HAL
open science

# Jugement éthique pour la décision et la coopération dans les systèmes multi-agents

Nicolas Cointe

## To cite this version:

Nicolas Cointe. Jugement éthique pour la décision et la coopération dans les systèmes multi-agents. Autre. Université de Lyon, 2017. Français. NNT : 2017LYSEM043 . tel-01851485

**HAL Id: tel-01851485**

**https://theses.hal.science/tel-01851485**

Submitted on 30 Jul 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°d'ordre NNT : 2017LYSEM043

**THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON**
opérée au sein de
**l'École des Mines de Saint-Étienne**

**École Doctorale** N° 488
**Sciences, Ingénierie, Santé**

**Spécialité de doctorat** : informatique
**Discipline** : intelligence artificielle

Soutenue publiquement le 18 décembre 2017, par :
**Nicolas Cointe**

---

# Jugement éthique pour la décision et la coopération dans les systèmes multi-agents

---

Devant le jury composé de :

| | | |
|---|---|---|
| OCCELLO, Michel | Professeur, Université Grenoble Alpes, Valence | Président |
| CHATILA, Raja | Professeur, Université Pierre et Marie Curie, Paris | Rapporteur |
| SABOURET, Nicolas | Professeur, Université Paris-sud, Orsay | Rapporteur |
| TESSIER, Catherine | Maître de Recherche, ONERA, Toulouse | Examinatrice |
| VILLATA, Serena | Chargée de Recherche, INRIA, Sophia Antipolis | Examinatrice |
| BOISSIER, Olivier | Professeur, École des Mines de Saint-Étienne | Directeur de thèse |
| BONNET, Grégory | Maître de conférences, Université Caen-Normandie | Co-directeur de thèse |

| Spécialités doctorales | Responsables : | Spécialités doctorales | Responsables |
|---|---|---|---|
| SCIENCES ET GENIE DES MATERIAUX | K. Wolski Directeur de recherche | MATHEMATIQUES APPLIQUEES | O. Roustant, Maître-assistant |
| MECANIQUE ET INGENIERIE | S. Drapier, professeur | INFORMATIQUE | O. Boissier, Professeur |
| GENIE DES PROCEDES | F. Gruy, Maître de recherche | SCIENCES DES IMAGES ET DES FORMES | JC. Pinoli, Professeur |
| SCIENCES DE LA TERRE | B. Guy, Directeur de recherche | GENIE INDUSTRIEL | N. Absi, Maitre de recherche |
| SCIENCES ET GENIE DE L'ENVIRONNEMENT | D. Graillot, Directeur de recherche | MICROELECTRONIQUE | Ph. Lalevée, Professeur |

## EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

| | | | | |
|---|---|---|---|---|
| ABSI | Nabil | MR | Génie industriel | CMP |
| AUGUSTO | Vincent | CR | Image, Vision, Signal | CIS |
| AVRIL | Stéphane | PR2 | Mécanique et ingénierie | CIS |
| BADEL | Pierre | MA(MDC) | Mécanique et ingénierie | CIS |
| BALBO | Flavien | PR2 | Informatique | FAYOL |
| BASSEREAU | Jean-François | PR | Sciences et génie des matériaux | SMS |
| BATTON-HUBERT | Mireille | PR2 | Sciences et génie de l'environnement | FAYOL |
| BEIGBEDER | Michel | MA(MDC) | Informatique | FAYOL |
| BLAYAC | Sylvain | MA(MDC) | Microélectronique | CMP |
| BOISSIER | Olivier | PR1 | Informatique | FAYOL |
| BONNEFOY | Olivier | MA(MDC) | Génie des Procédés | SPIN |
| BORBELY | Andras | MR(DR2) | Sciences et génie des matériaux | SMS |
| BOUCHER | Xavier | PR2 | Génie Industriel | FAYOL |
| BRODHAG | Christian | DR | Sciences et génie de l'environnement | FAYOL |
| BRUCHON | Julien | MA(MDC) | Mécanique et ingénierie | SMS |
| CAMEIRAO | Ana | MA(MDC) | Génie des Procédés | SPIN |
| CHRISTIEN | Frédéric | PR | Science et génie des matériaux | SMS |
| DAUZERE-PERES | Stéphane | PR1 | Génie Industriel | CMP |
| DEBAYLE | Johan | MR | Sciences des Images et des Formes | SPIN |
| DEGEORGE | Jean-Michel | MA(MDC) | Génie industriel | Fayol |
| DELAFOSSE | David | PR0 | Sciences et génie des matériaux | SMS |
| DELORME | Xavier | MA(MDC) | Génie industriel | FAYOL |
| DESRAYAUD | Christophe | PR1 | Mécanique et ingénierie | SMS |
| DJENIZIAN | Thierry | PR | Science et génie des matériaux | CMP |
| DOUCE | Sandrine | PR2 | Sciences de gestion | FAYOL |
| DRAPIER | Sylvain | PR1 | Mécanique et ingénierie | SMS |
| FAUCHEU | Jenny | MA(MDC) | Sciences et génie des matériaux | SMS |
| FAVERGEON | Loïc | CR | Génie des Procédés | SPIN |
| FEILLET | Dominique | PR1 | Génie Industriel | CMP |
| FOREST | Valérie | MA(MDC) | Génie des Procédés | CIS |
| FRACZKIEWICZ | Anna | DR | Sciences et génie des matériaux | SMS |
| GARCIA | Daniel | MR(DR2) | Sciences de la Terre | SPIN |
| GAVET | Yann | MA(MDC) | Sciences des Images et des Formes | SPIN |
| GERINGER | Jean | MA(MDC) | Sciences et génie des matériaux | CIS |
| GOEURIOT | Dominique | DR | Sciences et génie des matériaux | SMS |
| GONDRAN | Natacha | MA(MDC) | Sciences et génie de l'environnement | FAYOL |
| GONZALEZ FELIU | Jesus | MA(MDC) | Sciences économiques | FAYOL |
| GRAILLOT | Didier | DR | Sciences et génie de l'environnement | SPIN |
| GROSSEAU | Philippe | DR | Génie des Procédés | SPIN |
| GRUY | Frédéric | PR1 | Génie des Procédés | SPIN |
| GUY | Bernard | DR | Sciences de la Terre | SPIN |
| HAN | Woo-Suck | MR | Mécanique et ingénierie | SMS |
| HERRI | Jean Michel | PR1 | Génie des Procédés | SPIN |
| KERMOUCHE | Guillaume | PR2 | Mécanique et Ingénierie | SMS |
| KLOCKER | Helmut | DR | Sciences et génie des matériaux | SMS |
| LAFOREST | Valérie | MR(DR2) | Sciences et génie de l'environnement | FAYOL |
| LERICHE | Rodolphe | CR | Mécanique et ingénierie | FAYOL |
| MALLIARAS | Georges | PR1 | Microélectronique | CMP |
| MOLIMARD | Jérôme | PR2 | Mécanique et ingénierie | CIS |
| MOUTTE | Jacques | CR | Génie des Procédés | SPIN |
| NEUBERT | Gilles | | | FAYOL |
| NIKOLOVSKI | Jean-Pierre | Ingénieur de recherche | Mécanique et ingénierie | CMP |
| NORTIER | Patrice | PR1 | Génie des Procédés | SPIN |
| O CONNOR | Rodney Philip | MA(MDC) | Microélectronique | CMP |
| OWENS | Rosin | MA(MDC) | Microélectronique | CMP |
| PERES | Véronique | MR | Génie des Procédés | SPIN |
| PICARD | Gauthier | MA(MDC) | Informatique | FAYOL |
| PIJOLAT | Christophe | PR0 | Génie des Procédés | SPIN |
| PINOLI | Jean Charles | PR0 | Sciences des Images et des Formes | SPIN |
| POURCHEZ | Jérémy | MR | Génie des Procédés | CIS |
| ROUSSY | Agnès | MA(MDC) | Microélectronique | CMP |
| ROUSTANT | Olivier | MA(MDC) | Mathématiques appliquées | FAYOL |
| SANAUR | Sébastien | MA(MDC) | Microélectronique | CMP |
| STOLARZ | Jacques | CR | Sciences et génie des matériaux | SMS |
| TRIA | Assia | Ingénieur de recherche | Microélectronique | CMP |
| VALDIVIESO | François | PR2 | Sciences et génie des matériaux | SMS |
| VIRICELLE | Jean Paul | DR | Génie des Procédés | SPIN |
| WOLSKI | Krzystof | DR | Sciences et génie des matériaux | SMS |
| XIE | Xiaolan | PR0 | Génie industriel | CIS |
| YUGMA | Gallian | CR | Génie industriel | CMP |

Mise à jour : 03/02/2017

*À ma Solène*

# Remerciements

Je tiens à remercier ici les personnes qui, au long de ces années de travail, m'ont accompagné tant dans ma démarche de recherche que dans ma vie personnelle.

Je souhaite remercier Olivier qui a démontré empiriquement l'existence des directeurs de thèses impliqués et constructifs, ce qui m'a valu de la part de quelques doctorants externes au laboratoire l'expression de leur jalousie. La qualité de cette thèse aurait assurément été amoindrie sans ses questionnements méthodologiques et sa grande expertise scientifique. Il serait injuste en outre de réduire son mérite à ses seules qualités d'encadrant et omettre de saluer son humanisme, sa bienveillance et son humilité.

Je remercie également Grégory qui m'a accompagné durant ces années malgré l'inconfort imposé par nos communications à distance. La régularité et la qualité de nos échanges ont constitué une aide précieuse pour maintenir un cap dans l'exploration d'un sujet aussi passionnant que celui abordé ici. Je tiens particulièrement à saluer sa capacité à mettre nos travaux en perspective simultanément du point de vue d'une multitude de paradigmes philosophiques et de disciplines scientifiques.

Plus largement je souhaite remercier la totalité des membres du projet EthicAa pour les discussions passionnantes et inspirantes qui ont animé nos réunions. Les premières difficultés liées aux différences méthodologiques entre nos disciplines ont rapidement cédé la place à des réflexions dans lesquelles la sémantique des concepts employés était enrichie de nos compréhensions diverses du sujet.

Je tiens à remercier les membres du jury pour avoir consacré le temps nécessaire à l'évaluation de mon travail. La pertinence de leurs questions et la qualité de leurs remarques m'ont permis d'accroître sensiblement la qualité de la version finale de ce document.

