comme une connaissance commune, même si nous pourrions envisager dans des travaux futurs la coexistence de plusieurs ontologies. Ces connaissances peuvent être échangées de nombreuses manières comme du partage ou de l'inférence, les modalités de cet échange ne sont pas abordées dans cet article.

Nous distinguons quatre catégories de jugement : (1) le *jugement pour la décision* dans lequel l'agent juge de ses propres actions pour décider celle qui doit être réalisée ; (2) le *jugement aveugle* dans lequel l'agent juge a_j n'a pas d'autre information sur l'agent jugé a_i que son comportement observé ; (3) le *jugement partiellement informé* lorsque le juge a_j dispose d'informations partielles sur les connaissances de l'agent jugé a_i ; (4) le *jugement parfaitement informé* lorsque l'agent juge a_j dispose de la totalité des informations existantes sur l'agent jugé a_j.

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'humain à la théorie de l'esprit (Kim, Lipson, 2009) (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 pour la décision

Un premier usage du jugement consiste à l'intégrer dans le modèle de décision d'un agent autonome. Pour ce faire, l'agent doit être conçu de manière à ce que seules les actions jugées éthiques puissent être décidées par l'agent. Ainsi, l'agent présenterait un comportement qu'il juge éthique à tout moment. Notons toutefois que si ses connaissances (A_{a_i}, CK_{a_i}, GK_{a_i} ou RK_{a_i}) évoluent au cours du temps, ce jugement pourrait être contredit par un nouveau jugement. L'usage du modèle de jugement est illustré en section 5 dans une preuve de concept.

4.2. Jugement éthique aveugle

Un second type de jugement peut être effectué sur un autre agent 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 $\langle P_{a_j}, \succ_{e,a_j}\rangle$ afin d'évaluer le comportement de l'agent jugé a_i. C'est un jugement *a priori* et a_i est jugé comme ayant effectué une action injuste ou immorale si $\alpha_{a_i} \notin \mathcal{A}_{r,a_j}$ ou $\alpha_{a_i} \notin \mathcal{A}_{m,a_j}$.

4.3. Jugement éthique partiellement informé

Le troisième type de jugement tient 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 connaissance contextuelle (c.-à-d. CK_{a_j}), (*ii*) de la connaissance du bien (c.-à-d. GK_{a_j}) et A_{a_j} ou (*iii*) de la connaissance du juste (c.-à-d. RK_{a_j}) de l'agent jugé. Remarquons que, dans le second cas, A_{a_i} est nécessaire car, à l'inverse des principes éthiques, les règles morales peuvent porter directement sur des actions spécifiques.

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

Jugement considérant la situation

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

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 consistantes avec sa propre théorie du bien (c.-à-d. 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 responsabilité 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) [6]. Cela permet d'évaluer comment l'agent a_i concilie ses désirs et sa morale dans une situation en comparant l'ensemble des actions justes $\mathcal{A}_{r,a_j}$ et $\mathcal{A}_{r,a_i}$ respectivement générées en utilisant P_{a_j}, $\succ_{e,a_j}$ et P_{a_i}, $\succ_{e,a_i}$. Par exemple, si $\mathcal{A}_{r,a_j} = \mathcal{A}_{r,a_i}$ 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.

4.4. 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

6. Si la situation et la théorie du bien sont toutes deux partagées, il s'agit d'un jugement pleinement informé (voir 4.4).

connaissances de l'agent jugé. Un jugement pleinement informé 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 système multi-agent à l'aide du framework JaCaMo (Boissier *et al.*, 2013), où les agents sont développés en Jason et les artefacts utilisent Cartago. Le code source est disponible en ligne [7]. L'environnement est une simulation de marché financier sur lequel des actifs sont cotés, achetés et vendus par un ensemble d'agents autonomes. Les agents éthiques emploient leur modèle de jugement en guise de modèle de décision (comme présenté en section 4.1) pour sélectionner et effectuer des investissements. La section 5.1 introduit le contexte de la gestion éthique d'actifs financiers et les fonctionnalités de notre simulation. La morale et l'éthique utilisées dans cette application sont définies en section 5.2.

