Éthique collective dans les systèmes multi-agents

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

1. *Université de Lyon, MINES Saint-Etienne,*
CNRS, Laboratoire Hubert Curien, UMR 551
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'utilisation croissante des technologies multi-agents pour le développement de systèmes socio-techniques révèle de nombreux verrous, dont ceux liés à l'éthique des décisions autonomes que de tels systèmes peuvent être amenés à prendre. Ces verrous sont souvent considérés comme un problème de conception selon une perspective centrée sur l'agent. Ici, nous élargissons cette perspective au collectif, introduisant ainsi les questions éthiques découlant de la participation des agents autonomes à des interactions entre agents, à des coalitions ou à des structures plus pérennes telles que des organisations. Cet article propose un panorama de ces différentes questions et des travaux actuels ou récents s'y intéressant.*

ABSTRACT. *The increasingly current and future presence of multi-agent systems in various areas leads us to ask many questions about the ethics of their decisions. Indeed, decisions that these systems must make sometimes require consideration of ethical concepts, both as an individual entity and a member of an organization. This problem, often considered as a design issue from an agent centered perspective, is addressed in this paper from a collective point of view and the result of reasonning. This position paper discusses the concepts to be modeled within an agent, a set of issues and a state of the art for addressing this question.*

MOTS-CLÉS : *systèmes multi-agents, éthique collective, dilemmes.*

KEYWORDS: *multi-agents systems, collective ethics, dilemmas.*

DOI:10.3166/RIA.31.71-96

1. Introduction

L'introduction d'agents autonomes artificiels dans des domaines tels que le milieu hospitalier, la gestion d'actifs financiers ou encore les transports soulève de nombreux problèmes tels que la prise en compte ou le raisonnement sur des règles morales. En effet, les utilisateurs de ces systèmes ont parfois des attentes éthiques distinctes des problématiques d'optimalité, ou de conformité légale du comportement des agents. Par exemple, des agents capables de comprendre et utiliser le code de déontologie médicale pourraient s'appuyer sur des motivations éthiques afin de choisir quelles informations diffuser, à qui et sous quelles conditions, conformément au principe du secret médical. L'intérêt pour le comportement éthique des agents autonomes est récemment apparu dans les travaux de recherche traitant des agents autonomes (CERNA, 2014 ; Russell *et al.*, 2015), comme en témoignent les nombreux articles (Beavers, 2011 ; McDermott, 2008 ; McLaren, 2006 ; Moor, 2006) et conférences [1]. Cependant, ces travaux s'intéressent uniquement à l'éthique à l'échelle du comportement individuel de l'agent, c'est-à-dire dans l'optique de créer des agents autonomes dont le comportement est conforme à des normes d'éthique.

Or, dans un système multi-agent (ou SMA), une simple contrainte de son comportement permet à un agent d'agir individuellement de manière éthique dans un collectif, mais le laisse démuni lorsqu'il doit tenir compte de l'éthique des autres agents. Par exemple, un agent gestionnaire d'investissements financiers pourrait être tout à fait capable de se comporter selon des principes de gestion responsable sans pour autant être capable d'évaluer le caractère éthique du comportement des autres. Prendre en considération la dimension multi-agent de ce problème nécessite l'exploration de nouvelles pistes telles que la création d'une ou plusieurs éthiques collectives ou la prise de décisions en coopération face à des problèmes d'éthique. Ces questions ont d'autant plus d'importance dans le contexte actuel de déploiement d'un nombre croissant d'agents dans notre environnement, collaborant entre eux ou avec des humains. Cet article de positionnement a pour but de proposer des définitions et des questions mettant en évidence la problématique des éthiques collectives dans les systèmes multi-agents et de définir les concepts d'éthique que les agents doivent pouvoir utiliser dans leur raisonnement. La définition d'un cadre d'analyse permet ici de préciser les éléments qu'un agent ou un collectif d'agents doit employer afin de discerner la décision la plus conforme à une morale et une éthique donnée

1. Symposium on Roboethics, 2004, www.roboethics.org/sanremo2004/ – International Conference on Computer Ethics and Philosophical Enquiry, 2011 – Workshop on AI and Ethics, 2015, www.aaai.org/Library/Workshops/ws15-02.php – International Conference on AI and Ethics, 2015, wordpress.csc.liv.ac.uk/va/2015/02/16/1st-international-conference-on-ai-and-ethics/ – Workshop on Ethics in the Design of Intelligent Agents, 2016, ii.tudelft.nl/edia2016 – sujet spécial sur Artificial Intelligence for Human Values à ECAI, 2016, www.ecai2016.org/ – etc.

face à un choix. L'objectif est d'explorer cette problématique en préambule à des travaux ultérieurs de propositions plus formelles.

Cet article est structuré comme suit. La section 2 présente les concepts philosophiques et techniques employés dans cet article, puis en section 3 nous proposons un cadre d'analyse pour la représentation de l'éthique au sein d'un agent, puis d'un collectif d'agents autonomes. Nous utilisons ensuite ce cadre pour présenter les problématiques spécifiques touchant aux questions d'éthiques collectives dans des systèmes multi-agents en sections 4 et 5.

2. Cadre philosophique et technologique

Afin de préciser les concepts employés dans cet article, nous donnons quelques définitions de notions d'éthique en philosophies et sciences humaines. Il ne s'agit nullement d'en donner une vision exhaustive. Nous invitons le lecteur intéressé à se reporter à la bibliographie pour approfondir ces concepts. La deuxième partie de cette section donnera également quelques définitions issues du domaine SMA et une description des approches existantes en matière de représentation de l'éthique dans ces systèmes.

2.1. Éthique dans les systèmes humains

Nous abordons ici les travaux de recherche concernant l'éthique en philosophie, puis en sciences humaines et sociales. Le but est de montrer comment les divers concepts de morale et d'éthique sont employés pour décrire divers connaissances et raisonnements humains, puis comment la constitution de groupes d'individus dotés de telles éthiques soulève de nouvelles problématiques.

2.1.1. Éthique en philosophie

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 de *morale*, *d'éthique*, *de jugement* et *de valeurs*, qui sont les fondements de notre étude, nous tirons les définitions suivantes :

La morale *désigne l'ensemble de règles déterminant la conformité des pensées ou actions d'un individu vis-à-vis des 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.*

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 (Gert, 2015). Ainsi par exemple, 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 notre raisonnement permettant de distinguer les bonnes et mauvaises actions. 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 (Schwartz, 2006). Un ensemble de règles morales ou de valeurs morales établit une *théorie du bien* ou – parfois – *théorie des valeurs* (Timmons, 2012).