Pour leur accueil dans cette métropole stéphanoise, dont j'ignorais les us et coutumes et espère ne pas avoir pris l'accent, je souhaite remercier l'ensemble de mes collègues de l'École des Mines et du laboratoire Hubert Curien. Plus particulièrement, je souhaite mentionner ceux qui m'ont supporté au quatrième étage de l'institut Henri Fayol et que je préfère citer ici dans l'ordre très arbitraire de leur numéro de bureau au jour de ma soutenance en ne mentionnant qu'une partie infime de leurs mérites. Je remercie donc Amro pour mon initiation aux arcanes de la politique moyen-orientale, Antoine pour nos argumentations bancales sur l'interprétation de la théorie de la relativité générale,

Khadim pour avoir été la victime régulière et consentante de mon humour douteux, Noor pour nos riches débats sur la compatibilité de la conduite de véhicules et la consommation du chanvre indien sur les îles de l'océan éponyme, Roland pour nos discussions en patientant devant l'imprimante ou la porte de l'ascenseur, Flavien qui a su me comprendre sans même avoir besoin de mots lorsque je frappais à sa porte le jeudi matin, Gauthier avec qui j'ai couru des kilomètres par pur plaisir ou pour échapper à des organismes parfaits, Maxime pour avoir revisité l'esthétique de mon faciès dans mon sommeil malgré ses modestes talents de dessinateur, Mihaela pour ne pas s'être trop moquée de mon ignorance de la faune, Bruno pour son accueil des doctorants et son souci de leur bien-être au laboratoire, Marie Line pour avoir tenté d'achever mon éducation à grands renforts de tartes aux pralines, Nilou pour m'avoir accompagné jusqu'en haut des monts du lyonnais et m'avoir récupéré dans un sac poubelle une quarantaine de kilomètres plus loin, Jean François pour m'avoir initié à pacman et Archlinux, Oudom pour avoir survécu à trois ans dans le même bureau que moi et avoir toujours sportivement répondu à mes facéties, Philippe pour m'avoir appris à observer le grand lamas d'Hercule, Xavier pour m'avoir inspiré nombre de nouveaux projets qui ont rempli mes tiroirs et rentabilisé mon compte Amazon premium et Michel pour ses précieuses leçons d'ergonomie du poste de travail qu'il a apprises à ses dépens.

Je tiens à ajouter Agnès à l'équipe en la remerciant pour m'avoir fait découvrir le réseau industriel local, les crêpes de MiXiT, l'art de décoller des post-it par le côté et la méthodologie de travail qui va avec.

Je remercie également tous ceux qui ont séjourné au laboratoire durant ma thèse. Merci tout particulièrement à Jomi pour son aide précieuse dans l'utilisation de JaCaMo afin d'implémenter mes preuves de concept. Je salue également Amélie et son goût pour l'exploration des frontières de l'esthétique en CSS, Andrei pour ses précieux conseils sur ma fin de thèse, Chadha pour avoir apporté un peu de féminité à nos soirées de doctorants informaticiens, Lamia pour m'avoir fait découvrir duolingo, Gustavo pour avoir vécu avec moi quelques péripéties à Singapour et Jingyu pour m'avoir appris à rouler des makis presque corrects et pour son accueil chaleureux à Yokohama.

Je remercie également l'ensemble du département informatique du GREYC. Mes passages à l'université de Caen ont constitué pour moi autant d'occasions de prendre du recul sur ma démarche, profiter de compétences complémentaires et m'exercer aux mots fléchés et à la conquête de galaxies. Merci tout particulièrement à Bruno et Alexandre qui m'ont accordé une place entre leurs écrans respectifs.

Je tiens à m'excuser auprès de tous ceux qui me sont chers pour mon manque de disponibilité à leurs côtés durant ces quelques années et l'impact négatif qu'aurait eu ce travail sur mon caractère déjà peu commun. Je remercie tout particulièrement ma Solène pour tout le réconfort qu'elle m'a apporté. Je remercie également l'ensemble de mes amis qui ont considéré que nos liens méritaient de survivre à cet exil. Dans un ordre aléatoire (généré par un programme disponible sur mon serveur pour éviter toute jalousie), je remercie donc Fanouille, Tiff, Knarf, Gortrik, Snake, Toinou, Piba, BBJ et

Seb. Je remercie également ma famille et ma belle-famille pour m'avoir toujours accueilli à bras ouverts lors de mes rares retours en Normandie.

Je remercie chaleureusement mes scouts pour les camps inoubliables que nous avons vécus en bravant les catastrophes naturelles à 15 000 ou en arpentant les Pays-Bas à vélo. Merci pour l'aide que vous m'avez apportée pour la préparation des camps, et pour l'intensité de ces moments de divertissement si précieux. Je remercie évidement les autres chefs et l'ensemble de l'équipe de groupe et de territoire pour toutes ces aventures vécues avec intensité et sérénité.

Pour avoir joué un rôle décisif à leur insu dans mon épanouissement, je souhaite aussi remercier ceux qui m'ont initié au monde de la recherche scientifique et de la philosophie. Je remercie à ce titre madame Farias Endelin, enseignante de philosophie au Lycée Thomas Hélye de Cherbourg, pour m'avoir donné goût à cette discipline et avoir ainsi manifestement influencé mon choix de sujet de thèse. Je remercie également monsieur Babigeon, et monsieur Haquin, qui m'ont respectivement accueilli dans leurs laboratoires pour mes stages de DUT et de licence et m'ont montré que la recherche scientifique était le lieu idéal pour déchaîner mon insatiable curiosité.

Je remercie sincèrement l'Agence Nationale de la Recherche pour son financement. Puisse-t-elle encore longtemps financer des doctorants dans d'aussi bonnes conditions.

Enfin je tiens à te remercier, lectrice ou lecteur, pour ta patience et ton intérêt pour ce travail. Il m'aurait été pénible d'écrire ces pages pour qu'elles tombent dans l'oubli sitôt ma thèse soutenue.

Je présente toutes mes excuses pour toutes celles et ceux que j'ai oublié de mentionner ici et qui auraient mérité d'y figurer. Merci par avance pour votre indulgence.

Bonne lecture

# Avant-propos

Cette thèse s'inscrit dans le cadre du projet EthicAa (pour « Ethique et Agents Autonomes ») [1], financé par l'Agence Nationale pour la Recherche sous la référence ANR-13-CORD-0006. Ce projet poursuit plusieurs objectifs que sont la définition de ce que doivent être des agents autonomes éthiques et des systèmes multi-agents éthiques, la définition et la proposition de solutions pour les problèmes éthiques pouvant apparaître chez ces agents. Ce projet a permis de réunir des partenaires académiques et industriels, issus du domaine de l'Intelligence Artificielle ainsi que des Sciences Humaines et Sociales.

Cette thèse en intelligence artificielle présente des modèles afin de permettre à des agents autonomes de raisonner sur des concepts de morale et d'éthique. L'une des principales difficultés rencontrées durant cette démarche de recherche est due aux importants écarts méthodologiques entre l'intelligence artificielle et les disciplines liées à la philosophie morale. En effet, si la première de ces disciplines s'appuie aussi souvent que possible sur des concepts formellement définis, la seconde est constituée d'approches et de courants pouvant s'opposer ou se compléter et reposant sur des concepts qui ne sont pas explicitement définis. En raison de l'absence de définitions courtes, formelles et consensuelles dans la littérature philosophique, il a donc été nécessaire pour élaborer l'état de l'art de cette thèse d'effectuer un long travail de synthèse. Les définitions employées ici sont proposées avec le souci de faciliter la compréhension pour les spécialistes de l'intelligence artificielle et de structurer notre propos tout en permettant de représenter autant d'approches de l'éthique que possible.

Le lecteur qui ne serait pas familier de la littérature de sciences humaines pourrait être surpris par quelques aspects méthodologiques tel que l'emploi de dates d'éditions récentes pour des auteurs anciens (par exemple, (Platon, 1966) ne signifie pas que Platon a écrit son ouvrage en 1966, mais que l'on se réfère pour plus de commodité à une édition plus récente).

Enfin, il est nécessaire de préciser que ces travaux ne proposent aucun apport en philosophie et restent pleinement dans le domaine de l'intelligence artificielle en proposant des modèles appuyés sur des propositions antérieures et évalués par une démarche expérimentale. La problématique philosophique de la définition de ce que devraient être la morale et l'éthique d'un agent autonome n'est pas traitée ici.

1. Voir http://ethicaa.org/ pour plus d'informations.

# Sommaire

Introduction . . . . . 1
I L'éthique et la morale . . . . . 5
II Approches existantes . . . . . 27
III Modèle de jugement des actions . . . . . 51
IV Jugement des autres. . . . . 79
V Mise en œuvre et expérimentations . . . . . 99
Bilan et perspectives . . . . . 125
Bibliographie . . . . . 131

# Table des matières

Remerciements . . . iii
Avant-propos . . . vii
Sommaire . . . ix
Table des matières . . . xiv
Table des figures . . . xvi
Liste des tableaux . . . xvii
Introduction . . . 1
I L'éthique et la morale . . . 5
I.1 En philosophie . . . 6
I.1.1 La morale . . . 6
I.1.1.1 Les règles morales . . . 8
I.1.1.2 Les valeurs . . . 9
I.1.1.3 Les dilemmes moraux . . . 10
I.1.2 L'éthique . . . 11
I.1.2.1 Les principes éthiques . . . 12
I.1.2.2 Éthique des vertus . . . 13
I.1.2.3 Éthique déontologique . . . 13
I.1.2.4 Éthique conséquentialiste . . . 14
I.1.2.5 Dilemmes éthiques . . . 15
I.1.2.6 Le jugement . . . 15
I.2 En neurologie, psychologie et sciences cognitives . . . 17
I.2.1 Organisation de la cognition éthique . . . 17
I.2.2 Rôle des émotions et du raisonnement . . . 19
I.2.3 Facultés sociales et théorie de l'esprit . . . 20
I.3 En sciences sociales . . . 21
I.3.1 Théorie du développement moral . . . 21

I.3.2 Les valeurs morales en sociologie . . . 22
I.4 Synthèse . . . 24
II Approches existantes . . . 27
II.1 Problèmes éthiques dans les systèmes multi-agents . . . 28
II.1.1 Le jugement éthique pour la décision . . . 28
II.1.2 Influence de l'éthique sur les interactions entre agents . . . 29
II.1.2.1 Jugement du comportement des autres . . . 29
II.1.2.2 Usage du jugement pour la coopération . . . 30
II.2 Grille de lecture . . . 31
II.2.1 Représentations implicites ou explicites des éthiques . . . 31
II.2.2 Approche rationaliste ou intuitionniste . . . 32
II.2.3 Généricité . . . 32
II.2.4 Caractère opérationnel . . . 33
II.3 Approches procédurales . . . 33
II.3.1 Ethical governor . . . 33
II.3.2 Value sensitive design . . . 35
II.4 Approches numériques . . . 36
II.4.1 Case Supported Principle-Based Paradigm . . . 36
II.4.2 Jugement par évaluation d'expressions . . . 38
II.5 Approches déclaratives . . . 39
II.5.1 Représentation logique de principes éthiques . . . 39
II.5.2 Morale et logique déontique . . . 40
II.5.3 Responsabilité morale, éthique et causale . . . 41
II.5.4 Jugement dans une architecture BDI . . . 42
II.5.5 Argumentation formelle avec des valeurs morales . . . 45
II.6 Synthèse . . . 46
III Modèle de jugement des actions . . . 51
III.1 Préambule . . . 52
III.1.1 Fondements . . . 52
III.1.2 Scénario illustratif . . . 55
III.2 Présentation globale du modèle de jugement . . . 57
III.3 Reconnaissance de situation . . . 59
III.4 Évaluation de la possibilité et de la désirabilité des actions . . . 60
III.4.1 Évaluation de la possibilité . . . 60
III.4.2 Évaluation de la désirabilité . . . 62