5.1. Finance éthique

L'échange d'actifs sur des marchés financiers fait l'objet de nombreuses réflexions éthiques [8]. L'une des principales raisons à ces interrogations provient du fait que les décisions des agents autonomes, auxquels sont confiés les décisions d'achats et de ventes d'actifs appartenant à des humains, impactent l'économie réelle (Koopman, Székely, 2009). Certains analystes considèrent l'usage de techniques d'automatisation des activités financières comme introduisant en soi de nombreux effets pervers tels que des formes de manipulation de marchés, de concurrence déloyale envers les petits investisseurs et des brusques mouvements baissiers par effet cascade. D'autres arguent que cela réduit la volatilité, accroit la transparence, la liquidité et la stabilité des marchés avec un coût d'exécution des ordres plus faible (Aldridge, 2009). Comme le montrent certains rapports (Bono *et al.*, 2013), de nombreux fonds d'investissements sont intéressés par la promotion de certaines valeurs morales au travers de leurs pratiques. Les fonds d'investissement éthiques se multiplient et semblent peu à peu prendre une place significative sur les marchés. Toutefois, si des indicateurs objectifs efficaces permettent de mesurer la performance de ces fonds en termes de rendement, l'éthique de leur comportement reste plus difficile à quantifier et reste au moins partiellement sensible aux valeurs propres à l'évaluateur.

Dans cette preuve de concept illustrée par la figure 2, nous considérons une place de marché sur laquelle des agents autonomes peuvent échanger des actifs présents dans leurs portefeuilles. Ces actifs sont à la fois des devises et des participations de capitaux privés. Une place de marché est représentée au sein d'un agent par un tuple

7. https://cointe.users.greyc.fr/download/
8. http://sevenpillarsinstitute.org/

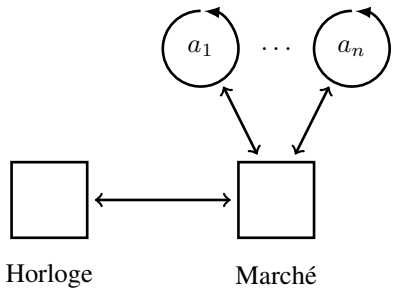

Figure 2. Architecture du système de place de marché pour la preuve de concept de notre proposition : des agents a_i interagissent sur une place de marché partagée dont le fonctionnement est rythmé par une horloge commune

$\langle$ `name`, `id`, `type`, `matching` $\rangle$ comprenant le nom du marché (`name`), un code unique d'identification (`id`), le type d'actifs échangés (`type`) et l'algorithme utilisé pour faire correspondre les ordres d'achat et de vente (`matching`).

L'ensemble des actions à la disposition d'un agent sont des ordres d'achat (`buy`), de vente (`sell`) ou d'annulation (`cancel`). Ils consistent respectivement en l'échange de devises contre des participations, l'échange de participations contre des devises et l'annulation d'un ordre en attente d'exécution sur le marché. Pour chaque ordre de vente ou d'achat, l'agent peut spécifier un prix limite ou accepter le prix actuel du marché. De même, le volume de titres échangés est précisé lors de l'envoi de l'ordre. Les titres de participation sont cotés sur le marché par des structures classiques de *Central Limit Order Book* (CLOB) (Aldridge, 2009). Un CLOB conserve et trie par ordre de prix l'ensemble des ordres d'achat et de vente, placés respectivement du côté de la demande (*bid*) et de l'offre (*ask*) du marché. Chaque agent peut ainsi insérer ses ordres d'un côté ou de l'autre et le CLOB se comporte alors selon les règles simples qui suivent :

– Si aucun ordre ne se trouve de l'autre côté au prix correspondant à l'ordre inséré, l'ordre arrivant est inséré,

– Si un ordre est présent de l'autre côté au prix correspondant, les deux ordres sont exécutés et le reste du plus grand des deux, s'il y en a un, est réinséré dans le CLOB (et peut éventuellement correspondre à un autre ordre).

