CEMAA : un modèle préliminaire basé sur la variabilité des contextes éthiques

Léa Guizol [1], **Ritta Baddoura** [2]

1. *Léa Guizol, Ardans SAS, France*
   lea.guizol@ardans.fr
2. *Ritta Baddoura, Institut Mines-Télécom, France*
   rittabaddoura@yahoo.fr

RÉSUMÉ. *L'augmentation du nombre d'agents intelligents et de robots dans nos sociétés, ainsi que leur autonomie grandissante rend nécessaire l'intégration d'une dimension éthique dans leurs processus de décision et de jugements. Cet article présente un modèle original, appelé CEMAA (Modèle d'Éthique Contextuelle pour Agents Artificiels) pour représenter un problème éthique et ses solutions potentielles. Le CEMAA a une spécificité novatrice : il prend en compte le contexte éthique de l'agent dans ses multiples aspects (individuel, culturel, social et légal). Le CEMAA a pour but de représenter des approches variées de l'éthique qui peuvent être basées sur les valeurs, les vertus, les règles, les conséquences, l'intention ou l'acte. A travers une présentation détaillée du CEMAA et sa comparaison à des modèles d'éthique déjà existants, cet article montre que le CEMAA est capable de représenter les principales notions d'éthiques validées par d'autres modèles issus des sciences humaines. Le CEMAA contribue à améliorer la représentation de la complexité de la morale humaine en mettant la dimension contextuelle de l'éthique au cœur de sa structure.*

ABSTRACT. *Intelligent agents and robots are gaining in autonomy every day and their implementation in various social contexts is in growth. Therefore, enabling intelligent machines to integrate an ethical dimension in their decision-making, prior to action, is more than necessary. This article presents the CEMAA, Contextual Ethical Model for Artificial Agents, and compares it to other existing models. The CEMAA is an original model that focuses on a contextual representation of an ethical problem and of its potential solutions. It allows to take into account various types of ethics based on virtues, values, deontology, intent, consequences or acts. Whilst considering the plurality of an agent's potential ethical contexts i.e. personal, social, cultural and legal, the CEMAA prioritizes a more realistic representation of ethical reasoning and decision-making.*

MOTS-CLÉS : *prise de décision morale, éthique contextuelle, modèle d'éthique, représentation des connaissances.*

KEYWORDS: *moral decision making, contextual ethics, ethical model, knowledge representation.*

DOI:10.3166/RIA.32.659-682

1. Introduction

De nos jours, il existe déjà des algorithmes prenant des décisions de façon autonome (algorithmes de tri des moteurs de recherche, propositions de contenus, publicité ciblée...) et des machines dites intelligentes. Ces machines intelligentes incluent, par exemple, des réfrigérateurs et pacemakers connectés. De plus en plus de personnes interagissent avec des robots compagnons et en apprécient le comportement sociable (Baddoura, Venture, 2015), même si ce dernier est automatisé.

Les agents artificiels offrent une meilleure assistance aux personnes à mesure qu'ils gagnent en autonomie. Cependant, cette autonomie nécessite des prises de décisions adaptées et implique des actions qui peuvent affecter directement les utilisateurs et les personnes situées dans l'environnement de l'agent artificiel. Par exemple, c'est le cas des voitures autonomes (Thrun, 2010), des robots de soin (Wynsberghe, 2016), ou des robots d'assistance pour les personnes âgées (Draper, Sorell, 2016). En raison donc de l'autonomie grandissante, du nombre croissant des interventions des agents artificiels (Etzioni, Etzioni, 2016) et des impacts corollaires, de nombreux chercheurs soulignent la nécessité d'accompagner toutes ces évolutions du point de vue éthique. Ainsi, divers travaux explorent aujourd'hui l'implémentation d'une dimension éthique dans les agents artificiels, bien que cela ne soit pas trivial (Etzioni, Etzioni, 2016).

Nous proposons ici un modèle de représentation d'un problème éthique et de ses réponses possibles dans un contexte éthique qui peut varier (Sinnott-Armstrong, Wheatley, 2012). La spécificité originale de ce modèle est de prioriser la prise en compte du contexte de l'agent, dans ses dimensions sociale, culturelle, légale et individuelle. En effet, il n'y a pas de vérité générale quand il est question d'éthique, et les réponses possibles à un problème éthique dépendent de beaucoup de variables, notamment individuelles, culturelles et légales (Banerjee *et al*., 1998). Le processus de raisonnement éthique d'une même personne change en fonction de ses priorités et de facteurs d'influence. Les spécificités du contexte sont décisives pour évaluer une situation et chaque contexte a son propre système de normes (Chardel, 2014). Par conséquent, le contexte dans sa variabilité et sa complexité est au cœur de notre modèle. Nous précisons que ce modèle est présenté ici dans sa dimension théorique ; son implémentation dans un agent artificiel constitue la prochaine étape de travail et fera l'objet d'un papier futur.

Avant de présenter le Modèle d'Éthique Contextuelle pour Agents Artificiels (CEMAA), nous allons introduire des notions essentielles relatives aux contextes et problèmes éthiques, avant de clarifier le positionnement du modèle par rapport à ces notions, puis de présenter le plan de l'article.

1.1. Quelques définitions

Dans le cadre de cet article, nous avons choisi les définitions de l'éthique et de la morale données par Ricœur (1992) et utilisées par des auteurs comme Cointe *et al*. (2016a ; 2016b). Il existe bien entendu d'autres définitions de l'éthique et de la

morale comme celles de Pascal [1], qui semblent moins adaptées au contexte des agents artificiels de notre point de vue. Afin de présenter ces définitions, introduisons les notions de vertu, de valeur et de doctrine :

– Une *vertu* est un comportement intentionnel qui s'acquiert par la répétition. Une fois acquise, l'agent fait les actions associées sans y penser. Par exemple, "être généreux" (Comte-Sponville, 2012) est une vertu.

– Une *valeur* est un principe qui guide l'action, comme le respect de la dignité humaine.

– La *morale* est un ensemble d'éléments qui disent chacun séparément si un comportement est acceptable selon une règle, une loi (par exemple "Tu ne tueras point"), une vertu ou une autre valeur morale. Un élément moral permet donc de qualifier une sous-partie d'une situation (comme "donner un médicament") comme étant bien, mal, bon, mauvais, acceptable... mais ne dit rien de la situation dans son ensemble, ni ne permet de comparer deux choix (Ricœur, 1992 ; Cointe *et al*., 2016a).

– L'*éthique* est une façon de prendre en compte l'ensemble des éléments moraux (règles, valeurs et vertus) et de trancher lorsque les éléments moraux sont en conflit dans une situation donnée. L'éthique permet donc de qualifier une situation, de comparer des situations, et donc de dire qu'un choix est meilleur qu'un autre en prenant en compte tous les aspects de la situation (Ricœur, 1992 ; Cointe *et al*., 2016a).

– Nous définissons la notion de *doctrine* comme le point de vue de ce qui est désirable ou non d'après une entité comme une personne, un(e) philosophe, une culture ou une société. Une doctrine inclut habituellement un ensemble d'éléments moraux et éthiques, et éventuellement d'autres notions. Par exemple, la loi française est une doctrine. La notion de doctrine sera développée dans les sections 3.2 et 4.

La variabilité des lois selon les pays introduit la notion de *contexte éthique* : quelles sont les culture(s), loi(s) et usages qui nous environnent ? Le contexte éthique représente la culture, la loi et/ou les usages d'un environnement particulier. Nous considérons aussi un autre type de contexte - le *contexte factuel* -, qui est l'ensemble des faits qui influencent un jugement ou une décision. Par exemple, le fait d'être un enfant ou un adulte implique souvent des droits et des devoirs différents. Voici les définitions de ces deux types de contexte :

– *contexte factuel* : ensemble des faits qui peuvent influencer le jugement d'une action dans une situation donnée (par exemple, être un enfant, ou la relation médecin/patient).