III.5 Évaluation de la moralité des actions . . . 63
III.5.1 Système de valeurs morales . . . 64
III.5.2 Supports de valeurs . . . 65
III.5.3 Évaluation des supports de valeurs . . . 66
III.5.4 Règles morales . . . 67
III.5.5 Évaluation de la moralité . . . 69
III.6 Évaluation de l'éthique des actions . . . 70
III.6.1 Principes éthiques . . . 71
III.6.2 Fonction d'évaluation de l'éthique . . . 72
III.6.3 Préférences éthiques . . . 73
III.6.4 Fonction de jugement . . . 73
III.7 Exemple récapitulatif . . . 74
III.8 Synthèse . . . 76
IV Jugement des autres . . . 79
IV.1 Préambule . . . 80
IV.1.1 Notations . . . 80
IV.1.2 Scénario illustratif . . . 80
IV.2 Typologie des jugements . . . 81
IV.2.1 Niveau d'information du jugement . . . 81
IV.2.2 Temporalité du jugement . . . 83
IV.3 Images de la moralité et de l'éthique d'un agent . . . 84
IV.3.1 Image de la moralité des actions d'un agent . . . 84
IV.3.2 Image de l'éthique des actions d'un agent . . . 88
IV.4 Coopération fondée sur le jugement des autres . . . 92
IV.4.1 Construction de la confiance . . . 92
IV.4.2 Éthique de la confiance . . . 93
IV.4.3 Utilisation de la confiance pour la coopération éthique . . . 94
IV.5 Synthèse . . . 96
V Mise en œuvre et expérimentations . . . 99
V.1 Implémentation du modèle de jugement . . . 100
V.1.1 Le framework JaCaMo . . . 100
V.1.2 Reconnaissance de situation . . . 102
V.1.3 Évaluation de la désirabilité et de la possibilité des actions . . . 103
V.1.4 Moralité des actions . . . 104
V.1.5 Évaluation de l'éthique des actions . . . 106

V.1.6 Jugement des autres agents . . . . . 109
V.2 Application à la gestion d'actifs financiers . . . . . 110
V.2.1 Les agents autonomes sur les marchés financiers . . . . . 111
V.2.2 L'éthique dans la finance . . . . . 112
V.2.3 Description du simulateur . . . . . 113
V.3 Évaluation de l'influence du jugement sur le comportement individuel . . . . 116
V.3.1 Description métier du domaine . . . . . 116
V.3.2 Paramétrage moral des agents . . . . . 117
V.3.2.1 Valeurs . . . . . 117
V.3.2.2 Supports de valeurs . . . . . 118
V.3.2.3 Règles morales . . . . . 118
V.3.3 Initialisation . . . . . 119
V.3.4 Comportement des agents . . . . . 120
V.4 Évaluation du jugement des autres . . . . . 121
V.4.1 Initialisation . . . . . 122
V.4.2 Agrégation d'images . . . . . 122
V.5 Synthèse . . . . . 123
Bilan et perspectives . . . . . 125
Bibliographie . . . . . 131
Bibliographie . . . . . 131

# Table des figures


I.1 Le triangle de l'éthique, illustration courante du rôle de la théorie du juste. 12

I.2 Représentation synthétique des concepts employés dans le jugement . . . . 16

I.3 Localisation des aires cérébrales impliquées dans le jugement (Greene, Haidt, 2002) . . . . . 17

I.4 Modèle social-intuitionniste du jugement (Haidt, 2001) . . . . . 20

I.5 Représentation graphique du modèle de valeurs de Schwartz (Schwartz, 1994) . . . . . 24

I.6 Représentation synthétique de la diversité des théories du bien et du juste. 25

II.1 Espace des actions selon (Arkin, 2009) . . . . . 34

II.2 Architecture de l'Ethical Governor (Arkin, 2009) . . . . . 34

II.3 Architecture d'agent mettant en oeuvre CPB (Anderson *et al.*, 2017) . . . 37

II.4 Phase d'apprentissage (Yamamoto, Hagiwara, 2014) . . . . . 38

II.5 Phase de jugement (Yamamoto, Hagiwara, 2014) . . . . . 38

II.6 Cadre déclaratif modulaire de (Berreby *et al.*, 2017a ; 2017b) . . . . . 41

II.7 Architecture d'agent BDI émotionnel (Battaglino *et al.*, 2013) . . . . . 44

II.8 Proposition d'architecture d'agent moral (Coelho, Rocha Costa, Trigo, 2010)45

III.1 Illustration du scénario pris en exemple : Robin Hood . . . . . 55

III.2 Intégration du modèle de jugement dans l'architecture BDI . . . . . 57

III.3 Modèle de reconnaissance de situation . . . . . 60

III.4 Modèle d'évaluation de la désirabilité et de la possibilité . . . . . 61

III.5 Modèle d'évaluation de la moralité . . . . . 64

III.6 Modèle d'évaluation de l'éthique . . . . . 71

III.7 Représentation globale du modèle de jugement . . . . . 75

IV.1 Modèle de coopération entre agents autonomes fondé sur l'éthique . . . . . 92

V.1 Représentation générale d'un système multi-agent conçu avec JaCaMo . . . 101
V.2 Architecture du système de place de marché pour la preuve de concept de notre proposition : des agents $a_i$ interagissent sur une place de marché partagée dont le fonctionnement est rythmé par une horloge commune . . . 113
V.3 Exécution d'un ordre lors de son ajout sur le marché . . . . . 114
V.4 Représentation en chandeliers de l'évolution du cours d'un titre lors d'une expérimentation. . . . . 115
V.5 Évolution de la composition du portefeuille d'actifs d'un agent écologiste . 120
V.6 Évolution des images des agents en sortie de la fonction d'agrégation éthique générées au sein d'un agent de type ethic-both . . . . . 122

# Liste des tableaux

I.1 Fonctions connues des aires cérébrales impliquées dans le jugement . . . . . 18
I.2 Récapitulatif de la théorie du développement moral de Kohlberg (Kohlberg, Hersh, 1977) . . . . . 22

II.1 Tableau récapitulatif des travaux existants . . . . . 47

III.1 Tableau récapitulatif des évaluations lorsque Robin Hood est pauvre. . . . . 73
III.2 Tableau récapitulatif des évaluations lorsque Sheriff of Nottingham n'est ni pauvre ni riche. . . . . 76

IV.1 Tableau récapitulatif des évaluations lorsque Robin Hood n'est ni pauvre ni riche. . . . . 90

V.1 Nombre de transactions par types d'agents en dix simulations de 30 minutes121

# Introduction

## Contexte

Le déploiement d'*agents autonomes artificiels* dans des systèmes où ils interagissent entre eux et éventuellement avec des humains pose un certain nombre de problèmes lorsque ces agents doivent prendre des décisions qui, pour un humain, soulèvent une ou plusieurs interrogations éthiques.

Le terme d'agent autonome artificiel, que nous appellerons par simplicité *agent* ou *agent autonome* dans la suite de cet ouvrage, désigne un concept classique en intelligence artificielle (Jennings *et al.*, 1998). Un *agent autonome artificiel* est un logiciel ou une machine, situé dans un environnement, capable d'effectuer des actions de manière autonome et flexible afin d'atteindre des objectifs.

Le fait que l'agent soit « situé dans un environnement » signifie qu'il acquiert des informations sur son environnement au moyen de perceptions et qu'il peut agir en retour sur cet environnement au moyen d'actions. L'autonomie est un concept plus difficile à définir mais, dans ce contexte, nous pouvons considérer que cela désigne la capacité à prendre des décisions et agir sans intervention d'un être humain ou d'un autre agent. Enfin la flexibilité signifie que l'agent est *réactif* (il réagit aux changements dans son environnement), *pro-actif* (il ne se contente pas de réagir, mais il est capable de prendre des initiatives pour atteindre ses objectifs) et *social* (en plus d'interagir avec son environnement, il peut interagir avec d'autres agents autonomes et humains).

La modélisation d'agents en interaction dans un même environnement est appelée un *système multi-agent* que nous définissons, en accord avec (Ferber, 1995) comme un ensemble d'agents en interaction les uns avec les autres, situés dans un environnement commun et éventuellement construisant ou prenant part à une organisation.

## Problématique

La délégation de décisions à de tels agents dans de nombreux domaines tels que la santé, la finance ou les transports, dans lesquels les humains perçoivent dans la prise de certaines décisions une problématique éthique, pose la question de l'intégration de cette dimension éthique au processus décisionnel de l'agent. Par exemple, dans le domaine médical, un agent autonome chargé de veiller sur un patient et l'accompagner dans sa

démarche de soins (Kang *et al.*, 2005) devrait pouvoir tenir compte dans son processus de prises de décision des éléments de déontologie et des valeurs auxquels doivent se conformer les membres du personnel médical (liés par exemple au respect de la dignité du patient, au respect de sa vie privée ou à son droit d'accès sur les données qui le concernent). Les préoccupations éthiques intervenant dans la réflexion semblent dépendantes du domaine applicatif. Par exemple, les notions de responsabilité et de transparence jouent un rôle important dans la gestion de capitaux et la protection de l'intégrité physique des usagers semble primer dans les applications liées au transport.

La communauté académique s'est saisie de cette problématique en intelligence artificielle et de nombreux articles, évènements, workshops [2] et standards (Chatila *et al.*, 2017) témoignent du vif intérêt pour ces questions. Parmi les nombreux travaux et communications traitant de cette problématique, citons tout d'abord (Ruvinsky, 2007), qui définit le champ de l'*éthique computationnelle* comme l'intégration de simulation informatique et de théories d'éthique. Plus spécifiquement, l'éthique computationnelle est une simulation à base d'agents qui donne une perspective computationnelle aux théories d'éthique. Ces approches utilisent des modèles informatiques et des systèmes multi-agents pour générer des sociétés d'agents capables d'adopter divers principes éthiques.