L'exemple présenté sur la figure 3 illustre l'insertion d'un ordre de vente de treize participations au prix p. Avant l'insertion, la meilleure demande est au prix p et la meilleure offre est à $p + 1$. L'ordre inséré correspond en termes de prix à un ordre situé de l'autre côté. Le plus grand des deux est alors partiellement exécuté, le plus petit est exécuté dans sa totalité (c.-à-d. Bid8 est supprimé), et le reste du plus grand est placé dans le CLOB (c.-à-d. Ask5 cf. représentation à t+1). À l'issue de l'insertion,

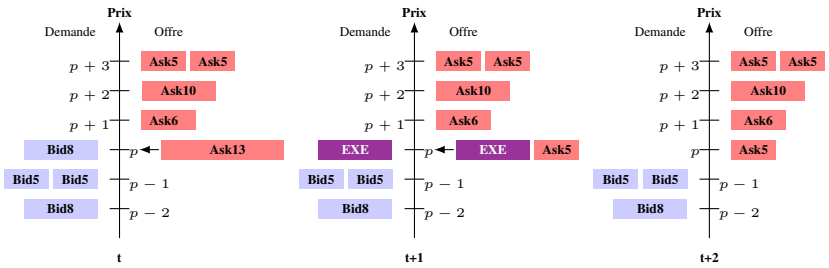

Figure 3. Exécution d'un ordre lors de son ajout sur le marché

la nouvelle meilleure demande est de prix $p - 1$ et la nouvelle meilleure offre est de prix p. Tous ces changements sont perçus par les agents par la mise à jour de croyances sur l'évolution du marché et l'état de leur portefeuille. Les croyances sur le marché au sein d'un agent sont de la forme suivante :

```
indicators(Date,Marketplace,Asset,Close,Volume,
    Intensity,Mm,Dblmm,BollingerUp,BollingerDown)

onMarket(Date,Agent,Portfolio,Marketplace,
    Side,Asset,Volume,Price)

executed(Date,Agent,Portfolio,Marketplace,
    Side,Asset,Volume,Price)
```

Les agents perçoivent chaque minute un ensemble de statistiques sur chaque actif coté via le prédicat `indicator`. Ce prédicat comporte dix arguments que sont le volume (la quantité de titres échangés), deux moyennes mobiles, calculées sur deux durées différentes, l'écart-type des prix sur les dernières vingt minutes, le prix pratiqué lors du dernier échange, et les indicateurs de Bollinger (BollingerUp (resp. BollingerDown), c'est-à-dire le prix moyen plus (resp. moins) le double de l'écart-type). Les agents sont également tenus informés des ordres en attente sur le marché et de leur exécution.

Les titres (`Asset`) présents dans le portefeuille de l'agent sont représentés par des croyances tenues à jour à chaque changement telles que :

```
own(PortfolioName,Broker,Asset,Quantity).
```

En raisonnant sur ces croyances présentes dans $\mathcal{B}_{a_i}$, un agent est alors capable d'évaluer la possibilité de passer un ordre d'achat ou de vente (en vérifiant s'il pos-

sède ou non assez de devises ou de titres) afin de générer l'ensemble des actions possibles $\mathcal{A}_{c_{a_i}}$. L'objet de cette preuve de concept n'étant pas de montrer une quelconque performance dans la sélection des investissements, la fonction d'évaluation de la désirabilité génère l'ensemble $\mathcal{A}_{d_{a_i}}$ à l'aide d'une méthode simple et classique basée sur la comparaison des moyennes mobiles et des bandes de Bollinger.

Dans le cadre de cette preuve de concept, nous considérons trois types d'agent pouvant évoluer sur la place de marché :

– *Agent aléatoire* envoyant au marché des ordres dont les prix et les volumes sont choisis au hasard, dans le but de générer une activité et simuler le « bruit » des places de marché réelles. Chaque agent aléatoire est en charge de la gestion d'un titre donné.