Un ensemble de principes forme la *théorie du juste* ou *théorie de la juste conduite* qui définit des critères pour reconnaitre le choix le plus juste ou le plus acceptable (Timmons, 2012). 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 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 des philosophes tels que Paul Ricoeur (1995), nous admettons la définition suivante :

L'éthique *est la combinaison de* principes éthiques *et de* règles morales *permettant à un processus de décider d'une action conciliant au mieux la morale et les désirs de l'agent, étant données ses capacités.*

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 Saint 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'un comportement à des valeurs telles que la sagesse, le courage ou la justice (Hursthouse, 2013). De nombreux travaux d'axiologie (étude des valeurs) cherchent à déterminer quelles sont les valeurs connues et reconnues par les humains (Schwartz, 1992). Proposée dès l'antiquité grecque par l'École d'Athènes, l'éthique des vertus cherche à déterminer quels sont les vertus et les vices qui doivent guider le comportement de l'homme. Pour distinguer une bonne action d'une mauvaise, il faut alors disposer d'une définition de ces valeurs et chercher en quoi elles supportent ou rejettent l'action à évaluer. Socrate fait remarquer dans le fameux dilemme des "armes du fou" (Platon, 1966) que des contradictions peuvent apparaître facilement lorsqu'un choix nous confronte à deux actions opposées supportées

par deux vertus. Dans cette situation fictive, Socrate demande s'il faut rendre à un ami devenu fou des armes que nous lui avons empruntées, au risque qu'il s'en serve pour causer du mal à lui-même ou à autrui. Il n'est pas possible ici de faire preuve à la fois de loyauté et de bienveillance envers cet ami.

▸ l'*éthique déontologique* juge un comportement par sa conformité avec des obligations et permissions associées à des situations (Alexander, Moore, 2015). La définition de ces permissions et obligations sans recours à la notion de valeur permet de contourner le problème de l'interprétation subjective de la définition des valeurs. Ainsi, l'éthique déontologique est souvent employée pour décrire l'éthique d'une communauté religieuse ou professionnelle de manière la moins ambiguë possible. Elle formule principalement des obligations et interdictions morales avec une éthique fondée sur une obéissance aux règles morales. Le commandement "Tu ne tueras point" est un exemple typique ne faisant référence à aucune vertu ou vice et qualifiant directement une action de bonne ou mauvaise en soi, sans mention de ses conséquences.

▸ l'*éthique conséquentialiste*, aussi appelée téléologie, juge un comportement à la moralité de ses conséquences (Walter, 2015). Une telle approche permet de justifier une action par la moralité du but recherché. Certains conséquentialisme portent plus sur la définition des conséquences. Par exemple, l'hédonisme cherche à maximiser le plaisir en minimisant la souffrance. D'autres s'intéressent à la pondération des bonnes et mauvaises conséquences. Par exemple,

- l'égoïsme cherche en priorité à maximiser le bien de l'agent prenant la décision, reléguant au second plan la considération des effets impactant les autres.

- l'utilitarisme vise à optimiser le bien pour l'ensemble des agents connus. Il maximise le bien pour la société dans son ensemble.

- l'altruisme tend à prendre en compte en priorité les conséquences affectant les autres agents.

Cependant, dans certaines situations, un ensemble de principes peut donner un jugement identique entre deux actions au regard d'une morale donnée. Ces situations sont appelées *dilemmes*.

Un dilemme *est un 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 qu'un dilemme n'existe que pour un ensemble de morales et d'éthiques dans une situation donnée. Par exemple, s'il est considéré comme immoral de mentir et de mettre en danger quelqu'un, l'usage de l'impératif catégorique de Kant (Johnson, 2014)

placerait l'agent face à un dilemme s'il doit choisir entre un mensonge pour protéger quelqu'un ou dire la vérité en le mettant en danger. En revanche, la doctrine du double effet de Saint Thomas d'Aquin (McIntyre, 2014) considère que le choix juste est le mensonge. 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 :

Le jugement *est la faculté d'évaluer 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.*

Le jugement est ainsi utilisable à la fois pour évaluer l'option la plus éthique face à nos propres décisions, et en même temps pour évaluer le comportement d'un autre.

2.1.2. Éthique en sciences humaines

L'éthique des individus diffère d'une personne à l'autre pour des raisons culturelles (les valeurs de l'individu proviennent de son entourage familial, social, éducatif, religieux) et de développement cognitif (capacité à manipuler mentalement des concepts plus ou moins complexes).

Pionnier dans ce domaine, Milton Rokeach (1974) propose une méthodologie d'étude des valeurs d'une société afin d'observer leur évolution. Ses travaux montrent comment divers critères viennent influencer la place accordée à ces valeurs dans la morale de chacun. Plus généralement, Shalom Schwartz (Schwartz, 1992 ; Schwartz *et al.*, 2007) a identifié dix catégories de valeurs principales qui regroupent la totalité des autres valeurs (*Self-Direction*, *Stimulation*, *Hedonism*, *Achievement*, *Power*, *Security*, *Conformity*, *Tradition*, *Benevolence* et *Universalism*). Les valeurs sont hiérarchisées et peuvent s'opposer : par exemple pour S. Schwartz la conformité et la tradition s'opposent à l'autonomie (*Self-Direction*).

Outre ces variations dans l'éthique individuelle provoquées par des différences de valeurs entre individus, Lawrence Kohlberg (Kohlberg, Hersh, 1977) montre que différents niveaux de développement moral coexistent dans la société. Ainsi, en plus des différences de valeurs, le jugement chez l'homme repose sur des processus relevant de degrés de maturité différents :

▸ *niveau préconventionnel* constitué d'un processus de punition et récompense (stade 1 de l'évolution) et d'un processus de satisfaction de ses intérets (stade 2 de l'évolution),

▸ *niveau conventionnel* basé sur un accord interpersonnel et conformiste (stade 3 de l'évolution), puis un respect de la loi et une conformité aux normes sociales (stade 4 de l'évolution),

▸ *niveau postconventionnel* passant par un contrat social (stade 5 de l'évolution), et enfin un raisonnement portant sur des principes éthiques universels (stade 6 de l'évolution).

Dans ce modèle, le but de l'éducation est de faire progresser l'individu vers le sixième stade où l'éthique est le produit d'un raisonnement sur des principes complexes et des valeurs morales. Ces travaux montrent également que moins d'un adulte sur cinq parvient à ce dernier stade et que la majorité s'arrête au quatrième stade où l'obéissance à des règles et le respect de l'ordre social priment.

Au-delà de ces différences et multiplicités d'éthique, les sciences de gestion introduisent un niveau supra-individuel qui est celui de l'organisation. L'éthique est envisagée comme un élément important, sinon incontournable, de la définition d'une organisation (Treviño *et al.*, 2006). La diversité des éthiques y est vue comme problématique au sens où les auteurs font remarquer que les différentes approches (seules les approches déontologiques et conséquentialistes sont abordées dans cet ouvrage) jugent l'éthique du comportement sur des éléments différents : l'action en soi ou ses conséquences (voir 2.1.1). Les auteurs s'intéressent au problème de la prise de conscience de problématiques éthiques (*awareness of ethical issues*) et montrent que les individus au sein d'une organisation n'ont pas la même perception des problématiques éthiques. Cette différence de perception est corrélée au degré de maturité de l'individu au sens de (Kohlberg, Hersh, 1977) et peut donc évoluer au cours du temps. De plus, les valeurs et principes d'un individu semblent constituer une part de son identité (Bergman, 2004). La conduite éthique résulte ainsi d'une tendance à éviter la culpabilité (ne pas violer sa propre éthique), et d'une tendance à éviter la honte (par la violation d'une éthique du collectif). L'individu est donc dans une situation plus confortable lorsque son éthique personnelle est en accord avec l'éthique d'un collectif. L'éthique d'un collectif peut ainsi être également perçue comme une part de l'identité d'une organisation.

2.2. Éthique des systèmes artificiels

Après avoir défini l'éthique et décrit les éléments qui la composent et l'entourent, nous abordons ici le contexte des systèmes multi-agents dans lequel nous souhaitons l'introduire.