Autrement dit, l'éthique computationnelle est donc l'intégration, dans le raisonnement des agents autonomes, de processus permettant d'employer des concepts empruntés à la philosophie. Elle ne doit pas être confondue avec la *roboethique* qui rassemble des travaux de philosophie traitants de la problématique de ce que devrait être l'éthique d'un agent autonome.

# Objectifs

L'éthique computationnelle propose déjà à l'heure actuelle de nombreux modèles d'agents autonomes permettant d'intégrer divers concepts de philosophie de diverses manières dans le raisonnement des agents. Cependant, il semble que la préoccupation première de ces travaux soit la représentation d'un processus interne à l'agent ne lui permettant de prendre des décisions qu'en regardant l'adéquation entre des théories philosophiques et sa situation sans tenir compte de la dimension sociale de l'éthique. En effet, l'éthique ne se cantonne pas à une description de la manière de juger personnellement de la meilleure action à effectuer dans une situation, mais elle permet également aux humains de juger le comportement des autres et influencer ainsi les interactions sociales.

Nous répondons ici aux nouveaux enjeux et aux nouvelles interactions entre agents

2. Symposium on Roboethics - www.roboethics.org,
International Conference on Computer Ethics and Philosophical Enquiry - philevents.org/event/show/15670,
Workshop on AI and Ethics, AAAI conference - www.cse.unsw.edu.au/~tw/aiethics,
International Conference on AI and Ethics -
wordpress.csc.liv.ac.uk/va/2015/02/16/
Workshop on Ethics in the Design of Intelligent Agents www.ecai2016.org/content/uploads/2016/08/W15-edia-2016.pdf
Workshop « Éthique et IA » lors de la Plateforme Intelligence Artificielle 2015 pfia2015.inria.fr/journees-bilaterales/ethique-et-ia et 2017 pfia2017.greyc.fr/ethique/presentation

issues des problématiques liées à la cohabitation au sein d'un même environnement d'agents autonomes ayant potentiellement des éthiques différentes. Nous cherchons à montrer comment l'éthique d'un agent peut être employée non seulement pour prendre des décisions, mais également pour juger les autres et servir de cadre à un mécanisme de coopération fondée sur l'éthique. Cette thèse défend l'idée que l'emploi d'une approche déclarative et rationaliste de l'éthique est la plus appropriée pour concevoir un modèle de jugement permettant à un agent autonome de prendre des décisions en lien avec l'éthique de leur comportement et celui des autres agents avec lesquels il peut coopérer.

Nous définissons dans une architecture BDI les états mentaux et les connaissances permettant de représenter la situation connue de l'agent et ses connaissances du bien et du juste, ainsi que les raisonnements permettant de mettre en œuvre un jugement de l'éthique des actions. Nous évaluons ensuite ce modèle par un ensemble d'expérimentations dans un cadre applicatif réaliste en observant le comportement des agents et leurs représentations mentales de l'évaluation du comportement des autres.

# Plan

Ce document est organisé en cinq chapitres. Les deux premiers présentent un état de l'art apportant un ensemble de définitions permettant de préciser la problématique, puis positionnant les divers travaux de la littérature issue des sciences humaines et sociales ainsi que de l'Intelligence artificielle et des systèmes multi-agents. Les deux chapitres suivants détaillent la contribution de cette thèse en présentant un modèle de jugement pouvant être utilisé comme processus décisionnel et permettant de proposer un cadre de coopération fondé sur l'éthique. Un cinquième chapitre présente la mise en œuvre dans une plateforme multi-agent et des expérimentations montrant l'implémentation de ce modèle et l'évaluation de son influence sur le comportement des agents.

Le premier chapitre se concentre sur la présentation des concepts d'éthique employés par la suite et des interactions entre ces concepts. La première partie de ce chapitre s'attache les définir dans le cadre de réflexions philosophiques en mettant en lumière les divergences entre les diverses théories proposées par la littérature et en montrant comment ces concepts s'articulent entre eux. La suite de ce chapitre apporte un complément à cette vision globale de l'éthique en montrant comment la neurologie, la psychologie, les sciences humaines et les sciences sociales apportent des éléments supplémentaires à la compréhension des processus mentaux à l'œuvre dans le jugement éthique et pointent des différences entre individus. Ce chapitre aboutit sur un modèle informel du jugement et met en évidence les différentes théories philosophiques que devrait permettre de représenter un modèle opérationnel.

Le second chapitre précise les problématiques d'éthique computationnelle en lien avec les objectifs de nos travaux à la lumière des définitions du premier chapitre, puis il expose quelques modèles proposés par des travaux d'intelligence artificielle. Ces travaux sont comparés à l'aide d'une grille de lecture afin de mettre en évidence des caractéristiques que sont la généricité des modèles présentés, afin de permettre de paramétrer

les agents avec des éthiques diverses et de pouvoir utiliser ces modèles indépendamment du domaine applicatif, le caractère rationnel du jugement et la représentation explicite des connaissances employées afin de pouvoir expliquer et justifier le comportement des agents. Nous ajoutons un critère qui est le caractère opérationnel des propositions présentées, afin de pouvoir évaluer expérimentalement les résultats. Nous utilisons cette étude comparative pour affiner nos objectifs et définir un ensemble de principes fondateurs pour nos contributions.

Le troisième chapitre présente notre proposition de modèle de jugement en détaillant et définissant les composants utilisés et en montrant comment il peut être employé comme processus de décision. Ce modèle propose de distinguer la morale et l'éthique, la première produisant des évaluations du caractère bon ou mauvais des actions en fonction de définitions de valeurs morales et de règles morales données en paramètre à l'agent, et la seconde permettant de déterminer l'action juste dans une situation au regard d'un ensemble ordonné de principes éthiques. Les représentations de ces éléments de connaissances sont distinctes des fonctions permettant de les employer afin d'assurer la généricité du modèle par rapport aux approches qu'il est possible d'employer pour exprimer la morale de l'agent et par rapport aux spécificités du domaine dans lequel l'agent est déployé. Ces définitions sont accompagnées d'exemples permettant d'illustrer la modélisation des concepts présentés au premier chapitre.

Le quatrième chapitre apporte des éléments supplémentaires au modèle de jugement proposé afin de mettre en place un cadre de coopération multi-agent fondé sur l'éthique. Ce modèle de coopération entre agents s'appuie sur un modèle de confiance dans lequel les agents se construisent une représentation du comportement des autres par des jugements progressifs. Le paramétrage de ce jugement des autres peut-être lui même soumis à une certaine éthique, ce qui permet de définir l'éthique du comportement social de l'agent.

Le cinquième chapitre présente la mise en œuvre et des expérimentations menées pour montrer comment le modèle proposé peut être mis en œuvre et implémenté dans des agents autonomes, dans un domaine applicatif réaliste. Nous illustrons le fonctionnement du modèle de jugement dans le cadre d'agents autonomes chargés de gérer des portefeuilles d'actifs financiers sur un marché d'actions. Ces agents autonomes reçoivent des représentations différentes de valeurs morales et règles morales et nous évaluons la cohérence entre ces connaissances et les comportements produits. Nous montrons également les images produites par le jugement du comportement des autres agents et évaluons la cohérence entre l'évolution de ces images et les connaissances employées pour les produire. Les résultats de ces expérimentations permettent de vérifier un certain nombre de propriétés du modèle et de son implémentation.

Cette thèse s'achève sur une conclusion présentant une prise de recul synthétisant notre démarche ainsi que les contributions de nos travaux. Elle présente également les perspectives nouvelles que soulève ce travail.

# CHAPITRE I

# L'éthique et la morale


**I.1 En philosophie** . . . . . **6**
I.1.1 La morale . . . . . 6
I.1.2 L'éthique . . . . . 11
**I.2 En neurologie, psychologie et sciences cognitives** . . . . . **17**
I.2.1 Organisation de la cognition éthique . . . . . 17
I.2.2 Rôle des émotions et du raisonnement . . . . . 19
I.2.3 Facultés sociales et théorie de l'esprit . . . . . 20
**I.3 En sciences sociales** . . . . . **21**
I.3.1 Théorie du développement moral. . . . . 21
I.3.2 Les valeurs morales en sociologie . . . . . 22
**I.4 Synthèse** . . . . . **24**


Ce chapitre a pour but de présenter au lecteur les concepts de morale et d'éthique permettant de définir la terminologie employée dans les chapitres suivants, présenter les interactions entre ces éléments et montrer la diversité des approches philosophiques permettant de les utiliser. La présentation de ces concepts s'appuiera sur une vision pluridisciplinaire des problématiques liées au jugement dans divers travaux de philosophie, de psychologie, et de sciences humaines et sociales.

Le problème considéré ici est celui de la définition de l'ensemble des éléments intervenant dans le raisonnement permettant de prendre une décision éthique et répondre aux problèmes posés en introduction de ce mémoire. Nous cherchons ainsi à mettre en évidence à la fois les étapes de ce raisonnement et la nature des connaissances employées. L'objectif de cette analyse est d'identifier et définir les composants du jugement éthique d'une action chez l'être humain afin de discuter dans le prochain chapitre de la production d'un tel jugement chez l'agent autonome.

La section I.1 définit les concepts employés en philosophie en montrant comment ces éléments s'articulent au sein d'un raisonnement et en quoi la manière dont ils sont employés et les connaissances utilisées divergent selon les écoles de pensée. La section

I.2 vient ensuite montrer comment des travaux de psychologie et de sciences cognitives proposent des modèles permettant de représenter et comprendre davantage ces processus. Enfin la section I.3 présente des travaux de sciences sociales portant sur la diversité des éléments intervenant dans le raisonnement éthique au sein de la société humaine. Ce chapitre s'achève sur une courte synthèse rappelant les principales notions et les apports des diverses disciplines dans la compréhension du raisonnement éthique.

## I.1 En philosophie

Dès l'antiquité, les philosophes ont élaboré des théories sur la capacité humaine à définir et distinguer le bien et le juste du mal et de l'injuste. Bien que la littérature philosophique propose une grande variété de réponses, employant des termes pouvant changer de signification d'un auteur à l'autre et employant ou non des concepts et raisonnements radicalement différents, nous nous attachons ici à montrer les éléments récurrents d'une part et les grandes catégories d'approches d'autre part. Cette section est structurée en deux parties. La première partie traite de la morale, que nous distinguons de l'éthique faisant l'objet de la seconde. Nous justifions et expliquons notre choix d'effectuer cette distinction, que nous savons sujette à débats, en début de section I.1.2.

### I.1.1 La morale

La majorité des êtres humains semble dotée de facultés lui permettant de savoir si une action, dans une certaine situation est bonne ou mauvaise. Nous appelons *morale* ou *théorie du bien* (*theory of the good*, voir (Timmons, 2012)) cette capacité à distinguer le bien du mal.