– *Agent sans éthique* doté uniquement d'une fonction d'évaluation de la désirabilité et n'ayant pour but que de spéculer : si le prix du marché monte (La moyenne mobile sur la durée la plus courte est au dessus de celle calculée sur la durée plus longue) il se comporte en acheteur, sinon, il se comporte en vendeur. Si le cours du titre est en dehors des indicateurs de Bollinger, ces règles sont inversées (ce cas est pris pour un signe de retournement de tendance). La même stratégie est employée par les agents éthiques pour évaluer l'ensemble $\mathcal{A}_{d_{a_i}}$ de leurs actions désirables.

– *Agent éthique* implémentant le modèle de jugement éthique sur ses propres actions afin de prendre une décision sur la sélection ou non de cette action.

Précisons qu'un agent n'ayant aucune règle morale et pour seul principe la satisfaction de ses désirs serait un agent hédonique, ce qui est différent des agents sans éthique au sens où ils raisonnent sur l'éthique de leurs actions en les considérant explicitement comme amorales.

Dans la section suivante, nous concentrons nos propos sur l'implémentation des agents éthiques et la description de leurs théories du bien et du juste en suivant le modèle de modèle de jugement décrit dans la section précédente.

5.2. Paramétrage éthique

Nous considérons des agents disposant, dès l'origine de la simulation, de croyances particulières sur les sociétés cotées sur le marché. Ces croyances peuvent concerner par exemple le secteur d'activité (e.g. production d'énergie nucléaire, transport), les labels montrant une conformité à certaines exigences ou l'appartenance à des groupes de pression. Ces croyances permettent d'exprimer la moralité de l'échange d'un actif en fonction de considérations sur la société à laquelle il est relié.

Pour décrire la moralité d'un acte sur les marchés financiers, nous fournissons aux agents une implémentation de valeurs morales et de règles morales directement inspirées de documents mis à disposition en ligne par divers organismes [9]. Chaque agent éthique est alors doté d'un ensemble de valeurs hiérachisées : par exemple « *environ-*

9. http://www.ethicalconsumer.org/

mental reporting » est décrit comme un sous-ensemble de la valeur « *environment* ». Les valeurs sont décrites sous la forme de prédicats logiques tels que :

```
value("environment").
subvalue("promote_renewable_energy","environment").
subvalue("environmental_reporting","environment").
subvalue("fight_climate_change","environment").
```

Les agents disposent également d'un ensemble de supports de valeurs tels que « Échanger des actifs liés à la production d'énergie nucléaire est contraire à la valeur de promotion des énergies renouvelables », ce qui peut être décrit comme :

```
~valueSupport(buy(Asset,_,_,_),
              "promote_renewable_energy"):-
    activity(Asset,"nuclear_energy_production").
```

« Échanger des actifs d'une société labelisée FSC est en accord avec la valeur de conformité environnementale » décrit comme :

```
valueSupport(sell(Asset,_,_,_),
             "environmental_reporting") :-
    label(Asset,"FSC").
```

« Échanger des actifs liés à la production d'énergie nucléaire est conforme à la valeur de lutte contre les changements climatiques » décrit comme :

```
valueSupport(buy(Asset,_,_,_),
             "fight_climate_change") :-
    activity(Asset,"nuclear_energy_production").
```

Les agents sont également dotés de règles morales leur permettant de lier la promotion de valeurs morales à une valuation morale. Par exemple « Il est moral d'agir conformément à la valeur *environment* » est simplement représenté comme :

```
moral_eval(X,V1,moral):-
    valueSupport(X,V1) & subvalue(V1,"environment").

moral_eval(X,"environment",moral):-
    valueSupport(X,"environment").
```

À ce stade, un agent éthique est capable de déduire que, au regard de ses croyances et de sa théorie du bien, échanger des titres d'une société labelisée FSC est moral,

tandis qu'échanger des actifs d'un producteur d'énergie nucléaire est à la fois moral et immoral au regard d'une même valeur. L'agent a donc besoin d'une théorie du juste pour déterminer s'il est éthique ou non d'échanger le second titre.