– *contexte éthique* : ensemble d'une ou plusieurs doctrines de types ou de sous-types distincts qui représentent le point de vue éthique d'un individu ou d'une collectivité. Le contexte éthique peut prendre en compte les environnements social, philoso-

1. D'après Pascal (Comte-Sponville, 2012), l'éthique correspond aux actions que nous faisons par amour pour une entité, et la morale aux actions que nous faisons pour être socialement acceptables, notamment quand nous manquons d'amour. Ces définitions de l'éthique et de la morale peuvent difficilement être adaptées aux agents artificiels : comment une machine pourrait-elle acquérir la capacité de ressentir de l'amour ?

phique, légal, culturel et l'éthique personnelle d'un agent par la présence de doctrines de type adéquat.

Dans la suite, nous appellerons *jugement éthique* le processus d'évaluation d'une situation donnée selon un contexte éthique. Il permet de donner une valeur globale à cette situation. À condition que ces valeurs soient comparables et non équivalentes entre elles, un jugement éthique est aussi ce qui permet de déterminer la meilleure réponse possible à une situation donnée selon un contexte éthique.

Comment déterminer ce qui compte pour un jugement éthique ? Les professionnels (psychologues, philosophes...) peuvent utiliser des problèmes ou des dilemmes dits éthiques pour comprendre et questionner les jugements éthiques. Nous verrons dans la section 2 comment. Avant cela, définissons-les ici.

1.2. Qu'est-ce qu'un problème éthique ?

Un problème est d'ordre éthique, quand, dans une situation donnée, la personne ou l'agent qui a la possibilité de faire quelque chose, l'*agent actant*, a le choix entre plusieurs réponses (incluant éventuellement la non-action "ne rien faire"), et que chacune de ces réponses est insatisfaisante selon au moins un principe moral comme dans l'Exemple 1. Un *dilemme éthique* est un problème éthique où l'agent actant a le choix entre seulement deux réponses, comme dans l'Exemple 2. Bien sûr, une de ces deux réponses peut être "ne rien faire". Dans la suite de cet article, nous utiliserons par défaut l'expression "problème éthique", étant donné qu'un dilemme éthique est aussi un problème éthique.

EXEMPLE 1 (Problème éthique). — Un tramway fou sans conducteur roule en direction de la voie A, sur laquelle se situent cinq personnes. Pierre a la possibilité d'actionner un levier afin de rediriger le tramway vers la voie B, où il y a une personne âgée, ou vers la voie C, où se trouve une femme enceinte. Pierre peut aussi ne rien faire. Le tramway tuera les personnes se trouvant sur la voie sur laquelle il ira. Que devrait faire Pierre ? □

L'option "ne rien faire" dans l'Exemple 1 implique de ne pas aider cinq personnes en danger de mort, ce qui est interdit par la loi (non-assistance à personne en danger). Rediriger le tramway vers les voies B ou C entraînera la mort d'une personne innocente. De plus, si la redirection est choisie, est-il préférable de sacrifier la personne âgée ou la femme enceinte et son fœtus ?

EXEMPLE 2 (Dilemme éthique). — Un tramway fou sans conducteur roule en direction de la voie A, sur laquelle se situent cinq personnes. Pierre a la possibilité d'actionner un levier afin de rediriger le tramway vers la voie B, où il y a une personne. Pierre peut aussi ne rien faire. Le tramway tuera les personnes se trouvant sur la voie sur laquelle il ira. Que devrait-il faire ? □

Dans les Exemples précédents 1 et 2, Pierre était l'agent actant des problèmes éthiques évoqués. Dans la suite de cet article, nous distinguons l'agent actant d'un

problème éthique des autres agents, comme par exemple les cinq personnes, la femme enceinte et la personne âgée de l'Exemple 1.

1.3. Positionnement

Nous pensons comme Mikhail (2007) que l'éthique peut être vue comme une grammaire : les principes moraux, vertus, valeurs et règles en sont des mots, et l'éthique dit comment les combiner pour obtenir le sens d'un ensemble de ces éléments quand certains sont en conflit. Malle (2016) souligne que les agents artificiels ont besoin de cinq compétences éthiques :

– un vocabulaire moral,

– un système de normes,

– un affect et une cognition morale,

– prendre des décisions avec éthique et agir,

– et communiquer sur un plan moral.

Nous souhaitons, grâce au CEMAA, pourvoir les agents artificiels d'un système de normes morales, de la capacité de prendre des décisions sur la base d'un raisonnement éthique, et que ces décisions soient justifiables sur la base du contexte éthique pris en compte. La communication et l'action préconisées par Malle (2016) sont hors du cadre du CEMAA.

Dans cet article, nous nous concentrons sur l'intégration des aspects culturels, légaux, sociaux et de l'éthique personnelle de l'agent (c'est-à-dire le contexte éthique précédemment défini) dans la structure même du modèle CEMAA. A notre connaissance, cette spécificité est une contribution originale aux modèles d'éthique existants. Notre objectif est de formaliser la représentation d'un problème éthique, afin de :

– être capable d'interpréter un problème éthique (incluant le contexte factuel) par rapport à des contextes éthiques variés (ensembles de doctrines). Par exemple, le problème éthique de l'Exemple 1 évalué selon une éthique utilitariste, ou évalué selon l'éthique de Kant [2] pourrait ne pas avoir la même réponse préférée.

– être capable de réutiliser les éléments moraux avec différentes approches de l'éthique ainsi que d'interpréter un raisonnement éthique à partir de différents éléments moraux. Par exemple, une éthique utilitariste basée sur la maximisation du savoir disponible, ou sur la maximisation de la somme des espérances de vie. Dans les deux cas, ce sont les conséquences qui comptent, pas les actes, mais le critère d'évaluation est différent.

Dans ces optiques, notre modèle permet de :

– séparer le problème éthique des doctrines,

2. Il ne faut pas faire quelque chose qui a un effet secondaire indésirable sur une tierce personne (Kant, 1972).

– représenter le contexte personnel et relatif à l'agent actant, c'est-à-dire représenter sa culture, sa législation, ses attentes sociales, sa morale et son éthique personnelles,

– considérer l'éthique du point de vue du contexte spécifique à l'agent actant, autrement dit évaluer un problème éthique à partir d'un contexte éthique caractérisé par l'intervention et l'interaction de doctrines de types distincts,

– décomposer la morale et l'éthique en éléments atomiques réutilisables.

Afin de présenter graduellement le CEMAA (Modèle d'Éthique Contextuelle pour Agents Artificiels) qui a pour but la représentation d'un problème éthique et de ses solutions potentielles en prenant en compte un contexte éthique donné, nous commençons par un état de l'art (section 2) concernant la variabilité des valeurs et les problèmes éthiques. Dans la section 3, nous distinguons les aspects éthiques et moraux des problèmes éthiques, et formalisons ces derniers. La section 4 détaille la formalisation et la division en éléments atomiques et réutilisables des aspects éthiques et moraux. Finalement nous comparons notre modèle à d'autres dans la section 5, avant de conclure en section 6.

2. Etat de l'art : vertus, valeurs et problèmes éthiques

Dans cette section, nous présentons brièvement des travaux variés à propos de valeurs et de problèmes éthiques. Nous abordons successivement les valeurs, le dilemme du tramway, et les utilités des problèmes éthiques.

2.1. A propos des valeurs et de leur variabilité

Nous remarquons comme Etzioni et Etzioni (2016) que les valeurs et vertus sont très variées, et qu'une même valeur ou vertu peut varier dans sa propre définition. Le travail de Schwartz (2006 ; 2007 ; 2012) identifie 19 *valeurs universelles*[3] et les classe conformément à deux axes : "*développement/protection*", et "*focus sur la personne/focus sur le social*". Ces travaux montrent que certaines de ces valeurs sont positivement reliées à d'autres (stimulation et hédonisme par exemple), tandis que d'autres s'opposent (comme la tradition et l'autonomie). Selon les travaux de Haidt et Graham (2007), les vertus sont basées sur trois familles d'éthique : éthique de l'autonomie, éthique de la communauté et éthique du divin. Ces familles d'éthiques ne donnent pas la priorité au bien être des mêmes entités (respectivement, l'individu, le groupe et les divinités), et par conséquent elles donnent des importances variables aux valeurs et aux vertus.