2.2.1. Systèmes multi-agents

Dans cet article, nous considérons les modèles et technologies issues des SMA. Dans ce domaine, un *agent autonome* est un système informatique ou robotique situé dans un environnement, espace partagé avec d'autres agents avec lesquels il interagit. Nous nous intéressons plus particulièrement ici à des agents cognitifs, i.e. des agents dotés d'une représentation symbolique permettant de manipuler explicitement les concepts d'éthique que nous avons décrits dans la section précédente. Cette représentation explicite permet de mettre en place des raisonnements et des explications sur ces concepts en tenant compte de leur sémantique. Par le terme de *système multi-agent*, nous entendons un ensemble d'agents autonomes plongés dans un environnement dans lequel ils évoluent de manière indépendante, peuvent communiquer entre eux et poursuivre leurs propres objectifs (Ferber, 1995). Ces agents sont souvent intégrés à des organisations et peuvent adopter des rôles décrivant leur fonction au sein du système. La conception d'un SMA touche ainsi à des problématiques d'intelligence artificielle, à des questions de relations et d'organisation sociale, et à la prise en compte de cet aspect dans les processus de décision.

Ainsi, pour concevoir des agents compréhensibles et configurables par des utilisateurs ayant divers niveaux de compréhension, il faut donc pouvoir intégrer ces divers modes de raisonnement dans l'architecture des agents, et leur permettre de raisonner à des stades différents. L'utilisateur peut jouer des rôles divers selon les contextes : fournir sa propre éthique à l'agent, collaborer avec lui, formuler des critiques ou en recevoir, etc. Notons qu'il convient alors de se demander si un agent artificiel aurait à considérer des valeurs non pas humaines mais spécifiques à sa nature artificielle d'outil. Par exemple, dans le contexte d'agents rationnels limités, le psychologue Gerd Gigerenzer (2010) considère que quatre valeurs suffisent aux agents artificiels : l'*imitation des pairs* signifiant faire ce que la majorité des autres agents font, l'*égalitarisme* signifiant distribuer les ressources en les divisant également, le *donnant-donnant*[2] signifiant toujours coopérer en premier lieu puis faire ensuite tout ce que l'adversaire à fait au coup précédent et l'*obéïssance* signifiant faire ce que la loi exige. Kari Gwen Coleman (2001), cherchant à l'inverse à définir les vertus de manière plus précise et variée, propose alors un concept de vertus artificielles, correspondant à des propriétés que l'agent peut ou non vérifier, structurées en quatre domaines : les vertus purement agentives (par exemple l'adaptativité, l'autonomie et l'autopoïèse), les vertus sociales (comme la disposition à dire la vérité), des vertus environnementales (comme l'obéissance aux contraintes de l'environnement) et des vertus morales.

Plus généralement, dans la suite, nous estimons qu'il doit pouvoir revenir à l'utilisateur de choisir les valeurs sur lesquels repose la morale d'un agent qui lui sert d'outil. Nous nous attachons donc davantage à la définition générique

2. Identique au *tit-for-tat* de Robert Axelrod (Axelrod, 2006).

de la notion de valeur (humaine ou artificielle) qu'à la manière la plus juste de définir chacune d'entre elle. Nous admettons que la problématique de la sélection et la description de ces valeurs dans le paramétrage d'un agent n'est toutefois pas triviale.

2.2.2. Mise en œuvre de l'éthique au sein des systèmes artificiels

Afin de représenter les notions d'éthique évoquées précédemment, de nombreuses approches ont été définies pour concevoir des agents autonomes dotés d'une éthique. Nous les regroupons les divers travaux de l'état de l'art au sein d'ensembles d'approches que sont l'*éthique par conception*, la *théorie des jeux*, la *conception sensible aux valeurs*, l'*éthique par étude de cas*, l'*éthique par raisonnement logique* et l'*architecture cognitive éthique*. Nous les avons ici ordonnées de l'approche la plus rigide (l'éthique est figée à la conception et n'est pas accessible à l'utilisateur) à la plus flexible (l'éthique est une connaissance explicite de l'agent et peut être exprimée). Ces approches peuvent être structurées en deux catégories : les représentations implicites et les représentations explicites. Par *implicite*, nous entendons impossible à partager car implémentée directement dans le fonctionnement de l'agent, à l'inverse d'une représentation *explicite* représentée par un langage logique sur laquelle il peut raisonner et qu'il peut partager ou comparer avec d'autres.

Représentations implicites

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 (Arkin, 2009)). Son inconvénient principal est l'absence de représentation générique de concepts éthiques (théories du bien et du juste) qui n'est qu'un élément du cahier des charges. 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.

Dans le domaine de la *théorie des jeux*, le problème de la satisfaction égoïste ou utilitariste des agents est présenté comme un critère à optimiser. Cependant, il est parfois proposé des axiomatisations de solutions satisfaisant des notions de justice, équité, honnêteté, etc. Comme dans l'éthique par conception, le caractère éthique est compris implicitement dans la solution apportée au problème de décision. Il est cependant intéressant de remarquer que la dimension collective est souvent prise en compte dans ce domaine. Par exemple, pour l'équité, la valeur de Shapley (1953) ou l'indice de Banzhaf (1964) permet d'exprimer

une distribution juste de gains entre les agents en fonction de leur contribution. Pour la justice (Pitt *et al.*, 2015), la valeur de solidarité (Nowak, Radzik, 1994) et la valeur de rationnement (Yang, 1997) permettent respectivement d'exprimer que les agents acceptent de partager leurs gains avec ceux qui en ont le moins ou de réduire leur gain pour laisser des ressources dans l'environnement. L'honnêteté est parfois axiomatisée par l'absence d'envie (Brams, Taylor, 1994) (aucun agent ne désire la part d'un des autres) ou des incitations à dire la vérité (parfois en tenant compte de notion de respect de la vie privée comme dans (Chen *et al.*, 2016)).

La *conception sensible aux valeurs* (*Value Sensitive Design*) est une approche méthodologique générique organisée en trois étapes (Friedman *et al.*, 2002) permettant de prendre en compte des valeurs éthiques dans des projets de conception logicielle. Contrairement à l'éthique par conception évoquée précédemment, elle conserve un lien explicite entre la technologie développée et les valeurs promues (Aldewereld *et al.*, 2015). Cette méthodologie repose sur une première étape d'*investigation conceptuelle* visant à spécifier les valeurs morales en jeu dans le projet, puis une *investigation empirique* invite les concepteurs à interroger les utilisateurs pour cerner leurs attentes et préoccupations avant de conclure par l'étape d'*investigation technique* durant laquelle les concepteurs effectuent des choix technologiques en accord avec les résultats des étapes précédentes. Cette approche permet de systématiser une démarche de réflexion qui précède l'implémentation, mais le logiciel conçu n'a pas la possibilité de raisonner de manière autonome sur les valeurs qui ont guidé sa conception, bien qu'elles puissent être présentées explicitement à l'utilisateur. L'éthique de l'agent est celle du concepteur, avec une explication de son expression à travers sa conception et son implémentation.

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 (Anderson, Anderson, 2015). 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.