Dans ce but, les agents éthiques sont dotés de principes éthiques tels que l'éthique d'Aristote (dont l'implémentation est une adaptation de celle proposée en Answer Set Programming par (Ganascia, 2007b)) et des principes plus simples tels que la description d'une action parfaite `perfectAct` « Il est juste de commettre des actions possibles désirables et morales » ou encore le non-renoncement à ses désirs `desireNR` « Il est juste de commettre des actions possibles, non immorales et désirables ». Ces principes et leur ordre de préférence sont décrits sous la forme suivante :

```
ethPrinciple("perfectAct",Action):-
    possible_eval(Action, possible) &
    desire_eval(Action,desired) &
    not desire_eval(Action,undesired) &
    moral_eval(Action,_,moral) &
    not moral_eval(Action,_,immoral).

prefEthics("aristotelian","perfectAct").
prefEthics("perfectAct","desireNR").
prefEthics("desireNR","dutyNR").
```

Une règle `tPrefEthics(PE2,PE1)` permet ensuite de définir une relation de transitivité sur les éléments de l'ensemble $\succ_{e_{a_i}}$ des préférences éthiques.

Chaque agent peut ainsi utiliser plusieurs principes éthiques afin de juger, dans une situation et au regard d'une théorie du bien, de l'action juste, c'est-à-dire celle qui satisfait au mieux les principes éthiques dans un ordre lexicographique. Le jugement est implémenté par les règles suivantes :

```
existBetter(PE1,X):-
    ethPrinciple(PE1,X) &
    tPrefEthics(PE2,PE1) &
    ethPrinciple(PE2,Y) &
    not ethPrinciple(PE1,Y).

ethicalJudgment(X,PE):-
    ethPrinciple(PE,X) &
    not existBetter(PE,X).
```

5.3. Résultats expérimentaux

Cette section présente et commente les résultat d'une série d'expérimentations simulant un marché sur lequel quatre actifs, dont les cours sont animés par cinquante

agents aléatoires, sont échangés par 10 agents sans éthique et 10 agents éthiques. Nous cherchons à illustrer l'utilisation du modèle de jugement éthique, tel qu'il est décrit en section 3, comme processus de décision.

À l'initialisation, chaque agent reçoit un portefeuille contenant un nombre aléatoire d'actifs pour une valeur totale d'environ 500 €.

Les figures 4 et 5 montrent le résultat de l'expérience. La figure 4 montre l'évolution des volumes et prix des transactions effectuées sur le cours de l'action LEGRAND par les agents. Les chandeliers montrent l'évolution du prix, et les deux courbes situées au milieu représentent les moyennes mobiles évoquées en section 5.1. On remarque qu'il est rare que les chandeliers passent en dehors des bandes de Bollinger représentées par les courbes encadrant les évolutions de prix. Le volume représenté en histogramme dans la partie inférieure montre une activité souvent intensifiée lorsque les moyennes mobiles s'entrecroisent. Ces importants volumes d'échanges sont explicables en raison de la fonction d'évaluation de la désirabilité qui considèrent l'intersection des moyennes comme le signe d'un renversement de tendance (haussière ou baissière), entraînant un changement de position des agents (acheteur ou vendeur).