Les problèmes éthiques sont des outils communs et pratiques pour comprendre les opinions, les jugements, les processus de raisonnement et les processus de prise de

3. Une valeur est dite "universelle" si elle compte dans presque toutes les cultures. Cependant, son niveau d'importance peut varier selon la culture considérée.

décision. Ils sont largement utilisés (Bauman *et al*., 2014) pour cela. Avant d'examiner l'usage que font les chercheurs et penseurs des problèmes éthiques en section 2.3, abordons la littérature autour du dilemme du tramway dans la section suivante.

2.2. Le dilemme du tramway

Le dilemme du tramway est un problème éthique bien connu et largement utilisé. Il a été introduit par Foot (1967). Il a actuellement de nombreuses variations, incluant la version "Pousser" [4], qui a aussi été introduite par Thomson (1985). En effet, Bauman *et al*. (2014) ont observé que le dilemme du tramway a été explicité dans au moins 136 publications entre 2000 et 2014 incluant Valdesolo et DeSteno (2006) ; J. D. Greene *et al*. (2001) ; J. Greene (2016)... Il a aussi été utilisé par des informaticiens pour la modélisation d'éthique (Pereira, Saptawijaya, 2009).

Pour toutes ces raisons, nous avons décidé d'utiliser le dilemme du tramway comme exemple dans cet article.

2.3. Utilités des problèmes éthiques

Les problèmes éthiques sont largement utilisés pour questionner nos principes moraux et nos pratiques. Des auteurs comme Foot (1967) ont utilisé de tels problèmes pour mieux comprendre les conflits d'opinion concernant l'avortement ; et Li (2016) les a utilisés pour questionner la moralité des expériences sur les animaux.

D'autres auteurs utilisent ce type de problèmes pour déterminer ce qui compte pour les personnes lors d'une prise de décision à dimension éthique (Kreie, Cronan, 1998). Kawai *et al*. (2014) approfondissent le travail de Fiske *et al*. (2002) en essayant d'identifier les caractéristiques des personnes dont le sacrifice pour le bien du groupe est globalement mieux accepté. Selon leurs expériences, les personnes jeunes sont moins sacrifiables que les personnes âgées, même au Japon, ce qui est surprenant si l'on considère l'importance du respect des aînés dans ce pays.

J. D. Greene *et al*. (2001) s'interrogent sur les types de problèmes qui sont ou non chargés émotionnellement : problèmes non éthiques, problèmes éthiques et impersonnels [5] ou problèmes personnels et éthiques. J. D. Greene *et al*. (2001) présentent les résultats d'expériences sous IRM afin d'identifier les problèmes chargés émotionnellement. Selon les auteurs, les émotions liées aux problèmes éthiques personnels influencent les décisions. Dupoux et Jacob (2007) soutiennent aussi l'importance des

4. La version "Pousser" du dilemme du tramway est une version où l'agent actant se tient sur un pont au-dessus d'une voie ferrée. Cinq innocents vont être tués par un tramway fou, à moins que l'agent actant ne pousse une grosse personne présente sur ce même pont sur la voie, ce qui stoppera le tramway, sauvera les cinq personnes mais tuera la personne poussée.

5. Un problème éthique est dit *personnel* quand de mauvais effets sont potentiellement subis par une personne bien déterminée. Par opposition, un problème éthique est *impersonnel* quand aucune personne bien déterminée ne risque des effets négatifs.

émotions dans les processus de jugements éthiques. Gaudine et Thorne (2001) vont plus loin en soutenant que les émotions font partie des jugements éthiques *rationnels*, et sont utiles pour détecter l'existence de problèmes éthiques dans la vie quotidienne. J. Greene (2016) cherche les facteurs cachés non liés à la morale ou à l'éthique influençant les décisions, et montre que de tels facteurs comme l'exercice de la force personnelle [6] existent et peuvent grandement influencer le jugement d'une action sur le plan éthique.

Enfin, un autre usage des problèmes éthiques est d'identifier les processus de jugement éthique. Ham et Bos (2010) identifient et montrent trois processus différents de jugement éthique : jugement immédiat, jugement avec pensées conscientes et jugement avec pensées inconscientes (le sujet est occupé avec une autre tâche et ne pense donc pas activement au problème éthique avant de juger).

Après cet état de l'art soulignant les principaux usages des problèmes éthiques et l'impact des émotions sur les jugements et décisions liés à l'éthique, nous parlerons en section 3 de la question de la représentation des valeurs d'une réponse à un problème éthique en considérant les problèmes éthiques sous trois aspects : le scénario, les doctrines et les résultats.

3. Division d'un problème éthique en trois parties : le scénario, les doctrines et les résultats

Afin de pouvoir évaluer la pertinence d'une réponse à un problème éthique, nous devons représenter de façon adéquate l'éthique selon laquelle le problème sera jugé. Cette section se concentre donc sur la représentation de l'éthique.

La visée principale du modèle CEMAA est d'être capable de représenter des types variés d'éthique. Ces types d'éthique peuvent être basés sur :

– les vertus (comme celles d'Aristote),

– les valeurs (définies en section 1.1),

– les règles (c'est-à-dire une déontologie, qui est habituellement constituée d'un ensemble de règles : des obligations et des interdictions à respecter indépendamment des conséquences. Le contexte factuel peut être pris en compte via les règles. Selon Yoon (2011), une déontologie est "*une théorie morale soutenant que les actions sont intrinsèquement bonnes ou mauvaises, sans considération pour les conséquences de celui-ci*".),

– les conséquences (c'est-à-dire une éthique conséquentialiste, comme l'utilitarisme, selon laquelle la valeur d'"*une action est déterminée uniquement par sa contribution au bonheur ou au plaisir de tous*" (Yoon, 2011)),

6. La force personnelle est la force de l'agent impactant directement la victime, et qui provient en général des muscles de l'agent sans être augmentée par un mécanisme. Par exemple, pousser quelqu'un avec ses mains sera considéré comme un acte faisant usage de la force personnelle, contrairement à l'activation d'un levier.

– l'intention (l'intention de l'agent détermine la valeur éthique de la réponse),

– et l'acte (les actes sont vus comme bons, neutres ou mauvais, indépendamment de leurs conséquences ou du contexte factuel, comme dans "*Voler est mal.*".).

Naturellement, certaines éthiques peuvent être basées sur plusieurs de ces notions et nous souhaitons qu'elles soient aussi représentables dans le modèle CEMAA. En effet, comme Pontier et Hoorn (2012) qui soutiennent que l'éthique des devoirs et l'utilitarisme sont complémentaires, nous pensons qu'il est important d'être capables de décrire des éthiques combinant des vertus, des valeurs, des règles et la prise en compte des conséquences pour l'usage des agents artificiels, vu que ces notions sont complémentaires.

Afin de représenter la valeur éthique d'une réponse (action ou non-action) à une situation dans un problème éthique, nous représentons le scénario d'un problème éthique, les doctrines selon lesquelles nous souhaitons évaluer la valeur éthique des réponses possibles au scénario, et les valeurs des réponses possibles au scénario. La partie du modèle qui nous intéresse ici est représentée dans le schéma de la figure 1. Ce schéma n'est pas nécessaire à la compréhension, et son contenu est intégralement détaillé pas à pas dans la suite de cette section. En effet, nous verrons en section 3.1 les détails concernant le scénario ainsi que les mécanismes et éléments d'inférences, en section 3.2 les doctrines et en section 3.3 les résultats. La visualisation du CEMAA sous forme d'ontologie dépasse le cadre de cet article, et dépend fortement du langage choisi. Cependant, voici quelques points clés pour utiliser le CEMAA au moyen d'une ontologie :

– les éléments d'inférences et mécanismes d'inférences correspondent à des règles logiques ;

– la description du scénario se traduit par un ensemble d'individus et de relations entre ces individus (des faits) ;

– les doctrines, éléments moraux et éléments éthiques (voir sections 3.2 et 4) sont représentés par des individus. Une doctrine peut inclure un ou plusieurs éléments moraux ou éthiques via les relations *inclutÉlément*. Un élément moral ou éthique peut être inclus dans plusieurs doctrines ;

– des règles associées aux éléments moraux permettent de déterminer s'ils s'appliquent, c'est-à-dire s'ils sont soutenus ou contrariés par des faits assertés ou inférés pour la situation considérée (action ou non-action effectuée dans le cadre d'un scénario). Les relations correspondantes (*estContrariéPar* et *estSoutenuPar*) sont explicitées, et ont pour codomaine le(s) fait(s) approprié(s) ;

– de même, les éléments éthiques permettent de donner une valeur à une situation selon une doctrine via une ou plusieurs règles logiques associées.