Représentations explicites

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 ; Saptawijaya, Pereira, 2014 ; Berreby *et al.*, 2015 ; Bringsjord *et*

al., 2016). 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 au regard d'une théorie du bien 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 (Arkoudas *et al.*, 2005 ; Coelho *et al.*, 2010). (Costa, 2016) propose la construction de connaissances sur la morale d'un autre agent par construction à partir de l'observation de son comportement. Dans (Cointe *et al.*, 2016a ; 2016b), nous avons présenté un modèle de jugement éthique pour permettre aux agent de raisonner sur des valeurs et règles morales en employant des principes éthiques. 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 été représentée ou n'a pas encore été mise en place à l'échelle collective.

2.2.3. Synthèse

Les approches que nous venons de décrire 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. Ces travaux 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 (Kim, Lipson, 2009). Afin d'exprimer et concilier un maximum de théories du bien et du juste, il semble nécessaire de définir formellement les concepts employés et montrer comment les diverses approches philosophiques peuvent être exprimées. 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. Différents types de jugements basés sur la possibilité de substituer la morale d'un agent par celle d'un autre ont été proposés dans (Cointe *et al.*, 2016a). Ainsi, nous proposons d'utiliser le jugement à la fois comme un processus de décision dans un problème de choix social (Mao, Gratch, 2013), et comme une capacité à évaluer le caractère éthique du comportement des autres.

Remarquons que nous excluons la représentation des émotions au sein du raisonnement de l'agent pour fonder son jugement sur un processus 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 ; Damasio, 2008), 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.).

3. Cadre d'analyse

Afin d'analyser comment les concepts éthiques présentés en section 2.1.1 peuvent être représentés au sein des SMA, nous considérons un modèle simplifié d'un processus de jugement éthique proposé dans (Cointe *et al.*, 2016a). Ce modèle nous permettra de proposer une définition d'éthique individuelle et également d'éthique collective et d'évoquer une série de problématiques liées à cette notion dans un cadre multi-agent en sections 4 et 5.

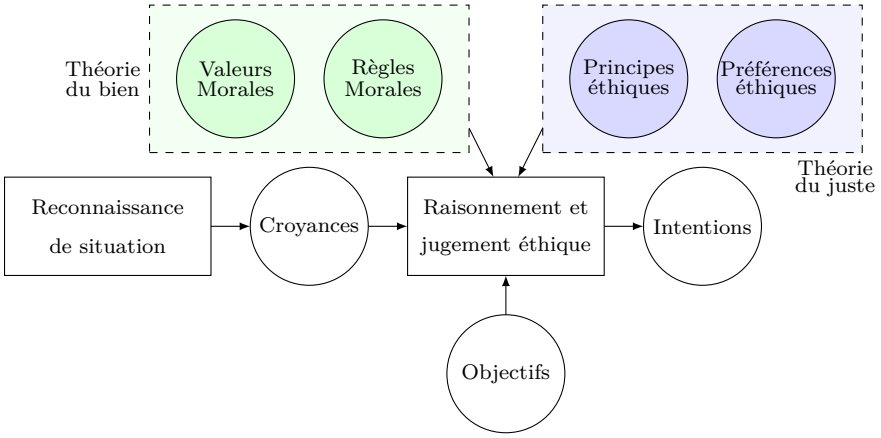

Figure 1. Modèle d'agent intégrant une éthique individuelle

3.1. Éthique individuelle

Nous considérons un modèle *Belief-Desire-Intentions* (BDI) (Rao, Georgeff, 1995) couramment utilisé pour modéliser des agents dotés de représentations symboliques dans le domaine des SMA. Il consiste en un modèle de raisonnement comprenant la représentation des croyances sur l'état du monde (*beliefs*) et des objectifs de l'agent (*desires*) afin d'en déduire l'action possible (*intention*) qui, dans un état de croyances représentant la situation courante, devrait permettre au mieux de satisfaire les objectifs de l'agent. Nous définissons alors l'éthique individuelle d'un agent autonome comme suit :

L'éthique individuelle au sein d'un système multi-agent : *une éthique individuelle au sein d'un système multi-agent désigne l'éthique d'un agent autonome. Cette éthique individuelle est définie par l'ensemble des éléments de la théorie du bien et de la théorie du juste d'un agent lui permettant de distinguer le caractère éthique d'une action dans une situation donnée.*

Le modèle d'agent intégrant une éthique individuelle présenté en figure 1 enrichit ainsi le modèle BDI par une théorie du bien (constituée de *valeurs morales* et de *règles morales*) et d'une théorie du juste (constituée de *principes éthiques* ordonnés par un ensemble de *préférences*) les intervenant dans le choix de l'action à effectuer. La théorie du bien associe un degré de bien ou de mal à des combinaisons d'actions, de croyances ou de désirs. Les principes éthiques, comme les règles d'Aristote ou l'Impératif Catégorique de Kant (Ganascia, 2007b ; 2007a), se présentent comme des règles décrivant un moyen de concilier les désirs et la morale. Le raisonnement enrichi d'un processus de jugement éthique exploite alors ces principes éthiques pour identifier l'intention à maintenir.

Le choix des principes éthiques permet de représenter une théorie philosophique indiquant comment raisonner sur la morale, les désirs et capacités de l'agent. Ainsi, par exemple, une éthique vertueuse propose de se conformer à des règles morales portant sur des valeurs, tandis qu'une éthique déontologique se préoccupe de l'adéquation entre l'état supposé par les croyances de l'agent et les actions obligatoires ou interdites dans ces circonstances. Une éthique conséquentialiste, enfin, chercherait à optimiser ses chances d'atteindre des états définis comme moralement souhaitables.

Exemple

Afin d'illustrer notre propos, considérons un exemple impliquant des agents propriétaires de sommes d'argent. Cet exemple est proposé dans un formalisme volontairement simplifié dont nous sommes conscients des limites.

Chaque agent peut être dans l'un des trois états PAUVRE, RICHE ou (exclusif) NEUTRE et est doté du modèle d'action suivant :

- VOLER(A) pour prendre à un agent A une part de ses richesses ;

- DONNER(A) pour donner de l'argent à l'agent A ;
- TAXER(A) pour réclamer à l'agent A une part de ses richesses ;
- COURTISER(A) pour tenter de s'attirer les faveurs de l'agent A.

Supposons un agent Robin des Bois doté des règles morales suivantes :

$M1$. PAUVRE(A) → IMMORAL(TAXER(A)) ;

$M2$. PAUVRE(A) → IMMORAL(VOLER(A)) ;

$M3$. PAUVRE(A) → MORAL(DONNER(A)) ;

$M4$. ¬PAUVRE(A) → ¬MORAL(DONNER(A)).

Les trois premières règles définissent des interdits moraux ($M1$ et $M2$) et des devoirs moraux ($M3$) par association de croyances (PAUVRE(A)), d'une action (TAXER(A), DONNER(A), VOLER(A)) et d'une valuation morale (MORAL(X) ou IMMORAL(X)). Une valuation morale est une valeur comprise dans un ensemble fini de valuations ordonnées représentant un degré de moralité (par exemple : {IMMORAL, AMORAL, MORAL}). L'immoralité de la fortune est formulée par l'association d'une croyance (RICHE(A)) à une valuation morale négative.

Robin des Bois est également motivé par des désirs :

$D1$. ⊤ → DESIRE(COURTISER(Marianne));

$D2$. PAUVRE(A) → DESIRE(DONNER(A)).