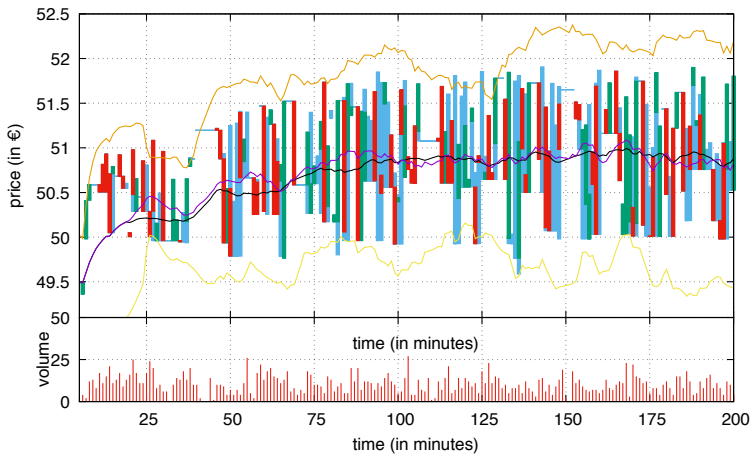

Figure 4. Evolution du titre LEGRAND. Les chandeliers représentent l'évolution du prix, avec les moyennes mobiles en son centre et les bandes de Bollinger de part et d'autre

La figure 5 représente l'évolution du portefeuille d'un agent éthique écologiste au cours de l'expérience. Son tracé, caractéristiques des agents dotés d'une éthique décrite dans la section précédente, indique par des couleurs l'évolution des proportions d'actions composant le portefeuille au cours de l'expérience. Le tableau 2 donne les résultats cumulés de dix simulations de trente minutes.

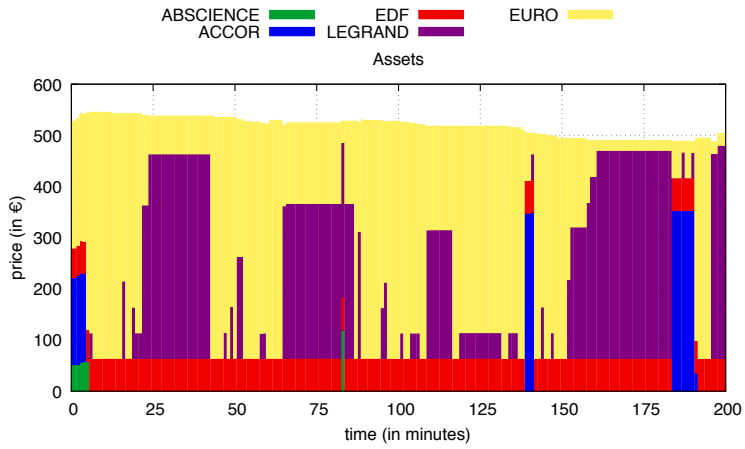

Figure 5. Évolution du portefeuille d'un agent éthique

Tableau 2. Nombre total de transactions après dix simulations de 30 minutes

	AB SCIENCE		ACCOR		EDF		LEGRAND	
	achats	ventes	achats	ventes	achats	ventes	achats	ventes
Total	6242	5928	7665	7349	0	0	12948	12284
	23 %		28 %		0 %		49 %	

Nous pouvons en premier lieu relever que le nombre de titres EDF dans le portefeuille est invariable au cours de la simulation. Ce phénomène s'explique par la morale de l'agent qui ne considère jamais comme juste d'échanger un tel actif en raison de supports de valeur concernant les producteurs d'énergie d'origine nucléaire.

Deuxièmement, nous pouvons aussi observer que, durant certaines périodes, le portefeuille intègre un grand nombre de titres de LEGRAND. De fait, l'agent ayant connaissance d'une labellisation de cette entreprise, une motivation morale s'ajoute à la recherche de bénéfice et en faire un investissement privilégié (près de la moitié des transactions selon le tableau 2). Ces actes d'achat et de vente seront en outre les seuls à satisfaire le principe `perfectAct` évoqué en section précédente.

Enfin, nous pouvons relever différentes périodes durant lesquelles l'agent effectue des investissements dans des titres n'ayant aucun lien avec les valeurs et règles morales de l'agent. Ces actifs sont choisis en raison du gain qu'ils apportent lorsqu'aucun investissement plus éthique n'est possible.