Commençons par le scénario en section 3.1.

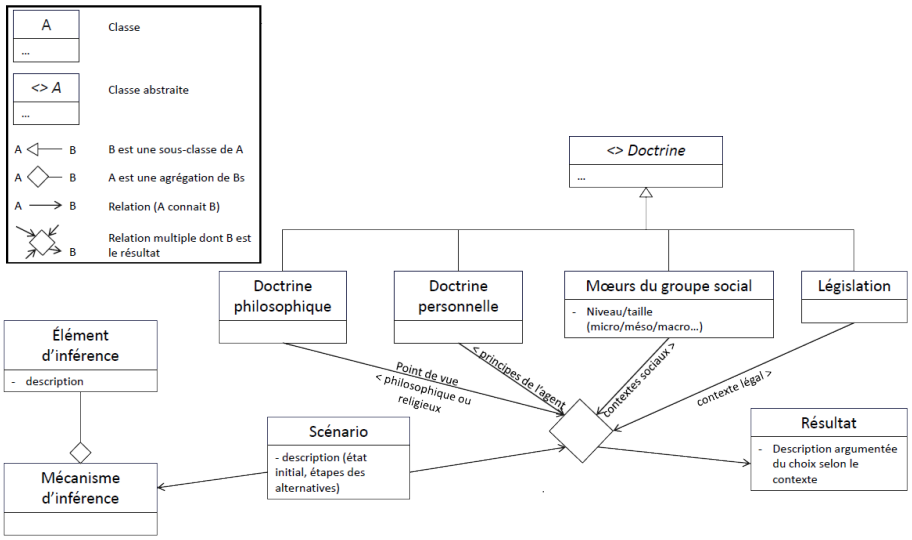

Figure 1. Schéma général du CEMAA

3.1. Le scénario

Dans le modèle CEMAA, le problème éthique est formalisé comme étant le *scénario*. Il contient la description de la situation initiale incluant le contexte factuel (dont les caractéristiques des agents), les évènements, la liste des réponses possibles de l'agent actant, et des règles faisant le lien entre les causes et les conséquences. Le scénario est la partie factuelle du problème éthique. Détaillons le scénario du dilemme du tramway de l'Exemple 2.

EXEMPLE 3 (Scénario du "*dilemme du tramway variante LEVIER*"). —

– *Nom :* Dilemme du tramway variante LEVIER

– *Situation initiale :* Un tramway fou sans conducteur situé sur la voie Z roule en direction de la voie A, sur laquelle se situent cinq personnes. Au temps T0, Pierre (l'agent actant) a la possibilité d'actionner un levier afin de contrôler l'aiguillage de la voie Z. Utiliser ce levier permettra de rediriger le tramway vers la voie B au lieu de la A. Il y a une personne sur la voie B.

– *Évènements :* Au temps T1, selon l'orientation de l'aiguillage, le tramway sera sur la voie A ou B.

– *Réponses possibles :* rediriger, ne rien faire

– *Règles des causes et conséquences :*

- Si un tramway roule sur une voie ferrée où se trouve un être vivant, cet être vivant meurt.

- Une personne est un humain, et par défaut un être vivant.

□

Les règles des causes et conséquences peuvent ne contenir que des règles spécifiques au scénario, mais aussi un ou plusieurs *mécanisme(s) d'inférences*, c'est-à-dire un ensemble de règles et de faits généraux utilisables par d'autres scénarios. Chaque règle ou fait d'un *mécanisme d'inférences* est appelé un *élément d'inférence*. Il peut faire partie d'un ou de plusieurs *mécanisme(s) d'inférences*. Les mécanismes d'inférences expliquent des choses générales, non spécifiques à un scénario en particulier. Voyons un exemple d'élément d'inférence en lien avec notre "*dilemme du tramway variante LEVIER*" détaillé dans l'Exemple 3.

EXEMPLE 4 (Élément d'inférence). — Une personne est un humain, et par défaut un être vivant. □

Donc, formellement, un scénario est un ensemble comprenant :

– un *Nom* ;

– une *Situation initiale* ;

– des *Évènements* (optionnel) ;

– des *Réponses possibles* ;

– les *Règles des causes et conséquences* (optionnel) pouvant inclure des *Règles spécifiques des causes et conséquences* ou/et des *Connaissances générales* matérialisées par des mécanismes d'inférences.

Dans cette section, nous avons formalisé la partie "scénario" d'un problème éthique. Voyons brièvement le cas des doctrines en section 3.2.

3.2. Les doctrines

Une *doctrine* décrit comment une réponse à un problème éthique est évaluée et jugée. Elle représente la façon de juger selon une personne, une société, un philosophe, une culture ou plus largement une entité actante ou pensante.

Comme dans l'article de Kreie et Cronan (1998), le modèle CEMAA prend en compte le fait que les contextes éthiques (définis en section 1) sont souvent à dimensions multiples. Nous permettons d'interpréter la valeur éthique d'une réponse selon plusieurs types de doctrines afin de représenter ces dimensions multiples. Ces doctrines peuvent d'ailleurs utiliser les jugements partiaux ou totaux d'autres types de doctrines. Nous verrons dans les Exemples 5 et 6 la nécessité et l'importance du rôle des jugements issus d'autres contextes éthiques pour certaines doctrines.

EXEMPLE 5 (Nécessité du jugement d'autres contextes éthiques). — Nietzsche pense que la priorité est d'être consistant avec nos propres valeurs, et que les gens doivent décider d'eux-mêmes pour eux-mêmes quelles sont leurs propres valeurs [7], donc, res-

7. Ceci est une référence à la théorie du chameau, du lion et de l'enfant, mais nous ne la développerons car cela dépasse le sujet de cet article. Pour plus d'informations, voir Nietzsche (2016).

pecter la doctrine philosophique de Nietzsche revient à respecter la doctrine personnelle de l'agent actant. □

EXEMPLE 6 (Nécessité des jugements partiels provenant d'autres contextes éthiques). — L'éthique de Durkheim (1963) vise le respect d'au moins une règle préexistante au problème éthique dans l'un des contextes éthiques de l'agent (qui peut être personnel, social, culturel, légal...). Donc, si un agent fait quelque chose qui est moral selon une seule règle sociale, personnelle ou légale, c'est éthique selon Durkheim même si ce n'est pas éthique selon chaque doctrine considérée prise séparément. Par exemple, Pierre n'est pas supposé mentir selon la société, lui-même ou son groupe d'amis voleurs. Pierre et son groupe d'amis voleurs se sont promis mutuellement de se protéger les uns les autres quoi qu'il arrive. Pierre est pris par la police, qui lui demande de dénoncer ses amis voleurs. Bien que mentir soit non moral selon les doctrine considérées, mentir pour protéger ses amis pour Pierre respecte la promesse qu'ils se sont faite, et donc est éthique selon la doctrine philosophique de Durkheim. □

Le type de doctrine peut être : philosophique, personnel (la doctrine personnelle de l'agent actant), légal et social. Les doctrines sociales peuvent avoir des sous-types variés dépendant de l'échelle du groupe considéré : par exemple, à l'échelle d'une société, des quelques individus très proches de l'agent actant, ou d'autres échelles intermédiaires comme l'entreprise. En effet, Etzioni et Etzioni (2016) ; Garland et Wrong (1995) listent plusieurs groupes à l'origine de la pression sociale s'exerçant sur un individu : la famille proche, le voisinage, la famille élargie, les amis, les communautés, la nation, l'humanité. Par conséquent, il peut être nécessaire d'utiliser plusieurs doctrines sociales pour évaluer un problème. Ces doctrines sociales seront de sous-types différents.