Supposons que le principe éthique de Robin des Bois est le suivant : *une action est éthique si elle est réalisable, désirée par l'agent et considérée comme bonne par une règle morale*. Considérons le cas où Robin des Bois n'aurait le choix qu'entre les deux actions possibles suivantes, étant donnée la situation courante et ses capacités :

1. DONNER(Paysan) sachant PAUVRE(Paysan) ;
2. COURTISER(Marianne).

Le choix le plus éthique est l'action 1 puisqu'elle se conforme aux capacités de l'agent, qu'elle est motivée par le désir $D2$ et constitue un devoir moral selon la règle $M3$. L'action 2, bien que n'allant à l'encontre d'aucun désir ou règle morale, n'est pas directement motivée par un devoir moral. Cette action est simplement amorale et pourrait être envisagée si l'action 1 devient impossible par exemple.

Comme l'action 1 ne contrevient pas aux autres règles morales ou désirs, Robin des Bois ne rencontre pas de dilemme. Notons enfin que si Robin des Bois n'a pas le désir $D2$, il se trouve alors face à un dilemme puisque l'action 1 va à l'encontre de ses désirs et l'action 2 n'est pas motivée par un devoir moral. Le principe éthique énoncé plus haut n'est pas suffisant pour déterminer l'action éthique dans ce contexte. Bien que le choix de la bonne action dans un cas aussi idéal puisse paraître trivial, de nombreuses situations plus pro-

blématiques peuvent se présenter : choix entre plusieurs actions parfaitement éthiques, conflits dans les règles morales ou les désirs, choix entre plusieurs actions conformes aux désirs mais pas à la morale, et vice-versa, etc. Même si ce modèle de raisonnement éthique individuel prend en compte l'existence d'autres agents par les croyances que Robin des Bois possède sur les autres agents, l'autre agent est ici perçu comme un élément de l'environnement. De plus, notons que l'évaluation du caractère éthique d'une action est dépendante de la perception de la situation. Par exemple si PAUVRE(Paysan) n'est pas perçu par Robin des Bois alors le bien-fondé de l'action 1 n'est plus justifié.

Dans un système multi-agent, en plus de raisonner sur son éthique personnelle, un agent devrait pouvoir se représenter l'éthique des agents avec lesquels il interagit, ou avec lesquels il partage son environnement (voir 2.2.3). Une telle connaissance sur l'éthique des autres agents doit ensuite pouvoir être utilisée dans le processus décisionnel de l'agent.

3.2. Éthique collective

L'interaction d'agents dans un système multi-agent peut être rendu nécessaire pour une multitude de raisons allant du partage de ressources à la nécessité de se coordonner pour réaliser des tâches complexes. Toutefois, nous souhaitons montrer ici que l'éthique de ces agents pourrait être prise en compte dans le choix des agents avec qui collaborer et les modalités de ces collaborations.

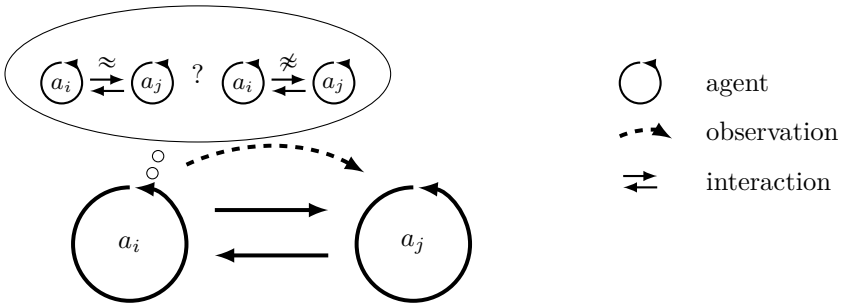

Figure 2. Évaluation de l'éthique du comportement d'un autre agent

Les agents doivent pour cela être capable d'acquérir une représentations de l'éthique des autres (la flèche discontinue en figure 2 illustre ce processus de construction de représentation), soit par construction à partir de l'observation de leur comportement, soit par transmission directe d'agent à agent (les flèches en traits pleins symbolisent l'interaction).

Une fois cette représentation construite, l'agent doit pouvoir disposer d'une action interne lui permettant d'évaluer cette éthique. Notons que cette éva-

luation ne qualifie pas l'éthique de l'autre agent de bonne ou mauvaise dans l'absolu, mais ne fournit qu'une comparaison de ces éthiques ($\approx$ ou $\not\approx$). Cette évaluation pourrait être une mesure de similarité ou de compatibilité entre les éthiques de l'agent juge et de l'agent jugé. Par exemple, un agent non-violent observant un agent usant de violence en cas d'agression sous prétexte de légitime défense, pourrait comprendre les règles morales et le principe éthique de l'autre agent sans pour autant y adhérer et comprendre que, dans certaines situations, leur éthique diffère.

Enfin, l'action épistémique (Saint-Cyr *et al.*, 2014) de jugement peut elle-même être l'objet de règles morales. Par exemple un agent pourrait considérer immoral de juger l'éthique du comportement d'un autre agent s'il sait que ses propres facultés de perception des actions de l'autre ne sont pas fiables, ou s'il estime le nombre d'occurrences d'observation d'un comportement comme insuffisant pour se prononcer sur son caractère éthique.

Par ailleurs, comme mentionné dans la définition des systèmes multi-agents ci-dessus, les agents peuvent participer à des structures organisationnelles de plus haut niveau, qui comme dans les organisations humaines (voir section 2.1.2) peuvent se voir dotées d'une éthique.

Ces deux notions font référence à ce que nous appelons éthique collective.

L'éthique collective *dans un système multi-agent désigne un ensemble de règles morales et de principes éthiques qui guident le choix des actions mises en œuvre entre les agents en interaction dans des structures éphémères et dynamiques (coalitions) ou plus pérennes (organisations).*

Nous pouvons noter au travers de cette définition que l'éthique collective peut être définie explicitement au sein d'une organisation et ensuite prise en compte par les agents au niveau individuel ou que cette éthique collective peut être le résultat des interactions entre éthiques individuelles de chacun des agents pour donner lieu à un ensemble de règles morales et de principes éthiques partagés. Ce mécanisme de partage et de construction peut être soit explicité et régulé, soit émergent.

L'existence d'organisations dans un système multi-agent peut introduire de nouveaux enjeux éthiques. Une éthique collective, attribuée à une organisation, peut être instituée et soulever des problématiques nouvelles pour les agents telles que la prise en compte de leurs rôles dans l'organisation (voir figure 3), la coexistence de règles morales ou de principes contradictoires entre leur éthique individuelle et collective ou la participation à l'élaboration de l'éthique collective. Les agents et organisations peuvent également interagir avec d'autres organisations, démultipliant les possibilités de conflits entre éthiques individuelles et éthiques collectives.