La nature et l'origine de cette morale sont des questions récurrentes touchant à des problèmes de compréhension du fonctionnement de l'esprit humain et des problèmes de métaphysique (au sens de la formulation de théories donnant sens à la place de l'Homme dans l'univers). La conception de la morale diffère selon les individus, les doctrines et les sociétés. Diverses théories s'opposent, que nous pouvons positionner vis-à-vis de deux critères :

- L'*universalisme moral* est une conception, s'opposant au *relativisme moral* (Gowans, 2016 ; Richardson, Williams, 2009), affirmant qu'une théorie du bien est valable indépendamment du temps, du lieu, et de tout groupe d'individus. Le relativisme, à l'inverse en fait un élément culturel, admettant des évolutions dans le temps et des nuances entre les communautés humaines.
- L'*absolutisme moral* est une conception tendant à attribuer aux actions une valeur morale bonne ou mauvaise en soi (Rai, Holyoak, 2013), c'est-à-dire indépendamment de tout contexte, tandis qu'à l'inverse d'autres conceptions telles que les *théories normatives* et le *contextualisme* admettent que déterminer si une action est bonne ou mauvaise nécessite d'en connaître le contexte, les conséquences et le but recherché par son accomplissement.

Ces conceptions de la morale sont fortement corrélées avec la conception métaphysique

du monde. Les principales religions font de la morale un don divin (par exemple les tables de la loi données à Moise pour les religions judéo-chrétiennes, le Coran dicté par l'Ange Gabriel à Mahomet pour l'Islam, le Veda révélé aux Rishis par Brahmā, etc.). Cela implique une vision universaliste d'une morale prenant la forme d'une doctrine infaillible, parfaite et sacrée qui ne peut ainsi être remise en cause par les croyants qui doivent chercher à l'interpréter et l'appliquer. De même, l'idée d'uniformiser les mœurs à l'échelle internationale par des moratoires (sur la peine de mort, l'exploitation des ressources, etc.) est l'effet sur le domaine juridique d'une vision universaliste de la morale. Cela implique l'existence d'un objectivisme moral, c'est-à-dire la possibilité de défendre rationnellement la totalité des éléments d'une théorie du bien. Certaines doctrines philosophiques, telles que l'impératif catégorique développé par Kant dans ses *Fondements de la métaphysique des mœurs* (Kant, 1785), sont absolutistes en ce sens qu'elles cherchent à attribuer dans tout contexte aux actions une valuation morale, arguant que même si l'objectif est louable, certaines actions (telles que le mensonge) sont nocives en elles-mêmes pour la société. À l'inverse, d'autres penseurs tels que Benjamin Constant (Constant, Kant, 2003 ; Constant, 2013) s'efforcent de montrer que la moralité d'une action est nécessairement dépendante du contexte, de ses conséquences et du but recherché par l'agent.

En plus de ces deux axes, les théories sur le fonctionnement de la morale sont à positionner en fonction de l'origine de la morale (Kauppinen, 2017) :

- L'*intuitionnisme* ou *sentimentalisme* défend l'idée d'une morale guidée par les émotions (ou passions) de l'individu. De ce point de vue, l'Homme agit moralement afin de s'éviter des souffrances et atteindre le bonheur.
- Le *rationalisme* est, à l'inverse, le fait de considérer la raison comme étant le moyen de déterminer les actions bonnes et mauvaises. Les passions sont des perturbations de cette analyse.

Parmi les philosophes ayant le plus fermement défendu et influencé la vision rationaliste, nous pouvons relever Descartes qui, dans son *Discours de la méthode* (Descartes, 1960 ; Rutherford, 2017) explique que la morale parfaite ne peut découler que d'une réflexion complète sur toutes les conséquences de nos actions et que, faute d'une telle connaissance complète et parfaite, il est nécessaire de se munir d'une *morale par provision*, c'est-à-dire une vision intuitive à rectifier par raisonnement pour tendre vers la véritable morale. Pour Kant et les philosophes influencés par l'idéalisme allemand, il est nécessaire que la morale se repose sur la *raison pratique* car, sans objectivisme moral, il n'est plus possible de justifier que la même morale soit imposée à l'ensemble de la société (Kant, 1785 ; Korsgaard, 1996).

À l'inverse, des penseurs plus matérialistes tels que David Hume (Hume, 2006), Baruch Spinoza (Spinoza, 1677) ou Friedrich Nietzsche (Nietzsche, 2015) conçoivent la morale d'un point de vue relativiste, sentimentaliste et intuitionniste, c'est-à-dire comme l'expression d'émotions et de convictions personnelles résultant d'une recherche intime du bien. Les morales élaborées par différents individus ou communautés peuvent alors être perçues comme équivalentes au sens où ce ne sont que des moyens différents de parvenir

au bonheur. David Hume (Hume, 2006) explique que « la raison est l'esclave des passions, elle ne peut prétendre à un autre rôle que la servir et lui obéir ». La morale est alors conçue comme une interprétation d'émotions telles que l'attirance, le dégout ou la peur ressentis lorsque nous imaginons ou faisons face à certains évènements.

Nous proposons une définition synthétique du concept de *morale* tenant compte de la diversité des positionnements sur les axes définis précédemment :

**Définition 1** (Morale)**.** *La* morale *désigne l'ensemble de règles déterminant la conformité des pensées ou actions d'un individu avec les mœurs, us et coutumes d'une société, d'un groupe (communauté religieuse, peuple, etc.) ou d'un individu pour évaluer son propre comportement ou celui d'autrui. Ces évaluations reposent sur les valuations de bien et de mal. Elle peut être universelle ou relative (c'est-à-dire indépendante ou non d'un lieu, d'une époque, d'un peuple), absolue ou contextuelle (c'est-à-dire indépendante ou non du contexte), rationnelle ou intuitive (c'est-à-dire produit d'un raisonnement indépendant ou non de réactions émotionnelles).*

Notons que la morale se distingue de la loi et du système légal. En effet, la morale ne comporte pas de pénalités explicites et ne repose pas nécessairement sur des règles officiellement établies (Gert, 2015). Outre la distinction de nature (la loi est imposée par l'État et fixe des peines pour les infractions), la différence entre loi et morale est parfois trouble. Par exemple, le code de déontologie médicale, en France, est intégralement retranscrit dans la loi (articles R.4127- 1 à 112 du code de santé publique), ce qui fait de toute infraction à ce code moral une infraction au regard de la loi. Cependant, même si de tels éléments sont communs aux ordres moraux et légaux, il existe des domaines où la morale peut inciter à des actes qui sortent du domaine de la loi (charité, relations amicales, etc.). Dans certains cas, la morale peut même aller à l'encontre de la loi, par exemple dans le cas du devoir moral de résistance contre un pouvoir tyrannique (Thoreau, 2016).

La morale décrit les droits et devoirs moraux des individus sous la forme d'un ensemble de règles. Chacun connaît des règles telles que « il est mal de mentir », « il est bon d'être loyal » ou « il est mal de tricher ». C'est sur ce type de règles que peut se fonder un raisonnement permettant de distinguer les bonnes et mauvaises actions. Les *règles morales* sont couramment soutenues et justifiées par des *valeurs morales* (liberté, bienveillance, sagesse, conformisme, etc.). Maintenant que nous avons défini la morale et montré la diversité des approches proposées, nous nous intéressons à ces concepts de *règles morales* et *valeurs morales*.

#### I.1.1.1 Les règles morales

La morale, lorsqu'elle est explicitée (c'est-à-dire énoncée à l'écrit ou à l'oral dans un langage naturel), peut prendre de nombreuses formes à la fois extensives (dans une démarche déontologique) et intensives (par l'expression de maximes ou de sentences). Par souci de simplicité, nous traitons par la suite essentiellement de la forme extensive.

Les règles morales telles que le commandement « Tu ne tueras point » décrivent la

moralité d'une action (ici le meurtre est immoral puisque désobéir à un commandement divin est nécessairement immoral pour un croyant). Nous considérons par la suite, comme équivalentes les formes impératives (exemple : « Tu ne commettras point d'adultère ») et les formes affirmatives (exemple : « Il est immoral de commettre un adultère ») de ces règles afin d'uniformiser leur écriture et expliciter la valuation de moralité attribuée à l'action par la règle.

La formulation d'ensembles de règles a, dans l'Histoire, souvent été employée pour exprimer le comportement attendu des membres d'une communauté. Les règles de vie monastique (par exemple : règle de Saint Benoît rédigée au sixième siècle) définissent les devoirs des membres de congrégations religieuses en mêlant à la fois des règles de vie et des devoirs spirituels (obéissance, humilité) relevant de la morale. La morale peut également être exprimée sous forme de maximes et de sentences (La Rochefoucauld, Vauvenargues, 1867). Une telle forme est principalement employée pour son aspect esthétique puisque la compréhension de la règle morale sous-jacente demande un effort d'interprétation. Plus récemment, nous assistons à la prolifération des codes de déontologie professionnelle dans les entreprises fixant et exprimant la morale du comportement des employés (Gendron, 2005).

Nous proposons de définir le concept de règle morale de la manière suivante :

**Définition 2** (Règles morales)**.** *Une* règle morale *est une norme attribuant à une action ou à un ensemble d'actions une valuation de bien ou de mal. L'application d'une règle morale peut être réduite à un ensemble de situations dans le cadre d'une approche contextualiste, ou à un ensemble d'individus dans un cadre relativiste.*

#### I.1.1.2 Les valeurs

Certaines règles morales telles que « Il est moral d'être généreux »ne désignent pas des actions directement mais font appel à des notions plus abstraites nommées *valeurs morales* promues ou trahies par des ensembles d'actions. Dans cet exemple, la générosité est une valeur morale employée pour qualifier toute action qui la promeut comme morale.

L'emploi de valeurs permet ainsi d'attribuer une valuation de moralité (tel que immoral, amoral, moral, etc.) dans une règle morale à un ensemble d'actions, en les désignant par la valeur qu'elles supportent ou trahissent. Par exemple, la règle « Il est moral d'honorer ses parents » (quatrième commandement dans le christianisme) fait appel à la valeur d'*honneur*. Toute action pour laquelle la personne portant le jugement estime qu'elle honore les parents sera ainsi évaluée comme moralement bonne au regard de cette règle morale.

La nécessaire connaissance de la sémantique des valeurs employées peut conduire à des divergences d'interprétations d'une même règle par des individus différents. Par exemple si deux individus sont en désaccord sur le fait qu'un mensonge par omission trahisse la valeur d'honnêteté, l'un considèrera que la règle « Il est immoral de ne pas être honnête » fait de cette forme de mensonge une action mauvaise tandis que l'autre ne considèrera pas cette action comme faisant partie de l'ensemble des actions qualifiées d'immorales

par cette règle.