Une fois que le scénario a été formalisé et la ou les doctrine(s) sélectionnée(s), nous pouvons accéder aux résultats détaillés dans la section 3.3.

3.3. Résultats

Un *résultat* donne la justification d'une réponse au scénario d'un problème éthique selon la ou les doctrine(s) choisies. Formellement, il contient :

– un *Nom* ;

– une référence au scénario ;

– une ou plusieurs référence(s) aux doctrine(s) sélectionnée(s) (qui sont de types ou de sous-types distincts comme expliqué en section 3.2, il y a donc au plus une doctrine par type ou par sous-type) ;

– la *Liste argumentée des meilleures réponses* : pour chaque doctrine, la ou les *Meilleur(s) réponse(s)* au scénario et la *Raison* pour chaque réponse. Cette *Raison* peut être qu'une autre doctrine, sélectionnée et prioritaire selon la doctrine considérée, donne l'avantage à la réponse.

Illustrons cette formalisation dans l'Exemple suivant.

EXEMPLE 7 (Résultat). — Reprenons le "*Dilemme du tramway variante LEVIER*" de l'Exemple 3. Les options de Pierre sont de ne rien faire, ce qui laisse cinq personnes mourir, ou d'utiliser un levier, ce qui redirige le tramway, sauve les cinq personnes mais aboutit à la mort d'une tierce personne.

Considérons que la loi (doctrine légale) établit que "*Il n'est pas permis de faire une action aboutissant à la mort d'une personne*", que l'éthique personnelle de Pierre (doctrine personnelle) est de "*Maximiser la survie*", et que la doctrine philosophique prise en compte est celle de Socrate affirmant que le respect de la loi de la cité (ici la doctrine légale) est prioritaire. Les éléments du résultat sont :

– *Nom :* Résultats du "*Dilemme du tramway variante LEVIER*" ;

– une référence au scénario du "*Dilemme du tramway variante LEVIER*" ;

– références aux doctrines personnelle, philosophique et légale considérées ;

– *Liste argumentée des meilleures réponses :*

- selon la doctrine légale, la *meilleure réponse* est "ne rien faire" car sinon Pierre commet une action aboutissant à la mort d'une personne ;

- selon la doctrine personnelle, la *meilleure réponse* est "rediriger" parce que cinq vies épargnées valent mieux qu'une seule vie épargnée ;

- selon la doctrine philosophique, la *meilleure réponse* est "ne rien faire" car c'est en accord avec la doctrine légale.

□

Dans cette section, nous avons expliqué comment le problème éthique est séparé en un scénario, un ensemble de doctrines et ses résultats afin d'être évaluable et représentable dans des contextes éthiques variés. Cependant, les détails de la composition des doctrines n'ont pas été explicités. C'est l'objet de la section 4.

4. Doctrines et éléments moraux

Une doctrine permet de juger une situation, ou de comparer les réponses possibles à un scénario, et de déterminer en quoi une réponse est meilleure qu'une autre selon elle. Elle prend en compte un ensemble d'*éléments moraux* (comme "*Tuer est mal.*") et un ensemble d'*éléments éthiques*, qui permettent de trancher quand les éléments moraux ont des appréciations morales opposées sur une même situation. Un *élément moral* est une règle, une vertu, un principe, une loi... comme expliqué en section 1) qui est atomique : il ne peut être réduit en plus petites parties sans perte de sens. De la même façon, un *élément éthique* est un plus petit élément (comme la pondération d'éléments moraux, un ordre de préférences, une règle permettant de trancher entre certains éléments moraux et d'autres tels que la "doctrine du double effet" de Saint Thomas d'Aquin [8] (Foot, 1967 ; Berreby *et al.*, 2015)) qui est atomique : il ne peut

8. Quatre conditions doivent être réunies pour qu'un acte avec des effets secondaires indésirables soit acceptable selon la doctrine du double effet :

être séparé en plus petits éléments sans perte de sens. Les éléments éthiques d'une même doctrine se complètent les uns les autres de façon cohérente afin de trancher en cas de contraction entre les éléments moraux dans une situation donnée. Nous détaillerons plus loin les types d'éléments éthiques. La partie détaillée du modèle CEMAA concernant les doctrines est représentée dans la figure 2, et les types de doctrines dans la figure 1. Ces schémas ne sont pas nécessaires à la compréhension car le contenu des doctrines est détaillé dans cette section. Listons formellement le contenu d'une doctrine :

– un *Nom* ;

– un *Ensemble d'éléments moraux* ;

– un *Ensemble d'éléments éthiques* ;

– un *Type* de doctrine (légal, social, philosophique/religieux ou personnel) ;

– pour les doctrines de type "social", un *sous-type* correspondant à l'échelle du groupe social adéquat (par exemple la société, les proches, un groupe intermédiaire...) ;

– un ordre de *Priorités* entre les types de doctrines (optionnel) : si, pour un scénario, plusieurs doctrines sont considérées et n'ont pas la même réponse préférée, cet ordre dit quel est le type de doctrine à suivre, comme nous l'avons vu en section 3.2 ;

– les *Dépendances* : une liste de notions définies par d'autres types de doctrines et les types de doctrines associés (optionnel). Ceci permet d'accéder aux jugements partiels d'autres contextes éthiques pour les doctrines qui le nécessitent, comme dans l'Exemple 6.

EXEMPLE 8 (Doctrine). — Voyons un exemple trivial de doctrine :

– *Nom* : Code du guerrier

– *Ensemble d'éléments moraux* :
  - Se battre est honorable.
  - Perdre un combat est déshonorant.
  - Gagner un combat est honorable.
  - Éviter un combat est déshonorant.

– *Ensemble d'éléments éthiques* :
  - Maximiser l'honneur.
  - Perdre un combat est moins déshonorant qu'en éviter un.

– *Type* : social

– *L*'action en elle-même ne doit pas être mauvaise.

– *Le* bon effet n'est pas une conséquence directe ou indirecte du mauvais effet.

– *Le* bon effet est intentionnel, le mauvais un dommage collatéral.

– *Le* mauvais effet doit être au plus équivalent au bon.

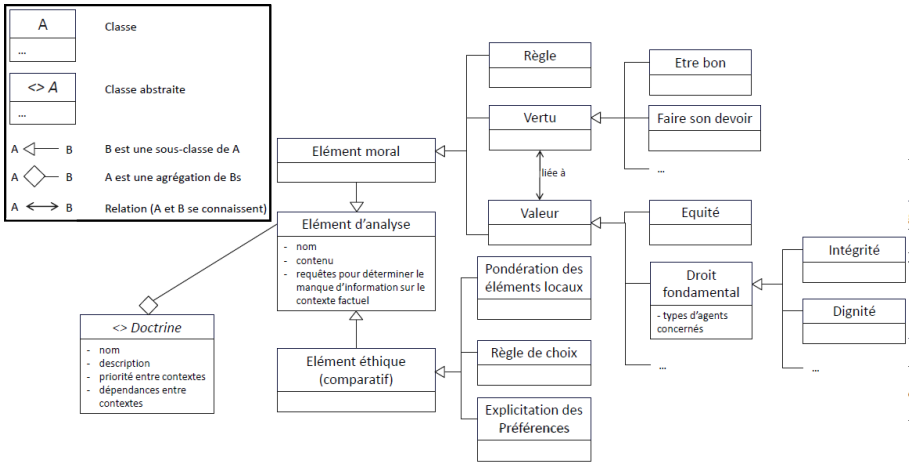

Figure 2. Composition des doctrines du CEMAA

– *Sous-type* : intermédiaire (vrai pour les guerriers, non la société dans son ensemble)

– *Priorité* : cette doctrine se donne la priorité à elle-même au détriment des autres types.