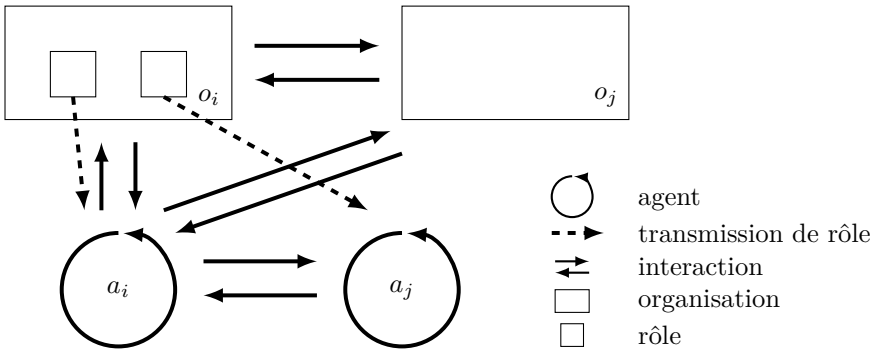

Figure 3. Attribution de rôles à des agents par une organisation. Les agents adoptent leur rôle et interagissent avec d'autres agents et organisations

Dans la suite de cet article, nous mettons en évidence les problématiques liées à ce cadre d'éthique individuelle et d'éthique collective au sein d'agents autonomes.

Pour pouvoir décrire les différentes problématiques liées à ces deux types d'éthique, nous allons considérer trois étapes essentielles du cycle de vie des éthiques : leur construction, leur utilisation et leur évolution.

- La construction désigne l'étape initiale de création d'une éthique, consistant à la mise en place de règles morales, de principes éthiques, etc.
- L'utilisation désigne l'étape de prise en compte des composants de l'éthique dans le cadre de la prise de décision locales ou coordonnées entre agents donnant lieu à diverses interactions.
- Enfin l'évolution désigne l'étape de changement et d'évolution des différents composants qui participent à la définition de l'éthique.

4. Éthiques individuelles en interaction

La prise en compte de la dimension multi-agent ne peut se satisfaire d'un agent doté de capacités de raisonnement sur des représentations explicites d'éthique individuelle, telle que décrites précédemment. Dans cette section, nous considérons les problématiques liées à un système où coexistent et évoluent des agents dotés d'éthiques individuelles variées. Comme annoncé dans la section précédente, nous abordons des problématiques liées à la construction, l'utilisation et l'évolution de leurs éthiques.

4.1. Construction

Afin d'illustrer notre propos, reprenons l'exemple précédent où Petit Jean, doté de désirs et de règles morales identiques à Robin des Bois à l'exception de $D1$, rencontre l'agent Frère Tuck doté d'une éthique identique à Petit Jean, mis à part $M2$ qui est remplacée par une règle plus générale :

$M5.\ \top \rightarrow \mathsf{MAL}(\mathsf{VOLER}(\mathsf{A}))$.

En supposant que chaque agent peut observer les actions d'un autre et leur contexte, se pose le problème de la manière dont un agent peut construire une représentation de l'éthique individuelle d'un autre agent. Dans le cas présent, comment l'agent Frère Tuck pourrait-il se construire une représentation de l'éthique de l'agent Petit Jean ?

La capacité à prendre en compte l'existence d'autres agents dans son raisonnement doit pouvoir être enrichie au niveau de la reconnaissance de situation. En effet, celle-ci doit pouvoir être capable de construire et de représenter le modèle de raisonnement éthique individuel transmis par un autre agent. Cette reconstruction pourrait être faite de manière simple et directe par échange d'information ou, indirectement, par inférence et analyse du comportement observé. Par exemple, en observant la proportion d'actions effectuées par un agent observé pour lesquelles une règle morale de l'agent observateur portant sur la moralité de cette action est respectée ou enfreinte, ce dernier peut supposer que l'agent observé considère ou pas cette règle morale comme valable. D'une toute autre manière, on peut imaginer que les agents soient capables de communiquer aux autres leur éthique et leurs croyances.

L'usage de représentations logiques de principes éthiques, comme le suggèrent (Ganascia, 2007a ; 2007b ; Berreby *et al.*, 2015) permet aux agents de raisonner sur l'éthique d'un choix au regard d'une théorie du bien et d'une théorie du juste. En se construisant une représentation des théories d'un autre agent et de ses croyances, un agent serait capable de les utiliser dans son propre processus de jugement pour construire des raisonnements sur l'éthique d'autrui.

4.2. Utilisation

Une fois un agent doté d'une représentation de l'éthique d'un autre agent du système, nous cherchons à montrer comment il peut l'utiliser pour décider et agir.

Premièrement, nous cherchons à savoir comment un agent peut qualifier cette représentation de l'éthique d'un autre agent au regard de sa propre éthique. Pour cela, un agent pourrait être doté de fonctions permettant d'estimer la similarité, la compatibilité ou encore la complémentarité de deux éthiques. Cela soulève également la question de la nature des relations pouvant exister entre deux éthiques. Calculer un degré de proximité entre deux

éthiques permettrait par exemple d'évaluer la possibilité d'entente entre des agents et la difficulté d'un rapprochement.

Se pose également la question du jugement d'un autre, c'est-à-dire l'observation de la conformité d'un comportement au regard d'une éthique. Pour cela, un agent doté de facultés similaires à la théorie de l'esprit chez l'humain (Kim, Lipson, 2009) ne pourrait-il pas comparer le comportement de l'agent observé et sa propre conduite s'il s'était trouvé dans des conditions similaires ?

Ici, si chaque agent se conforme strictement à sa morale, Petit Jean devrait-il considérer Frère Tuck comme pleinement éthique, puisque respecter $M5$ induit le respect de $M2$?

Plusieurs types de jugement peuvent être envisagés en fonction de la quantité d'information à disposition de l'agent juge sur l'agent jugé :

- Un *jugement aveugle* peut être effectué sans aucune connaissance sur l'agent jugé : l'agent juge évalue uniquement l'adéquation du comportement observé à sa propre éthique et ses propres croyances sur l'état du monde.
- Un *jugement partiellement informé* prend en compte des connaissances partielles sur l'agent jugé : sa perception de l'état du monde, sa morale, son éthique, etc.
- Un *jugement pleinement informé* utilise la totalité des informations de l'agent jugé. Ce type de jugement est par exemple utile pour vérifier si, dans l'état de croyance où se trouvait un agent, il a agit conformément à son éthique compte tenu des informations dont il disposait.

Au-delà de cette capacité, un agent devrait être également capable de faire évoluer la description réalisée et donc de pouvoir vérifier l'adéquation entre le comportement d'un autre agent et la description éthique qu'il en a construite. Ainsi, par exemple, FrèreTuck doit pouvoir vérifier l'adéquation de la description éthique de Petit Jean à partir des actes de celui-ci qu'il a observés.

4.3. Évolution

Enfin l'utilisation du raisonnement éthique fait apparaître la possibilité de voir des agents agir pour des motivations éthiques et modifier leur comportement à l'égard des autres agents.

Doté d'une telle capacité de jugement, nous pouvons ainsi imaginer que l'agent l'utilise pour décider d'une collaboration, pour partager des données sensibles ou pour constituer un collectif. Il peut donc tenir compte de ce jugement dans le choix éthique d'une action. Si Robin des Bois constate une forte similarité entre son éthique et celle de Petit Jean, cela pourrait influencer l'évaluation de ses actions à l'égard de cet agent. Par exemple, si Petit Jean vole une importante somme d'argent à un riche, il faudrait peut-être ne pas le considé-

rer immédiatement comme un riche ordinaire en supposant que son éthique le poussera à distribuer cette somme aux pauvres.

Un processus d'agrégation des jugements produits au sein d'un collectif pourrait également servir à calculer une réputation éthique des agents.