6. Conclusion

Afin d'agir collectivement en conformité avec une éthique et une morale donnée, un agent autonome a besoin d'être capable d'évaluer la justesse et la moralité de son comportement et de celui des autres. En se fondant 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'agents, 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 ne permet pas encore la représentation d'un degré d'éthique. En outre, bien que l'on puisse définir diverses valuations morales, nous n'avons pas expérimenté l'introduction d'incertitudes sur les conséquences. Il serait intéressant d'enrichir le modèle d'action et d'intégrer un raisonnement sur la causalité. Quelques travaux récents ont abordé la problématique de la prise en compte de la causalité dans le raisonnement éthique (Berreby *et al*., 2017).

De plus, la notion de principe éthique reste à approfondir pour refléter l'immense variété des théories philosophiques. Il serait intéressant d'adapter d'autres principes éthiques et d'illustrer les variations de jugement qu'ils produisent. De même, il serait possible d'envisager la représentation d'un degré d'éthique afin de fournir une information sur l'adéquation entre une action et une théorie du juste.

La suite de ce travail pourrait consister en la représentation de divers codes de déontologie dans d'autres domaines applicatifs (par exemple, en éthique médicale ou journalistique) afin de montrer la généricité 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, telle qu'une image construite par des jugements successifs, pour permettre à un agent de se positionner plus facilement face à leur diversité (Cointe *et al*., 2017).

Remerciements

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

Bibliographie

Aldridge I. (2009). *High-frequency trading: a practical guide to algorithmic strategies and trading systems* (vol. 459). John Wiley and Sons.

Alexander L., Moore M. (2015). Deontological Ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2015 éd.. http://plato.stanford.edu/archives/spr2015/entries/ethics-deontological/.

Anderson M., Anderson S. (2014). Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot*, vol. 42, n° 4, p. 324-331.

Arkin R. (2009). *Governing lethal behavior in autonomous robots*. CRC Press.

Arkoudas K., Bringsjord S., Bello P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI fall symposium on machine ethics*, p. 17–23.

Battaglino C., Damiano R., Lesmo L. (2013). Emotional range in value-sensitive deliberation. In *12th international conference on autonomous agents and multi-agent systems*, p. 769–776.

Berreby F., Bourgne G., Ganascia J.-G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *20th international conference on logic for programming, artificial intelligence, and reasoning*, p. 532-548.

Berreby F., Bourgne G., Ganascia J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, p. 96–104.

Boella G., Pigozzi G., Torre L. van der. (2009). Normative systems in computer science - Ten guidelines for normative multiagent systems. In *Normative multi-agent systems.*

Boissier O., Bordini R. H., Hübner J. F., Ricci A., Santi A. (2013). Multi-agent oriented programming with JaCaMo. *Science of Computer Programming*, vol. 78, n° 6, p. 747–761.

Bono S., Bresin G., Pezzolato F., Ramelli S., Benseddik F. (2013). *Green, social and ethical funds in Europe*. Rapport technique. Vigeo.

Bratman M. (1987). Intention, plans, and practical reason.

Coelho H., Rocha Costa A. da. (2009, October). On the intelligence of moral agency. *Encontro Português de Inteligência Artificial*, p. 12–15.

Coelho H., Trigo P., Rocha Costa A. da. (2010). On the operationality of moral-sense decision making. In *2nd brazilian workshop on social simulation*, p. 15–20.

Cointe N., Bonnet G., Boissier O. (2016a). Ethical judgment of agents' behaviors in multi-agent systems. In *15th international conference on autonomous agents & multiagent systems*, p. 1106-1114.

Cointe N., Bonnet G., Boissier O. (2016b). Jugement éthique dans les systèmes multi-agents. In *Journées francophones sur les systèmes multi-agents.*

Cointe N., Bonnet G., Boissier O. (2017). Coopération fondée sur l'éthique entre agents autonomes. In *Journées francophones sur les systèmes multi-agents.*

Damasio A. (2008). *Descartes' error: Emotion, reason and the human brain*. Random House.