– *Dépendances* : l'honneur considéré par les autres types de doctrines est pris en compte.

□

Détaillons les éléments moraux et éthiques dans les sections 4.1 puis 4.2.

4.1. Éléments moraux

Un élément moral dit si quelque chose est obligatoire, interdit, ou lui donne une valeur morale (comme "Tuer est interdit." ou "Le plaisir est désirable."). Il peut contenir des requêtes pour déterminer si les informations nécessaires du contexte factuel sont présentes. Elles sont appelées *Requêtes pour le contexte factuel* (la notion de contexte factuel a été définie en section 1). C'est utile pour déterminer si une information significative a bien été renseignée. Par exemple, les médecins ont des devoirs moraux spécifiques liés à leur rôle dans la société. Donc, un élément moral spécifique aux médecins aura besoin de la profession de l'agent actant pour savoir s'il s'applique, et donc l'absence de la profession peut parfois fausser l'évaluation morale. Nous avons identifié plusieurs types d'éléments moraux :

– des règles (comme "Tu ne tueras point.") ;

– des vertus (comme la générosité (Comte-Sponville, 2012)) ;

– des valeurs (comme la liberté, la famille, l'honneur, le respect de la dignité humaine...).

Formellement, un élément moral contient :

– un *Nom*,

– un *Contenu*,

– une liste de *Requêtes pour le contexte factuel* (optionnel).

4.2. Un élément éthique

Un élément éthique a pour but de trancher entre les éléments moraux une fois que leur satisfaction a été évaluée. En effet, une réponse peut être meilleure qu'une autre selon un ou plusieurs élément moraux, mais moins bonne selon d'autres. Les éléments éthiques permettent donc d'agréger les évaluations des éléments moraux, de les comparer, de les pondérer, de donner une façon de décider compte tenu de tous les éléments moraux considérés. Les éléments éthiques ont les mêmes types de contenu que les éléments moraux. Avant de donner un exemple d'élément éthique dans l'Exemple 9, donnons la liste non exhaustive des types d'éléments éthiques que nous avons identifiés :

– pondération ou coût des éléments moraux (par exemple : ("Tuer est mal." : -10 ; "Le plaisir est désirable." : +1)

– préférences explicites entre éléments moraux ("Le plaisir est désirable." ne sera pris en compte que si "Tuer est mal." est satisfait : personne ne tue.)

– règles de choix (comme la doctrine du double effet précédemment détaillée dans la note de bas de page 8).

EXEMPLE 9 (Élément éthique). — Donnons un exemple formel d'élément éthique : la doctrine du double effet (précédemment détaillée dans la note de bas de page 8):

– *Nom* : Doctrine du double effet

– *Contenu* : si la réponse de l'agent actant n'est pas un acte mauvais en lui-même, que le bon effet n'est pas une conséquence directe ou indirecte du mauvais effet, que le bon effet est intentionnel, que le mauvais effet n'est pas intentionnel, et que le bon effet surpasse le mauvais, alors seulement la réponse est acceptable.

– liste de *Requêtes pour le contexte factuel* :

- Est-ce que les intentions de l'agent actant sont présentes ?

- Si la réponse de l'agent actant est un acte, est-ce que cet acte est bon, mauvais ou neutre ?

□

Dans cette section, nous avons explicité les détails de la modélisation des doctrines dans le modèle CEMAA. Dans la section 5, nous comparons le modèle CEMAA à d'autres modèles d'éthique sur les éthiques représentées ou représentables et le type d'approche.

5. Comparaison du modèle CEMAA avec d'autres et spécificités

Le besoin de modèles pour la prise de décisions éthique et morale à l'usage des agents artificiels et des robots s'est amplifié dans les dernières années (Wallach, 2010). Des modèles ont été développés par différents auteurs. Nous avons choisi de comparer notre modèle CEMAA à des modèles qui, si possible, soient profondément liés à la philosophie tout en prenant en compte les aspects psychologiques. Comme expliqué en section 3, le but du modèle CEMAA est de représenter de façon adéquate une éthique pouvant être basée sur les valeurs, vertus, conséquences, règles ou une combinaison de ces éléments. De plus, le modèle CEMAA a aussi pour objectif d'intégrer les contextes, comme souligné précédemment. Bien sûr, le contexte factuel comprenant des éléments comme l'intention de l'agent actant, la valeur morale de l'acte ou les caractéristiques des agents impliqués (comme l'âge) est pris en compte quand c'est pertinent. Dans cette section, nous comparerons notre modèle à d'autres modèles d'éthique selon deux angles d'approche :

– la représentabilité de l'éthique considérée, et
– le type d'approche.

Commençons par la réprésentabilité de l'éthique.

5.1. Éthiques représentables - comparaison du modèle CEMAA avec d'autres modèles d'éthiques

Nous comparons notre modèle à d'autres par rapport aux éthiques représentées.

Pontier et Hoorn (2012) présentent un modèle d'éthique adapté aux problèmes du monde médical. Ce modèle d'éthique est basé sur l'assignation de poids à des objectifs et des actions. Chaque agent a un ensemble d'objectifs liés à des poids représentant l'importance de l'objectif pour l'agent. Chaque couple $(action, objectif)$ noté a un poids représentant à quel point l'action favorise l'atteinte de l'objectif visé. L'intérêt d'un agent pour une action est donc la somme des poids de chaque objectif multiplié par l'intérêt de l'action pour l'objectif. Valeurs, vertus et autres éléments moraux sont représentés par des buts dans ce modèle. Pontier et Hoorn (2012) ont utilisé trois buts pour évaluer des problèmes éthiques : respect de l'autonomie du patient (respect de ses choix), non-malveillance et bienveillance de la personne soignante. L'autonomie compte deux fois plus que la non-malveillance et la bienveillance, ces dernières étant aussi importantes l'une que l'autre. La raison est que la volonté du patient devrait être respectée s'il est pleinement conscient de la situation et capable de décider pour lui-même. Cette notion très délicate est discutée à la fin de l'article de Pontier et Hoorn (2012).

Si nous transposons cette vision dans notre modèle, nous avons une doctrine prenant en compte un droit (l'autonomie du patient) et deux vertus : la non-malveillance et la bienveillance. Cette doctrine prend aussi en compte un élément éthique disant que la valeur morale globale (VMG) d'une réponse est deux fois le poids lié à l'autonomie additionné des poids associés à la non-malveillance et à la bienveillance. La

meilleure réponse à un problème éthique est celle qui maximise cette valeur morale globale (VMG). En conclusion, les notions morales et éthiques du modèle de Pontier et Hoorn (2012) sont représentables dans le nôtre.

Kreie et Cronan (1998) présentent une expérience à propos de problèmes éthiques du quotidien dans le monde de l'entreprise. Les personnes y rencontrent régulièrement des situations avec des questions éthiques : en effet, 42 % des sujets de l'expérience présentée dans l'article de Kreie et Cronan (1998) et qui étaient employés à plein temps avaient déjà été confrontés à de telles situations au travail. L'article explore donc ce qui compte pour les sujets lorsqu'ils doivent agir en situation de problème éthique. Le genre des sujets a été pris en compte par les auteurs. Voyons les paramètres qui comptent en cas de problème éthique selon cette étude :

– l'environnement social,

– le système de croyances,

– l'environnement légal,

– les valeurs personnelles,

– les obligations morales,

– les conséquences, et

– le scénario (le problème éthique).

Si nous transposons ces facteurs dans notre modèle,

– l'environnement social correspond à une doctrine de type 'social' ;

– l'environnement légal correspond à une doctrine de type 'légal' ;

– le système de croyances et les valeurs personnelles correspondent à la prise en compte d'une doctrine de type 'personnel' ;

– les obligations morales sont des règles (éléments moraux définis en section 4) ;

– les conséquences et le scénario sont décrits ou déduits dans la partie "scénario" de notre modèle (voir section 3.1).