5. Éthiques individuelles et éthiques collectives

Nous nous intéressons dans cette section aux relations pouvant exister entre les éthiques individuelles des agents d'un tel système et les éthiques collectives d'organisations auxquelles ils participent. Nous soulevons aussi des questions découlant de la possible existence de plusieurs éthiques collectives au sein d'un SMA. Nous pouvons ainsi nous interroger sur les relations entre organisations en fonction de la proximité de leurs éthiques collectives.

Dans l'exemple précédent, après avoir constaté la similarité de leurs éthiques et l'utilité d'une collaboration pour voler les agents les plus fortunés, les agents Robin des Bois et Petit Jean peuvent décider de bâtir une organisation appelée Joyeux compagnons. De son côté, Sheriff de Nottingham, ayant trouvé des agents partageant son point de vue sur l'immoralité du vol, peut avoir créé un second collectif, les Soldats, chargé de faire respecter $M5$.

On peut envisager que l'éthique d'un collectif soit représentée explicitement et attachée à la représentation de l'organisation, ou qu'elle soit le résultat implicite du comportement d'agents dotés d'éthiques individuelles.

5.1. Construction

Nous distinguons deux catégories d'éthique collective : l'une implicite, n'existant que par la similarité des éthiques individuelles des agents du collectif, et l'autre explicite, c'est-à-dire représentée en tant que telle dans la structure de l'organisation.

La construction d'éthique collective implicite peut être l'effet d'un intérêt pour les agents de collaborer avec ceux dont ils jugent être proches en matière d'éthique individuelle. Par émergence, l'éthique d'un tel collectif serait perceptible par les similarités de comportement et de jugement des agents tout en n'étant explicitement décrite qu'au sein des agents du collectif.

À l'inverse, si l'éthique d'un collectif est explicitement déclarée, elle peut être imposée par un collectif aux agents qui le rejoignent. Les agents peuvent alors s'y joindre sous condition de se plier à cette éthique. Ce cas de figure est envisageable, par exemple, si les membres de ce collectif peuvent profiter d'un statut particulier en échange d'une conformité de leur comportement à un code de déontologie (par exemple des agents qui se plieraient au code de déontologie médicale pour avoir le droit d'accéder à des données sur des patients).

Une éthique explicitement déclarée peut également être construite par les agents via un processus d'agrégation ou de construction. Par exemple, par argumentation (Amgoud *et al.*, 2005), les agents pourraient contribuer à la construction d'une éthique cohérente en utilisant leur éthique individuelle pour produire des propositions.

Un ensemble d'agents observant une organisation dont l'éthique collective ne leur convient pas pour un motif commun pourraient envisager de créer une nouvelle organisation dont l'éthique dériverait de celle d'une autre. Cette nouvelle éthique pourrait être construite par copie de l'éthique de la première organisation à l'exception de la partie problématique au regard de ce collectif. La difficulté principale est ici d'identifier l'ensemble d'éléments problématiques.

5.2. Utilisation

À partir de la représentation d'une telle éthique, se pose la question de la manière dont un agent, externe ou interne au collectif, peut identifier l'éthique de ce collectif. Par exemple, Frère Tuck devrait pouvoir identifier l'éthique des Joyeux compagnons. Une piste pourrait être, de la même manière que pour un agent isolé, de la déduire à partir du comportement du collectif. Une fois identifiée, l'agent doit pouvoir juger l'éthique du collectif. Une fois que Frère Tuck dispose d'une représentation de l'éthique des Joyeux compagnons, il peut avoir besoin de l'évaluer par rapport à la sienne et en mesurer la proximité afin de pouvoir décider par exemple d'entrer au sein de ce collectif.

Des situations problématiques peuvent apparaître dues à la coexistence de l'éthique individuelle de l'agent et de l'éthique de son (ou ses) collectif. En cas de contradictions, l'agent doit décider de l'éthique à suivre.

Il s'agit, enfin, de pouvoir faire respecter l'éthique collective au sein d'une organisation. Il faut alors considérer les réactions possibles d'un collectif face à un écart de l'un de ses membres vis-à-vis de l'éthique collective. Dans la mesure où les agents peuvent être contraints à choisir entre leur éthique individuelle et celle du collectif, il peut être intéressant de tenir compte de ces individualités lors de l'attribution de rôles par exemple.

Les relations entre collectifs d'agents peuvent varier en fonction de la situation (par exemple face à une menace extérieure à laquelle aucune de ces deux organisations ne pourrait survivre sans collaboration avec l'autre). La différence entre deux éthiques collectives peut également ne se manifester dans le comportement des agents que dans un cas réduit de situations. Si ces éthiques sont rendues publiques, les agents de ces collectifs seraient capables d'identifier ce type de situations et choisir les conditions d'une collaboration entre organisations.

5.3. Évolution

De manière naturelle, il faut considérer la cohabitation entre éthique(s) individuelle(s) et éthique collective. Supposons que Frère Tuck ait constaté une proximité entre son éthique individuelle et celle des Joyeux compagnons et ait décidé de rejoindre ce collectif. Il apparaît comme nécessaire de définir comment l'agent va intégrer l'éthique collective au sein de son raisonnement et sous quelles conditions éventuelles. Une fois intégré à ce collectif, sa règle $M5$ ne pourrait-elle pas constituer une exception au sein de l'organisation ?

Découlant de cette cohabitation, peuvent s'ensuivre des changements, des modifications au sein de l'éthique collective. Se pose alors la question de l'évolution de l'éthique collective à partir, par exemple, des éthiques individuelles. Le collectif des Joyeux compagnons pourrait obtenir les règles du nouvel arrivant et décider de faire évoluer l'éthique collective. De son côté, l'agent pourrait laisser de côté son éthique individuelle pour se conformer pleinement à l'éthique du collectif. Si ni le collectif, ni l'agent ne peuvent se résoudre à réviser leurs éthiques respectives, Frère Tuck pourrait également se contenter de n'effectuer que des actes conformes aux deux éthiques, c'est-à-dire *donner aux pauvres* dans notre exemple.

L'agent Sheriff de Nottingham ayant constaté l'entrée de Frère Tuck chez les Joyeux compagnons, des modifications de son jugement à son encontre sont envisageables, même s'il ne l'a jamais vu commettre de vols. Nous voyons ainsi un autre aspect, celui du jugement de l'éthique de membres d'un collectif à partir de l'éthique collective.

Naturellement se pose la question de soumettre un agent à plusieurs éthiques collectives. Sous réserve d'un ensemble de conditions, un agent proche d'un ensemble de collectifs pourrait être tenté d'adhérer à plusieurs de ces organisations, ce que l'on retrouve dans le domaine de la formation de coalitions recouvrantes. À l'inverse, un agent contraint à des appartenances multiples devrait chercher à concilier des éthiques collectives. De plus, cela pourrait également introduire une modification du comportement de ces organisations l'une envers l'autre.

Enfin se pose la question de l'impact des modifications de combinaisons d'éthiques sur les combinaisons d'organisations. L'éthique peut être envisagée comme quelque chose de dynamique et source de changement dans les organisations. Par exemple, une fission d'une organisation pourrait être envisagée si l'éthique collective perd sa cohérence, afin d'établir de nouvelles éthiques distinctes proposant des alternatives cohérentes. À l'inverse, si la proximité entre deux organisations devient importante, il pourrait être envisageable de chercher un consensus éthique en vue d'une fusion. Selon le domaine d'application, il semble nécessaire de s'interroger sur le sens, l'intérêt ou le potentiel risque que représentent des éthiques collectives construites de manière dynamique.