Certaines écoles philosophiques, notamment socratiques et platoniciennes, font des valeurs morales l'élément central de la théorie du bien en désignant certaines d'entre elles comme des *vertus*, c'est-à-dire des valeurs morales dont la mise en pratique mène au bonheur de l'Homme. La promotion de valeurs telles que la sagesse, la vérité, la justice ou le courage doit alors être la principale préoccupation de l'Homme de bien.

Nous proposons alors la définition suivante :

**Définition 3** (Valeur)**.** *Une* valeur morale *est une croyance fondamentale pour l'expression de la morale chez l'individu. Les valeurs morales peuvent être promues ou trahies par diverses actions.*

L'*axiologie*, ou *théorie des valeurs*, est une branche de la philosophie s'attachant à organiser et ordonner ces valeurs (Schroeder, 2016). Certaines valeurs ont alors une plus grande importance dans la morale que d'autres (ce point sera approfondi en section I.3.2 grâce à des travaux de sciences sociales).

#### I.1.1.3 Les dilemmes moraux

Bien que l'humain soit capable de prendre en compte la morale dans ses décisions, les philosophes ont relevé l'apparition de situations dans lesquelles une décision est problématique puisque la morale seule ne permet pas de désigner la bonne décision. Nous appelons une telle situation *dilemme moral* que nous définissons de la manière suivante en s'appuyant sur (McConnell, 2014) :

**Définition 4** (Dilemme moral)**.** *Un* dilemme moral *est un choix entre plusieurs options que la morale évalue chacune comme à la fois bonnes et mauvaises, dans des proportions similaires, sans qu'il soit possible de les réaliser simultanément.*

**Exemple 1.** *Un cas célèbre de dilemme moral est celui communément nommé* Problème du tramway *(Foot, 1967) que nous pouvons formuler de la manière suivante :*

*« Imaginons le conducteur d'un tramway hors de contrôle qui se doit de choisir sa course entre deux voies possibles : cinq hommes travaillent sur la voie sur laquelle est engagé le tram et un homme est situé sur l'autre. La voie prise par le tram entraînera automatiquement la mort des personnes qui s'y trouvent. Le conducteur du tram dispose d'un levier permettant de changer de voie. »*

*Nous formulons les règles morales, implicites dans le problème exposé ici, de la manière suivante : « Il est immoral de causer volontairement la mort d'une ou plusieurs personnes », « Il est immoral, par inaction, de laisser mourir une ou plusieurs personnes ».*

*Dans cette situation, toute décision du conducteur (c'est-à-dire actionner ou non le levier) serait qualifiée d'immorale par l'une des deux règles.*

Les dilemmes moraux montrent que les règles morales peuvent être insuffisantes pour déterminer la décision appropriée face à une situation. La morale peut présenter des

contradictions, évaluer une action comme étant bonne et mauvaise simultanément ou qualifier toute option de mauvaise au regard d'au moins une règle morale. Ces dilemmes ont mis en évidence la nécessité d'un niveau supérieur de raisonnement permettant de juger de l'action juste à effectuer dans une situation.

### I.1.2 L'éthique

Même si la théorie du bien permet de qualifier les actes de bons ou mauvais, parfois simultanément, au regard de règles et de valeurs morales, cela est insuffisant dans certaines situations pour déterminer l'action qu'il convient d'effectuer. Que faire lorsqu'une action est à la fois bonne et mauvaise ? Un tel cas est par exemple celui de la légitime défense où il est nécessaire de causer du mal à quelqu'un pour empêcher qu'un autre mal soit causé à soi-même ou autrui. Un tel raisonnement, s'il est acceptable à l'échelle d'un individu, peut-il être transposé à l'échelle d'un état, par exemple pour justifier la torture de terroristes afin de sauver des vies ? Peut-on effectuer de mauvaises actions afin d'empêcher que d'autres en commettent des pires ? Doit-on faire passer toute bonne action que nous sommes en mesure de réaliser avant celles que nous désirons ? Comment prendre en compte l'existence d'alternatives possibles dans l'évaluation d'une action ?

Ces questions montrent que la théorie du bien, qui permet d'évaluer en quoi toute action est bonne et/ou mauvaise, est insuffisante pour choisir laquelle effectuer. Les nombreux principes éthiques élaborés par les philosophes afin de répondre à ces problématiques sont au cœur de ce que nous nommons l'*éthique*.

Notons tout d'abord que cette distinction entre éthique et morale ne fait pas l'unanimité en philosophie (Downie, 1980). La première étant de racine latine, l'autre de racine grecque, de nombreux auteurs en font de parfaits synonymes interchangeables à volonté. D'autres auteurs, tels que Paul Ricoeur, soutiennent au contraire qu'il existe une différence, définissant l'éthique comme une discipline complémentaire de la morale, qui devient nécessaire lorsque cette dernière conduit à des impasses pratiques (voir (Ricoeur, 1995), début de septième étude). De même, (Comte-Sponville, 2012), s'inspirant de l'idée des ordres pascaliens décrivant des ensembles d'éléments à prendre en considération dans une décision, fait de l'éthique l'un de ces quatre *ordres* qui guident la décision :

- l'ordre *économique, technique et scientifique* structuré intérieurement par l'opposition du possible et de l'impossible ;
- l'ordre *juridique et politique* structuré par l'opposition du légal et de l'illégal ;
- l'ordre *moral* structuré par l'opposition du bien, du devoir, au mal et à l'interdit ;
- l'ordre *éthique* structuré par l'opposition de la joie et de la tristesse.

Le premier ordre (économico-techno-scientifique) est limité par le second (juridico-politique), qui est lui-même limité par le troisième (la morale). C'est-à-dire que chacun de ces ensembles prévaut sur le précédent. L'ordre éthique, d'après ce philosophe, vient compléter la morale pour permettre d'accéder au bonheur.

Nous adoptons par la suite cette distinction entre la morale et l'éthique, considérant que cela contribue à clarifier la compréhension générale du problème. De plus, certains

travaux de sciences humaines que nous aborderons en section I.3 s'attachent à proposer des modèles de l'une ou de l'autre.

L'éthique peut également être appelée *théorie du juste* ou *théorie de la juste conduite* (*theory of the right*, voir (Timmons, 2012)). Elle nécessite l'existence d'une théorie du bien (voir section I.1.1) et permet de désigner, au regard de celle-ci, la ou les actions justes de celles qui sont injustes. L'éthique permet également de concilier l'ensemble des désirs avec la morale compte tenu des capacités de l'agent en déterminant dans quelle mesure l'un doit primer sur l'autre. La figure I.1 montre une illustration courante de l'éthique, comme un espace de réflexion regroupant les compromis possibles entre compétences, désirs et devoirs moraux d'un individu.

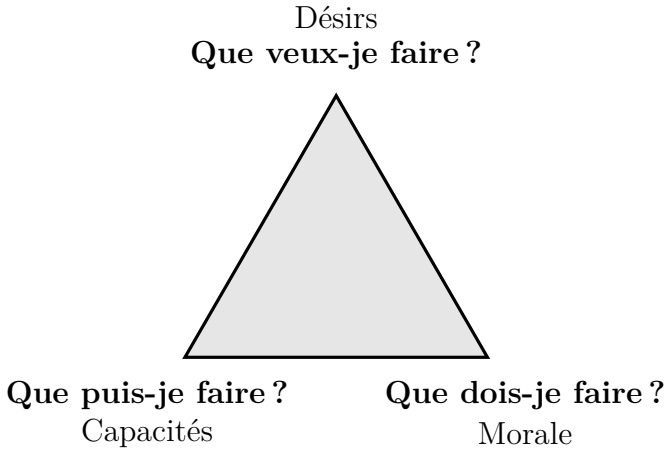


FIGURE I.1 – Le triangle de l'éthique, illustration courante du rôle de la théorie du juste.

Après lecture de ces travaux, nous proposons une définition synthétique de l'éthique dont nous allons définir les composants et les approches par la suite :

**Définition 5** (Éthique)**.** *L'*éthique *est la combinaison de* principes éthiques *permettant à un agent de déterminer dans une situation, au regard de connaissances et d'une morale, l'action qu'il est juste d'effectuer.*

### I.1.2.1 Les principes éthiques

Comme la littérature philosophique propose une grande variété de réponses au problème de l'évaluation de la justesse d'une action, nous définissons le concept de *principe éthique*, représentant tout type de raisonnement permettant d'évaluer cette justesse :

**Définition 6** (Principe éthique)**.** *Un* principe éthique *est un ensemble consistant de règles permettant de distinguer les actions justes compte tenu de la morale, des désirs et capacités de l'agent.*

Un principe éthique représente un raisonnement philosophique qui qualifie chaque action, dans une situation, de juste ou d'injuste. Comme de nombreux principes sont proposés dans la littérature philosophique, nous présentons quelques catégories classiques

d'approches de l'éthique. Les sections I.1.2.2, I.1.2.3 et I.1.2.4 présentent trois grandes catégories d'approches éthiques.

### I.1.2.2 Éthique des vertus

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 cherchent à déterminer quelles sont les valeurs morales 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 quelles sont les vertus (c'est-à-dire les valeurs qu'il est bon de promouvoir) et les vices (à l'inverse des vertus, les vices sont des valeurs qu'il est mal de promouvoir) qui doivent guider le comportement de l'Homme pour le conduire au bonheur. 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. Un principe éthique relevant de cette approche s'attache à définir comment évaluer des actions dont la moralité dépend du support ou de la trahison de valeurs morales. Certains philosophes désignent l'une de ces vertus comme cardinales. Par exemple, la sagesse est la plus grande des vertus pour Socrate tandis qu'il s'agit de la justice pour Platon. Le problème essentiel posé par l'éthique des vertus est donc celui de la définition du sens des valeurs et de l'ordre de leur importance.

**Exemple 2.** *Pour Platon, la justice (au sens de l'attribution des rôles dans la société) est la plus grande vertu car elle est celle qui assure l'ordre dans la société. Ainsi, par exemple, un individu qui serait face à un dilemme moral dans lequel l'une des actions est soutenue par la justice et l'autre par le courage (par exemple laisser agir des forces de l'ordre impuissantes ou venger soi-même un être cher) devrait toujours opter pour la première.*

### I.1.2.3 Éthique déontologique

L'*éthique déontologique* (du grec $\delta\epsilon o\nu$, signifiant « il faut ») juge un comportement par sa conformité avec un code de déontologie édictant à l'aide d'un ensemble de règles morales 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. Comme expliqué en section 4, le principal problème est l'apparition de dilemmes moraux par des contradictions entre des devoirs et interdits moraux, ou bien l'équivalence des actions au regard des règles morales. La principale difficulté posée par une telle approche est alors de déterminer à quels devoirs ou interdits accorder une priorité.