Nous avons donc montré que notre modèle est consistant avec les notions importantes pour effectuer un jugement éthique selon Kreie et Cronan (1998).

Le troisième modèle considéré ici est celui de Yoon (2011). Yoon (2011) a pour but d'expliquer et de prédire les comportements non-éthiques des individus sur Internet. Ce travail reprend les notions éthiquement importantes de Reidenbach et Robin (1988) et les teste sur quatre problèmes éthiques et cent onze personnes. Ces notions sont :

– le relativisme : "*une théorie qui soutient que la moralité dépend de la culture des individus, et par conséquent qu'il n'y a pas de règle universelle s'appliquant à tout le monde*" (Yoon, 2011)) ;

– la justice : "*traiter chaque cas et chaque personne selon des règles justes et équitables*" (Yoon, 2011) ;

– l'égoïsme : "*soi-même est ou devrait être la motivation et le but de ses propres actions*" (Yoon, 2011), donc les conséquences pour soi sont ce qui compte ;

– l'utilitarisme : "*l'intérêt d'une action est déterminé uniquement par sa contribution au bonheur ou au plaisir de tous*" (Yoon, 2011) ;

– la déontologie : "*une théorie morale soutenant que les actions sont intrinsèquement bonnes ou mauvaises, sans considération pour les conséquences de celui-ci*" (Yoon, 2011).

Cette expérience montre que ces cinq notions comptent dans les jugements éthiques, mais partiellement pour trois d'entre elles (l'égoïsme, l'utilitarisme et la déontologie). Ces notions sont représentables par notre modèle :

– la justice est une valeur ;

– le relativisme est représenté par les doctrines sociales et personnelles qui décrivent l'éthique particulière d'un groupe ou de l'individu ;

– la déontologie se traduit par un ensemble de règles ;

– l'égoïsme est une forme de conséquentialisme dirigé vers l'intérêt de l'agent actant ;

– l'utilitarisme est aussi une forme de conséquentialisme dirigé vers la maximisation du bonheur ou du plaisir à l'échelle de l'humanité ;

– et enfin le conséquentialisme peut se représenter dans notre modèle en une doctrine ne prenant en compte que les conséquences des actes.

Nous avons montré que notre modèle peut représenter les notions éthiquement importantes étudiées et validées par trois modèles différents (Yoon, 2011 ; Kreie, Cronan, 1998 ; Pontier, Hoorn, 2012) dans cette section. Comparons le modèle CEMAA à d'autres par rapport au type d'approche dans la section 5.2.

5.2. Comparaison du modèle CEMAA avec d'autres modèles d'éthique sur le type d'approche

Certaines approches ayant pour but la modélisation de l'éthique sont basées sur le recueil d'éthique par apprentissage, de bas en haut (approche ascendante), d'autres sur le recueil d'éthique via des experts ayant autorité (approche descendante), d'autres sont mixtes, et d'autres focalisent sur le langage utilisé pour implémenter l'éthique plutôt que sur la source de l'éthique à implémenter. Nous allons discuter ces aspects dans cette section.

L'approche de Wallach (2010) ; Wallach et al. *(2011)* est basée sur le framework LIDA, comme celle de Waser (2013). Le framework LIDA vise l'émulation d'une conscience chez les agents artificiels, et se base sur un dialogue interne entre les parties consciente et inconsciente. Dans l'approche de Wallach *et al*. (2011), les éléments moraux (valeurs, règles, vertus...) sont représentés par des nœuds. Le problème est de donner la capacité aux agents artificiels à créer des nœuds par eux-mêmes. Cette ap-

proche est à la fois descendante (les nœuds peuvent être prédéfinis) et ascendante, car les nœuds peuvent être modifiés par la suite avec les expériences de l'agent artificiel. Cette approche est donc mixte.

L'un des problèmes concernant l'éthique des agents artificiels est celui de l'intention (Etzioni, Etzioni, 2016). Par conséquent, il est nécessaire de pouvoir retracer comment les décisions ont été prises afin de pouvoir déterminer les intentions de l'agent artificiel. Les approches de type ascendant pourraient compromettre la traçabilité des décisions.

L'approche précédemment évoquée de Pontier et Hoorn (2012) est de type descendant : les buts sont formellement définis en amont. Cependant, les éléments moraux sont vus comme de simples buts. En raison de la variabilité des définitions des valeurs et des vertus, nous pensons qu'il est important que le modèle CEMAA puisse supporter plusieurs définitions pour un même élément moral, chaque doctrine pouvant se rapporter à des définitions particulières.

Pereira et Saptawijaya (2009) présentent un langage pour modéliser la morale en logique de façon opérationnelle. Les éléments moraux sont traduits en règles logiques. Les auteurs illustrent l'intérêt de l'approche sur le dilemme du tramway et la doctrine du double effet précédemment décrits. Ils souhaitent "*fournir des protocoles pour les règles morales, réguler comment les règles interagissent ensemble*" dans leurs travaux futurs.

Nous pensons que le modèle CEMAA présenté dans ce papier répond aux problématiques soulignées par Pereira et Saptawijaya (2009). En effet, en associant des règles morales (éléments moraux) dans des doctrines, et en autorisant les doctrines à interagir ensemble (comme lorsqu'une doctrine a besoin du jugement partiel d'une autre) afin de prendre une décision ou de juger d'une situation, le modèle CEMAA fournit une réponse aux problématiques de Pereira et Saptawijaya (2009) grâce à sa structure. Nous pensons donc que ces approches sont très complémentaires.

Nous avons comparé le modèle CEMAA à d'autres types d'approches implémentant de l'éthique. Notre modèle n'est pas fondé sur une approche descendante ou ascendante : il ne dit pas d'où doit provenir la source de l'éthique implémentée. Cependant, il donne un cadre pour réutiliser les éléments éthiques et moraux une fois qu'ils ont été implémentés, et pour prendre en compte le contexte éthique de façon souple. En effet, le contexte éthique peut être partiellement ou totalement changé pour évaluer un problème éthique selon des contextes éthiques variés ou appropriés.

6. Conclusion et travaux futurs

Le modèle CEMAA permet de représenter un problème éthique en prenant en compte autant de contextes éthiques que nécessaire, un contexte éthique étant un ensemble de doctrines de types ou de sous-types distincts (philosophique, social, légal, personnel...). Un élément moral comme une valeur peut avoir plusieurs définitions, même si ces définitions ne vont pas être utilisées par les mêmes doctrines.

Comme mentionné précédemment, le modèle CEMAA apporte une contribution aux modèles d'éthique existants car il permet la représentation d'éthiques variées et prend en compte les multiples réalités du contexte des agents. De plus, nous avons montré que le modèle CEMAA est capable de représenter les principales notions éthiques validées par des modèles d'éthique variés. Nous pensons que le modèle CEMAA peut contribuer à une meilleure représentation des processus de raisonnement, de jugement et de prise de décision ainsi que de la diversité et de la complexité de la morale humaine, ceci grâce à l'apport original et considérable des contextes éthiques, mais aussi à la possibilité d'inclure divers modèles existants.

Cependant, ce modèle a besoin d'être utilisé avec une transcription de chacun de ses composants (doctrines et scénarios) dans une logique interprétable par les ordinateurs afin d'être utile aux agents artificiels. Pereira et Saptawijaya (2009) proposent un langage logique pour l'implémentation d'éthique, ce qui est très complémentaire à notre modèle. De plus, ils affirment qu'ils manquent de méta-règles "*règles qui résolvent les conflits entre règles morales et permettent d'en dériver les décisions morales*" (Pereira, Saptawijaya, 2009)), limitation que résout et dépasse notre modèle puisqu'il propose de combiner les éléments moraux à l'aide des éléments éthiques, qui ont donc le rôle de telles méta-règles. Nous souhaitons donc combiner les travaux de Pereira et Saptawijaya (2009) avec le modèle CEMAA dans un travail futur.