Dennis L., Fisher M., Winfield A. (2015). Towards verifiably ethical robot behaviour. In *1st international workshop on AI and ethics.*

Ethical judgment. (2015, August). Free Online Psychology Dictionary.

Foot P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, p. 5–15.

Friedman B. (1996). Value-sensitive design. *Interactions*, vol. 3, n° 6, p. 16-23.

Friedman B., Kahn P., Borning A., Huldtgren A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, p. 55-95. Springer Netherlands.

Ganascia J.-G. (2007a). Ethical system formalization using non-monotonic logics. In *29th annual conference of the cognitive science society*, p. 1013–1018.

Ganascia J.-G. (2007b). Modelling ethical rules of lying with Answer Set Programming. *Ethics and information technology*, vol. 9, n° 1, p. 39–47.

Gert B. (2015). The definition of morality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall éd..

Greene J., Haidt J. (2002). How (and where) does moral judgment work? *Trends in cognitive sciences*, vol. 6, n° 12, p. 517–523.

Hursthouse R. (2013). Virtue ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall éd..

Johnson R. (2014). Kant's moral philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Summer éd..

Kant E. (1967). Sur un prétendu droit de mentir par humanité (1797). *Théorie et pratique/Droit de mentir*, p. 67–73.

Kim K.-J., Lipson H. (2009). Towards a theory of mind in simulated robots. In *11th annual conference companion on genetic and evolutionary computation conference*, p. 2071–2076.

Koopman G., Székely I. (2009). Impact of the current economic and financial crisis on potential output. *European Economy, Occasional Paper*, n° 49.

Mao W., Gratch J. (2013). Modeling social causality and responsibility judgment in multi-agent interactions. In *23rd international joint conference on artificial intelligence*, p. 3166–3170.

McConnell T. (2014). Moral dilemmas. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall éd..

McIntyre A. (2014). Doctrine of double effect. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter éd..

Professional Journalists S. of. (2014, September). *Code of ethics.*

Ricoeur P. (1995). *Oneself as another*. University of Chicago Press.

Rocha-Costa A. (2016). Moral systems of agent societies: Some elements for their analysis and design. In *1st workshop on ethics in the design of intelligent agents*, p. 32-37.

Rokeach M. (1973). *The nature of human values*. New York Free Press.

Saptawijaya A., L. Moniz Pereira L. M. (2014). Towards modeling morality computationally with logic programming. In *Practical aspects of declarative languages*, p. 104–119.

Scheff T. J. (2003). Shame in self and society. *Symbolic interaction*, vol. 26, n° 2, p. 239–262.

Schroeder M. (2016). Value theory. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Fall 2016 éd.. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2016/entries/value-theory/.

Schwartz S. (2012). An overview of Schwartz theory of basic values. *Online Readings of Psychology and Culture*, vol. 2, n° 1.

Schwartz S. H. (2006). Basic human values: Theory, measurement, and applications. *Revue française de sociologie*, vol. 47, n° 4, p. 249–288.

Schwartz S. H., Bilsky W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross cultural replications. *Journal of Personality and Social Psychology*, vol. 58, p. 878-891.

Stoeber J., Yang H. (2016). Moral perfectionism and moral values, virtues, and judgments: Further investigations. *Personality and Individual Differences*, vol. 88, p. 6–11.

Timmons M. (2012). *Moral theory: an introduction*. Rowman & Littlefield Publishers.

van Marrewijk M., Werre M. (2003). Multiple levels of corporate sustainability. *Journal of Business Ethics*, vol. 4, n° 2-3, p. 107-119.

Walter S. (2015). Consequentialism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter éd..

Wiegel V., Berg J. van den. (2009). Combining moral theory, modal logic and MAS to create well-behaving artificial agents. *International Journal of Social Robotics*, vol. 1, n° 3, p. 233–242.

Wiener Y. (1988). Forms of value systems: A focus on organisational effectiveness and cultural change and maintenance. *Academy of Management Review*, vol. 13, n° 4, p. 534-545.