**Exemple 3.** *Des philosophes tels que Kant et Aristote on tenté de répondre à ce problème par la définition de principes généraux permettant de répondre à certains dilemmes moraux. Pour Kant, par exemple, l'Homme doit toujours chercher à agir conformément à des règles qui, si elles étaient universelles, conduiraient au bien de la société entière. Ainsi par exemple, il est interdit de mentir pour éviter une situation immorale puisque la généralisation de ce comportement mènerait à société où tout le monde ment et où la confiance entre individus serait impossible (ce qui est mauvais selon Kant).*

#### I.1.2.4 Éthique conséquentialiste

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. Les nuances entre courants du conséquentialisme portent plus sur la définition de la moralité des conséquences. Par exemple,

- l'*hédonisme* cherche à maximiser le plaisir en minimisant la souffrance. Ce courant accorde davantage d'importance aux désirs qu'à la dimension morale des actions ;
- 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, c'est-à-dire à maximiser le bien pour la société dans son ensemble ;
- l'*altruisme* tend à accorder une plus grande importance aux conséquences affectant les autres agents.

Une action pouvant avoir de multiples conséquences bonnes et mauvaises, cette approche peut également mener à des dilemmes moraux qui sont la conséquence de combinaisons de règles morales et de connaissances sur les conséquences des actions. Les philosophes cherchent alors à déterminer s'il est juste d'effectuer une action pour ses conséquences morales, malgré ses conséquences immorales.

**Exemple 4.** *Thomas D'Aquin, par exemple, propose un principe appelé* Doctrine du double effet *désignant comme juste une action qui répond à quatre critères :*

1. *l'action est en elle-même morale ou neutre (au regard des règles morales) ;*
2. *l'agent ne doit pas désirer les conséquences immorales et les conséquences morales recherchées ne peuvent être obtenues sans ces conséquences immorales ;*
3. *les conséquences morales sont une conséquence directe de l'action, tout comme les conséquences immorales, c'est-à-dire que les conséquences morales ne doivent pas découler des conséquences immorales ;*
4. *La moralité des conséquences recherchées doit au moins compenser l'immoralité des conséquences immorales.*

*Ce principe est encore aujourd'hui celui qui permet de justifier le droit à la légitime défense : il est possible de faire usage de violence, au risque de blesser l'agresseur (ce qui constitue la conséquence immorale) dans le but de protéger soi-même ou un tiers (ce qui constitue la conséquence morale) uniquement si cette violence est inévitable (second point de la doctrine du double effet) et proportionnelle à l'attaque (quatrième point de la doctrine).*

#### I.1.2.5 Dilemmes éthiques

De manière analogue à la morale, certaines situations peuvent conduire à l'incapacité, pour l'agent, de distinguer l'action la plus juste d'un choix qui lui est présenté au regard d'un principe éthique.

**Définition 7** (Dilemme éthique)**.** *Un* dilemme éthique *est un choix entre plusieurs options qu'un principe éthique évalue comme justes ou injustes de manière égale.*

Ces cas rares sont souvent employés dans le débat philosophique pour montrer qu'un principe éthique ne suffit pas à départager en toute situation les actions disponibles.

**Exemple 5.** *Benjamin Constant, au cours de débats qui l'ont opposé à Kant et à l'idéalisme allemand (Constant, Kant, 2003 ; Constant, 2013), a proposé la description de la situation suivante : « Un agent A est caché chez un agent B pour échapper à un agent C, et C vient demander à B où se trouve A pour le tuer ». S'il est autant inacceptable pour la société que les personnes ne se portent pas mutuellement assistance ou qu'elle aient recours au mensonge, B est incapable de juger s'il doit révéler ou non la cachette de A en employant l'impératif catégorique de Kant (voir exemple 3). Cette situation est alors un dilemme si l'impératif catégorique est le seul principe éthique employé par le jugement. Si un autre principe était employé, par exemple la doctrine du double effet (voir exemple 4) en considérant la mort de A plus immorale que le mensonge, cette situation n'est plus un dilemme.*

Afin d'éviter les situations de dilemme éthique et les situations d'indécision, une dernière étape de raisonnement est nécessaire pour employer les principes éthiques et, en fonction de l'évaluation des actions à disposition, déterminer l'action juste.

#### I.1.2.6 Le jugement

L'emploi d'un ou plusieurs principes éthiques pour déterminer la justesse d'une action est appelé *jugement*. La faculté de jugement est au cœur de l'éthique et constitue l'étape finale pour prendre une décision éthique en évaluant chaque choix au regard de ses désirs, sa morale, ses capacités et principes éthiques. En accord avec quelques définitions consensuelles (*Ethical Judgment*, 2015) et les concepts précédemment évoqués, nous considérons la définition suivante de jugement :

**Définition 8** (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. Il est possible de juger une action *a priori*, c'est-à-dire avant de faire l'expérience de son exécution au regard des connaissances dont on dispose sur les conséquences possibles de cette action, ou *a posteriori*, c'est-à-dire une fois que celle-ci est effectuée.

Le jugement peut, comme toute action, faire l'objet de règles morales. Ainsi certaines valeurs comme la prudence, l'indulgence ou l'intransigeance peuvent décrire de bonnes et de mauvaises manières de juger (Hill, 2010) selon les informations à disposition, l'inclusion ou non de règles propres à l'agent jugé. Un agent juge ayant une conception relativiste de la morale pourrait ainsi trouver immoral de juger l'action d'un autre agent sans prendre connaissance de la morale de ce dernier.

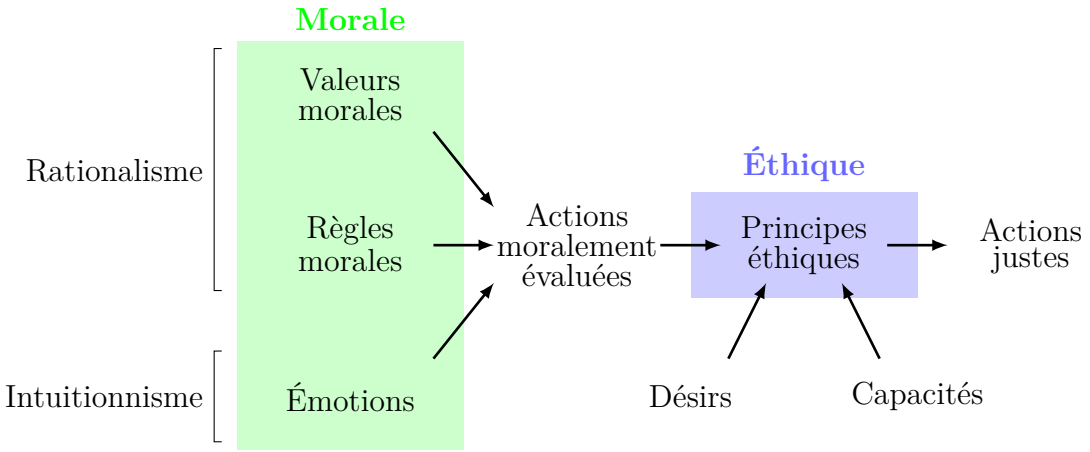


FIGURE I.2 – Représentation synthétique des concepts employés dans le jugement

La figure I.2 présente l'ensemble des concepts de philosophie évoqués dans cette section. Nous avons défini la morale comme étant ce qui permet de produire des évaluations d'actions, c'est-à-dire d'affecter aux actions des valuations de moralité. La morale peut être exprimée à l'aide de règles morales (dans un cadre rationaliste) ou avoir recours à des émotions (dans un cadre intuitionniste). Les évaluations peuvent s'appliquer à l'ensemble des êtres humains (universalisme) ou à une population donnée, une culture, une communauté (relativisme). De même, ces évaluations peuvent définir la moralité des actions de manière absolue, c'est-à-dire indépendamment de la situation (absolutisme) ou de façon restreinte à un ensemble de contextes (contextualisme).

L'éthique juge ensuite les actions justes au regard des valuations morales affectées aux actions, ainsi que des désirs et capacités de l'agent. Ce jugement a recours à un ou plusieurs principes éthiques décrivant des raisonnements issus de conceptions variées de ce qui est juste ou injuste. Ces principes peuvent représenter des approches vertueuses, accordant une grande importance aux valeurs promues ou trahies par les actions, ou bien des approches déontologiques, privilégiant le respect strict des règles morales par l'action elle-même, ou encore des approches conséquentialistes pour lesquelles les conséquences des actions importent davantage que les actions en elles-mêmes.

La suite de ce chapitre a pour but de montrer comment des travaux issus d'autres disciplines telles que la neurologie, les sciences cognitives, la sociologie et la psychologie, ont cherché à explorer ces divers éléments et montrer ce qui est commun ou différent entre les êtres humains.

## I.2 En neurologie, psychologie et sciences cognitives

L'apparition de modèles et de méthodes pour observer et comprendre le fonctionnement de l'esprit humain a rapidement mené à de nouvelles conceptions du fonctionnement de la morale et de l'éthique. Nous abordons dans cette section les principales théories de psychologie, de neurologie, et de sciences cognitives qui tentent d'élaborer un modèle du fonctionnement du jugement éthique.

### I.2.1 Organisation de la cognition éthique

L'une des premières préoccupations de la *neuro-éthique*, domaine s'intéressant au fonctionnement de l'éthique d'un point de vue neurologique, est de localiser les ensembles de neurones impliqués dans le jugement et montrer le lien entre ces régions cérébrales et d'autres fonctions cognitives déjà identifiées. Ces travaux ont pour cela eu recours à deux principales méthodes, la première d'un point de vue historique est l'étude de cas d'individus ayant subi des lésions suivies de troubles cognitifs (Damasio, 2008), mettant en évidence le rôle de ces aires cérébrales par la différence de comportement avec un individu normal, la seconde s'appuyant sur des techniques modernes d'imagerie médicale pour observer l'activité du système nerveux en fonctionnement dans diverses situations.

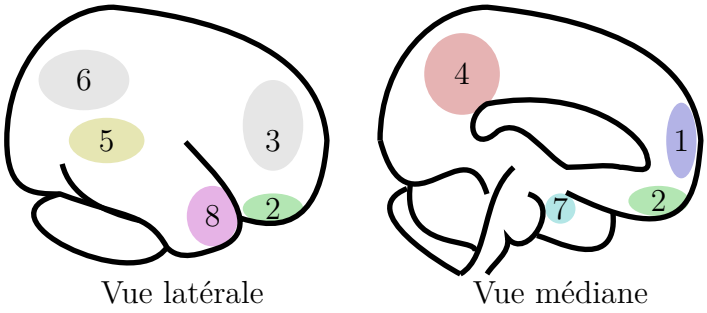


FIGURE I.3 – Localisation des aires cérébrales impliquées dans le jugement (Greene, Haidt, 2002)

La figure I.3 montre l'emplacement des aires cérébrales dont l'implication dans le jugement a fait l'objet d'études. La table I.1 décrit leur implication dans divers processus cognitifs. Ces illustrations sont une synthèse de plusieurs travaux d'état de l'art dans ce domaine (Greene, Haidt, 2002 ; Young, Dungan, 2012). La première constatation évidente est que cette faculté de jugement n'est pas localisée en un endroit unique, ni même sur un ensemble de zones regroupées dédiées à cette fonction (Parkinson *et al.*, 2011 ; Young, Dungan, 2012). En revanche cette faculté de jugement va, selon la situation, activer des aires cérébrales intervenant dans d'autres facultés (essentiellement émotionnelles et sociales), montrant que le jugement semble faire appel à ces capacités.