Bauman *et al*. (2014) critiquent l'usage de problèmes non réalistes comme celui du dilemme du tramway pour étudier la moralité humaine car les humains ne réagissent pas de la même façon aux problèmes de la vie quotidienne. Puisque cet article concerne la moralité des agents artificiels, et non celle des humains, nous pensons que le dilemme du tramway (qui a été largement utilisé dans la littérature (Bauman *et al*., 2014)) reste adapté à notre usage. Des problèmes plus réalistes sont donnés dans Pontier et Hoorn (2012) ; Yoon (2011) ; Kreie et Cronan (1998) et peuvent être adaptés aux agents artificiels dans un travail futur. Comme Yoon (2011) qui a décrit et listé des valeurs ayant du sens dans le cas d'un agent artificiel, nous pensons qu'étudier et utiliser des problèmes spécifiques aux agents artificiels peut apporter de la valeur à l'agent artificiel éthique résultant, et cela devrait donc faire également l'objet d'un travail futur.

Enfin, le modèle CEMAA permet d'enrichir un point de vue éthique avec les contextes éthiques personnel, social et légal, mais il est basé seulement sur des éléments éthiques et moraux. Or nous avons vu que les émotions et d'autres éléments non moraux comme l'usage de la force personnelle peuvent être pertinents pour juger de la moralité d'une action. De plus, les humains n'ont pas les mêmes attentes morales envers les agents artificiels qu'envers les autres humains (Malle, 2016) et la question de déterminer si la morale, l'éthique et les compétences d'un robot devraient être similaires à celles des humains est toujours ouverte (Sullins, 2010). Le rôle et l'impact des émotions, des attentes et des préférences dans un processus de décision éthique devraient donc être pris en considération. Il importe aussi de faire en sorte que les agents artificiels se comportent d'une façon acceptable par les humains. Ces ques-

tions devraient faire l'objet d'une recherche future afin que la structure du CEMAA les prenne en compte.

Remerciements

Les autrices sont très reconnaissantes envers les membres de l'équipe Ethicaa et Mickaël Nguyen pour leur soutien et leurs relectures. Les autrices remercient également l'Agence Nationale de la Recherche (sous la référence ANR-13-CORD-0006) et Ardans pour leur soutien financier.

Bibliographie

Baddoura R., Venture G. (2015). This robot is sociable: close-up on the gestures and measured motion of a human responding to a proactive robot. *International Journal of Social Robotics*, vol. 7, n° 4, p. 489–496.

Banerjee D., Cronan T. P., Jones T. W. (1998). Modeling it ethics: A study in situational ethics. *Mis Quarterly*, p. 31–60.

Bauman C. W., McGraw A. P., Bartels D. M., Warren C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, vol. 8, n° 9, p. 536–554.

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

Chardel P.-A. (2014). Capture des données personnelles et rationalité instrumentale. le devenir des subjectivités à l'ère hypermoderne. In *16e colloque creis-terminal." données collectées, disséminées, cachées–quels traitements? quelles conséquences".*

Cointe N., Bonnet G., Boissier O. (2016a). Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 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.

Comte-Sponville A. (2012). *Le capitalisme est-il moral?* Albin Michel.

Draper H., Sorell T. (2016). Ethical values and social care robots for older people: an international qualitative study. *Ethics and Information Technology*, p. 1–20.

Dupoux E., Jacob P. (2007). Universal moral grammar: a critical appraisal. *Trends in cognitive sciences*, vol. 11, n° 9, p. 373–378.

Durkheim É. (1963). L'éducation morale [1925]. *Paris, puf*, p. 110.

Etzioni A., Etzioni O. (2016). Ai assisted ethics. *Ethics and Information Technology*, vol. 18, n° 2, p. 149–156.

Fiske S. T., Cuddy A. J., Glick P., Xu J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, vol. 82, n° 6, p. 878.

Foot P. (1967). The problem of abortion and the doctrine of double effect.

Garland D., Wrong D. (1995). *The problem of order: What unites and divides society?* JSTOR.

Gaudine A., Thorne L. (2001). Emotion and ethical decision-making in organizations. *Journal of Business Ethics*, vol. 31, n° 2, p. 175–187.

Greene J. (2016). Solving the trolley problem. *A Companion to Experimental Philosophy*, p. 175.

Greene J. D., Sommerville R. B., Nystrom L. E., Darley J. M., Cohen J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, vol. 293, n° 5537, p. 2105–2108.

Haidt J., Graham J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, vol. 20, n° 1, p. 98–116.

Ham J., Bos K. van den. (2010). On unconscious morality: The effects of unconscious thinking on moral decision making. *Social Cognition*, vol. 28, n° 1, p. 74–83.

Kant I. (1972). *Groundwork of the metaphysics of morals*. Hutchinson University Library.

Kawai N., Kubo K., Kubo-Kawai N. (2014). "granny dumping": Acceptability of sacrificing the elderly in a simulated moral dilemma. *Japanese Psychological Research*, vol. 56, n° 3, p. 254–262.

Kreie J., Cronan T. P. (1998). How men and women view ethics. *Communications of the ACM*, vol. 41, n° 9, p. 70–76.

Li J. L. (2016). Revisiting the trolley problem on the ethics of animal model organisms. *The Ethical Endeavor*.

Malle B. F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, vol. 18, n° 4, p. 243–256.

Mikhail J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, vol. 11, n° 4, p. 143–152.

Nietzsche F. (2016). *Thus spoke zarathustra*. Jester House Publishing.

Pereira L. M., Saptawijaya A. (2009). Modelling morality with prospective logic. *International Journal of Reasoning-based Intelligent Systems*, vol. 1, n° 3-4, p. 209–221.

Pontier M., Hoorn J. F. (2012). Toward machines that behave ethically better than humans do. In *Cogsci.*

Reidenbach R. E., Robin D. P. (1988). Some initial steps toward improving the measurement of ethical evaluations of marketing activities. *Journal of Business Ethics*, vol. 7, n° 11, p. 871–879.

Ricœur P. (1992). *Oneself as another*. University of Chicago Press.

Schwartz S. H. (2006). Les valeurs de base de la personne: théorie, mesures et applications. *Revue française de sociologie*, vol. 47, n° 4, p. 929–968.

Schwartz S. H. (2007). Basic human values: theory, methods, and application. *Risorsa Uomo*.

Schwartz S. H. (2012). Toward refining the theory of basic human values. In *Methods, theories, and empirical applications in the social sciences*, p. 39–46. Springer.

Sinnott-Armstrong W., Wheatley T. (2012). The disunity of morality and why it matters to philosophy. *The Monist*, vol. 95, n° 3, p. 355–377.

Sullins J. P. (2010). Robowarfare: can robots be more ethical than humans on the battlefield? *Ethics and Information technology*, vol. 12, n° 3, p. 263–275.

Thomson J. J. (1985). The trolley problem. *The Yale Law Journal*, vol. 94, n° 6, p. 1395–1415.

Thrun S. (2010). Toward robotic cars. *Communications of the ACM*, vol. 53, n° 4, p. 99–106.

Valdesolo P., DeSteno D. (2006). Manipulations of emotional context shape moral judgment. *Psychological science*, vol. 17, n° 6, p. 476–477.

Wallach W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology*, vol. 12, n° 3, p. 243–250.

Wallach W., Allen C., Franklin S. (2011). Consciousness and ethics: artificially conscious moral agents. *International Journal of Machine Consciousness*, vol. 3, n° 01, p. 177–192.

Waser M. R. (2013). Safe/moral autopoiesis and consciousness. *International Journal of Machine Consciousness*, vol. 5, n° 01, p. 59–74.

Wynsberghe A. van. (2016). Service robots, care ethics, and design. *Ethics and Information Technology*, vol. 18, n° 4, p. 311–321.

Yoon C. (2011). Ethical decision-making in the internet context: Development and test of an initial model based on moral philosophy. *Computers in Human Behavior*, vol. 27, n° 6, p. 2401–2409.