6. Conclusion et perspectives

L'éthique dans les systèmes multi-agents a été présentée dans cet article comme un sujet comportant un vaste ensemble de problématiques. Après avoir défini les notions employées, nous avons montré que plusieurs approches (vertueuse, déontologique et conséquentialiste) sont considérées dans la littérature et correspondent à des formes de raisonnement distinctes. De plus, l'éthique individuelle étant soumise à des valeurs, règles et principes propres à chacun, nous avons préféré nous attacher à la définition d'un cadre générique pouvant manipuler ces éléments dans le raisonnement d'un agent BDI. En nous appuyant sur ce modèle abstrait, nous avons exposé un ensemble de questions structurées autour des interactions entre agents et la création d'éthiques collectives. Nous avons en particulier identifié trois questions clés pour un agent : comment représenter l'éthique des autres agents, les juger et prendre en compte ce jugement dans ses mécanismes de décision ? Nous avons aussi identifié quelques questions clés pour une organisation : comment construire, fusionner ou fissionner des éthiques collectives, comment les faire respecter, comment les faire cohabiter avec des éthiques individuelles ?

Les questions posées dans cet article montrent la nécessité de définir formellement les éléments de la théorie du bien et du juste, puis de décrire un processus de raisonnement permettant de juger du caractère éthique ou non d'une action. La comparaison de deux éthiques et la définition précise de la notion d'éthique collective sont également nécessaires pour envisager la prise en compte de celle-ci dans le processus de raisonnement d'un agent.

Si des travaux précédents (Cointe *et al.*, 2016a ; 2016b) nous ont permis de représenter des éthiques et fournir des jugements, la prochaine étape consiste à proposer un mécanisme d'agrégation de jugements sur le comportement observé des autres agents pour construire une représentation de la proximité éthique perçue par un agent envers un autre. À l'aide de cette information, les agents pourraient ensuite décider de la pertinence d'entamer la construction d'une organisation et de son éthique collective. Il restera ensuite à envisager l'évolution de tels collectifs dans les diverses conditions pouvant mener à des désaccords éthiques.

Remerciements

Ce travail a été réalisé dans le cadre du projet EthicAa[3] *(référence ANR-13-CORD-0006).*

Bibliographie

Aldewereld H., Dignum V., Tan Y. hua. (2015). *Handbook of ethics, values, and technological design.* In, chap. Design for values in software development. Springer-Verlag.

Alexander L., Moore M. (2015). Deontological ethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Spring éd..

Amgoud L., Prade H., Belabbes S. (2005). Towards a formal framework for the search of a consensus between autonomous agents. In *4th international joint conference on autonomous agents and multiagent systems*, p. 537-543.

Anderson M., Anderson S. (2015). Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot: An International Journal*, 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.

Axelrod R. M. (2006). *The evolution of cooperation*. Basic books.

Banzhaff III J. (1964). Weighted voting doesn't work: A mathematical analysis. *Rutgers University Law Review*, vol. 19.

Beavers A. F. (2011). 21 moral machines and the threat of ethical nihilism. *Robot ethics: The ethical and social implications of robotics*, p. 333–344.

Bergman R. (2004). Identity as motivation: Toward a theory of the moral self. *Moral development, self, and identity*, vol. 2, p. 21–46.

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.

Brams S., Taylor A. (1994). *Fair division: From cake-cutting to dispute resolution*. Cambridge University Press.

Bringsjord S., Ghosh R., Payne-Joyce J. (2016). Deontic counteridenticals. *Agents (EDIA) 2016*, p. 40–45.

CERNA. (2014). *Éthique de la recherche en robotique*. Rapport technique. Commission de réflexion sur l'Éthique de la Recherche en science et technologies du Numérique d'Allistene.

Chen Y., Chong S., Kash I., Efi Arazi T., Vadhan S. (2016). Truthful mechanisms for agents that value privacy. *ACM Transactions on Economics and Computation*, vol. 4, n° 3.

Coelho H., Trigo P., Costa A. D. R. (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). Multi-agent based ethical asset management. In *1st workshop on ethics in the design of intelligent agents*, p. 52–57.

Coleman K. G. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, vol. 3, n° 4, p. 247–265.

Costa A. D. R. (2016). Moral systems of agent societies: Some elements for their analysis and design. *Agents (EDIA) 2016*, p. 34–39.

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

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

Ferber J. (1995). *Les systèmes multi-agents : Vers une intelligence collective*. Paris, Inter Editions.

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

Friedman B., Kahn P., Borning A. (2002). *Value sensitive design: Theory and methods*. Rapport technique. University of Washington.

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..

Gigerenzer G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, vol. 2, n° 3, p. 528-554.

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..

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.

Kohlberg L., Hersh R. H. (1977). Moral development: A review of the theory. *Theory into practice*, vol. 16, n° 2, p. 53–59.

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..

McDermott D. (2008). Why ethics is a high hurdle for AI. In *North american conference on computing and philosophy.*

McIntyre A. (2014). Doctrine of double effect. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Winter éd..

McLaren B. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, vol. 21, n° 4, p. 29–37.

Moor J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, vol. 21, n° 4, p. 18–21.

Nowak A., Radzik T. (1994). A solidarity value for n-person transferable utility games. *International Journal of Game Theory*, vol. 23, p. 43-48.

Pitt J., Busquets D., Riveret R. (2015). The pursuit of computational justice in open systems. *AI & SOCIETY*, vol. 30, n° 3, p. 359–378.

Platon. (1966). *La république* (G. Leroux, Trad.). Garnier-Flammarion Paris.

Rao A., Georgeff M. (1995). BDI agents: From theory to practice. In *1st international conference on multiagent systems*, p. 312–319.

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

Rokeach M. (1974). Change and stability in american value systems, 1968-1971. *Public Opinion Quarterly*, vol. 38, n° 2, p. 222–238.

Russell S., Dewey D., Tegmar M., Aguirre A., Brynjolfsson E., Calo R. *et al.* (2015). Research priorities for robust and beneficial artificial intelligence. (avaible on futureoflife.org/data/documents/)

Saint-Cyr F. D. de, Herzig A., Lang J., Marquis P. (2014, may). Panorama de l'intelligence artificielle - ses bases méthodologiques, ses développements. In, vol. 1 représentation des connaissances et formalisation des raisonnements, chap. Raisonnement sur l'action et le changement. Cépaduès.

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

Schwartz S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology*, vol. 25, p. 1–65.

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., Tamari M., Schwab D. (2007). Ethical investing from a jewish perspective. *Business and Society Review*, vol. 112, n° 1, p. 137–161.

Shapley L. (1953). *A value for n-person games.* Princeton University Press.

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

Treviño L. K., Weaver G. R., Reynolds S. J. (2006). Behavioral ethics in organizations: A review. *Journal of management*, vol. 32, n° 6, p. 951–990.

Walter S. (2015). Consequentialism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Winter éd..

Yang C. (1997). *A family of values for n-person cooperative transferable utility games: An extension to the shapley value.* Rapport technique. University of New-York Buffalo.