Certains spécialistes de la cognition éthique distinguent deux types de jugements en fonction du contexte (Greene, Haidt, 2002) : si l'action jugée implique de causer suffisamment de mal à un individu en particulier, sans que cela ne soit justifiable par un autre effet de l'action permettant de protéger quelqu'un, alors le jugement est qualifié de *jugement personnel*, sinon il est qualifié de *jugement impersonnel*. Ici le premier

| Région | Rôles dans le jugement | Autres rôles connus | Pathologies dues aux lésions |
|---|---|---|---|
| 1. Cortex préfrontal | Jugement personnel et impersonnel, capacité à pardonner (Farrow *et al.*, 2001), vision d'images évoquant des émotions et des actes immoraux (Harenski, Hamann, 2006) | Attribution d'intentions et d'états mentaux à des personnages fictifs ou historiques (théorie de l'esprit), attribution d'émotion à des visages, réminiscence d'épisodes heureux/malheureux, prise de décision dans un contexte émouvant (Partiot *et al.*, 1995) | Prises de décisions sur le long terme irrationnelles, réactions agressives et impolies, diminution de l'empathie et des savoir être sociaux |
| 2. Cortex orbitofrontal et ventromédian | Jugement simple | Punition/récompense, souvenir d'épisode triste, Reconnaissance du contexte positif des mots, reconnaissance des expressions faciales colériques, sommeil | Prises de décisions sur le long terme irrationnelles, réactions agressives et impolies, diminution de l'empathie et des savoir être sociaux, difficultés avec des taches complexes de théorie de l'esprit |
| 3. Cortex préfrontal latéral | Réaction face à la malhonnêteté (Parkinson *et al.*, 2011) | | |
| 4. Gyrus cingulaire postérieur et précuneus | Jugement personnel et impersonnel, rancune, réaction face aux attitudes menaçantes (Parkinson *et al.*, 2011) | Écoute d'épisodes autobiographiques émouvants, menaces verbales, lecture d'histoires et vision de dessins, en particulier usant de la théorie de l'esprit, familiarité des visages, expression faciales de dégout, tristesse, vision de combats, réminiscence de souvenirs tristes (hommes) ou heureux, reconnaissance de mots neutres dans un contexte négatif, planification en contexte non-émotionnel, sommeil | Altération de la reconnaissance des visages familiers, peut-être syndrome de Capgras |
| 5. Gyrus angulaire | Jugement personnel, jugement simple, images morales | Visualisation de mouvements de membres (mains, visage, yeux, corps), visages tristes, films joyeux, tristes et dégoutants, dessins faisant appel à la théorie de l'esprit, attribution d'intentions et d'états mentaux à des personnages fictifs ou historiques, reconnaissance de mots neutres dans un contexte négatif, mémoire de paires de mots, vue et mémoire de films émouvants, sommeil | Altération du jugement à partir du regard chez le singe, peut-être syndrome de Capgras |
| 6. lobule pariétal | Jugement impersonnel, réaction face à la malhonnêteté (Parkinson *et al.*, 2011) | Mémoire de travail et autres tâches cognitives | |
| 7. Amygdale | Reconnaissance de situations immorales (Moll *et al.*, 2002), dégout (Parkinson *et al.*, 2011) | Reconnaissances d'images et de films émouvants, reconnaissance d'expressions faciales, évaluation rapide par punition/récompense, particulièrement suite à des stimulus visuels et négatifs | Jugement social pauvre à partir des expressions faciales et corporelles |
| 8. Pôle temporal | Jugement simple | Lecture d'histoires cohérentes avec personnages, attribution d'intentions et d'états mentaux à des personnages fictifs ou historiques, souvenir de lieux et visages connus, souvenir d'épisodes autobiographiques chargés d'émotions, reconnaissance d'émotions représentées, reconnaissance d'expressions faciales colériques ou tristes, vue et souvenir de films heureux, dégoutants et tristes | Altération de la mémoire autobiographique |

Ce tableau est une version retravaillée et actualisée de celui présenté par (Greene, Haidt, 2002).

TABLE I.1 – Fonctions connues des aires cérébrales impliquées dans le jugement

type de jugement s'applique aux actes considérés comme délibérément immoraux, sans contradiction ou dilemme possible. Les situations telles que la légitime défense sont en revanche de l'ordre du second type de jugement. Ces deux types de jugements semblent faire intervenir des ensembles d'aires cérébrales en partie distincts.

Notons enfin que le *cortex préfrontal ventro-médian* (incluant, au moins partiellement les zones 1, 2 et 3 sur la figure I.3 et dans le tableau I.1) est parfois appelé *cerveau émotionnel* tandis que le *carrefour temporo-pariétal* (incluant, au moins partiellement les zones 4, 5 et 6 sur la figure I.3 et dans le tableau I.1) est parfois désigné sous l'appellation de *cerveau social* (Young, Dungan, 2012). Les sections I.2.2 et I.2.3 fournissent une explication à cette distinction et montrent l'influence et le rôle de ces aires dans le jugement.

### I.2.2 Rôle des émotions et du raisonnement

Après avoir observé des patients aux comportements altérés suite à des lésions du cortex préfrontal ventro-médian, certains neurologues (Damasio, 2008) ont montré que ces aires cérébrales, manquantes ou non fonctionnelles chez ces patients, intervenaient dans le fonctionnement émotionnel et certains types de jugements, principalement ceux concernant les choix à long terme impactant la vie des patients (orientation professionnelle, gestion du patrimoine, etc.) ainsi que sur leur rapport avec les autres (politesse, pudeur, fierté, etc.). Les personnes présentant ces lésions semblent pleinement capables de raisonner et comprendre les raisonnements d'autrui sur leur attitude, mais tendent à prendre des décisions anormales en comparaison des individus sains. Ces études montrent que l'incapacité à ressentir certaines émotions chez ces patients (embarras, honte, culpabilité, fierté, etc.) lorsqu'ils envisagent les conséquences des choix à leur disposition entraine une évaluation erronée sans que ceux-ci ne s'en rendent compte au premier abord, même s'ils ont la connaissance de normes sociales réprouvant ces attitudes. La conception cartésienne purement rationnelle du jugement (voir section I.1.1), comme étant uniquement le fruit d'un raisonnement sur des règles et principes connus est alors remise en question.

Afin de concilier le rôle des émotions et celui du raisonnement logique, certains résultats de travaux tels que l'approche *socio-intuitionniste* (Haidt, 2001) proposent des modèles faisant intervenir à la fois une réaction émotionnelle et un raisonnement rationnel dans le jugement. La figure I.4 présente le modèle social-intuitionniste du jugement dans lequel un individu $A$, lorsqu'il perçoit une situation va (1) produire un jugement intuitivement, puis (2) raisonner sur ce jugement intuitif. (3) Ce raisonnement peut être communiqué à une personne $B$, (4) de même que le jugement. (5) Le raisonnement peut parfois amener la personne à revenir sur son jugement voire (6) générer une nouvelle intuition. Ce modèle étend le modèle intuitif en ajoutant une dimension sociale : $B$ procède de la même manière à la production d'états mentaux qu'il peut à son tour échanger avec $A$.

Ce type de modèle permet d'expliquer pourquoi une personne peut formuler un ju-

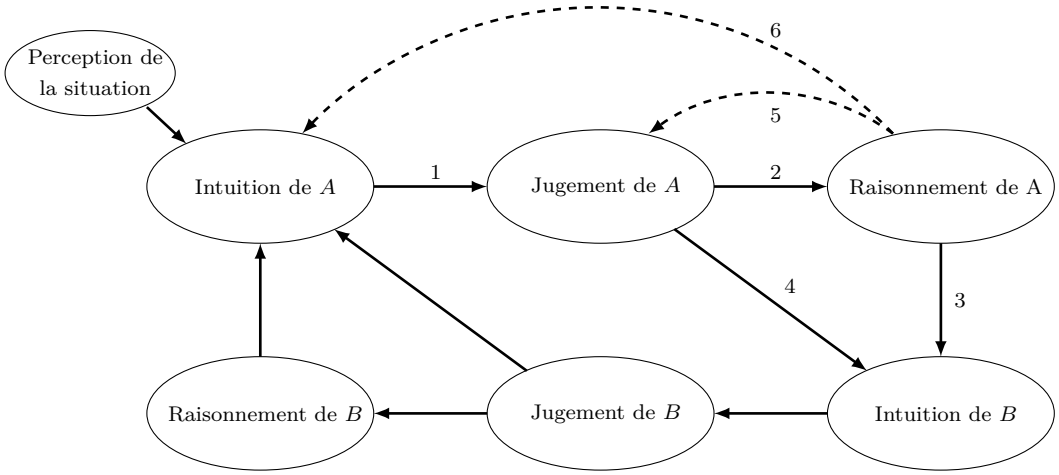


FIGURE I.4 – Modèle social-intuitionniste du jugement (Haidt, 2001)

gement *a priori* intuitivement et rapidement, puis modifier son jugement suite à un raisonnement plus long et complexe. Du point de vue de la biologie et de la théorie de l'évolution, cela est cohérent puisque les zones telles que l'amygdale (zone 7 sur la figure I.3), située dans la région centrale du système nerveux, sont capables de produire très rapidement des émotions (peur, dégoût, etc.) (Moll *et al.*, 2002) pouvant déclencher des réactions instinctives salutaires en cas de danger.

Quelques études approfondies de la cognition morale (Cova, 2014) montrent de nombreuses propriétés de la production intuitive de jugements. Par exemple, de simples variations de la perception de situation peuvent induire d'importantes différences de jugement. Certains travaux (Patil *et al.*, 2014) montrent par exemple que le degré de réalisme des situations auxquelles sont exposées les personnes lors de ces expérimentations ont une influence sur l'issue de jugements rapides.

Enfin, certaines personnes peuvent être incapables de trouver un raisonnement permettant de justifier un jugement intuitif (Haidt *et al.*, 2000). Ce sont alors des convictions, des valeurs personnelles impossibles à étayer par des arguments rationnels.

### I.2.3 Facultés sociales et théorie de l'esprit

Si le *cerveau émotionnel*, situé à l'avant de l'encéphale joue un rôle essentiel pour répondre à des stimulus émotionnels, le *cerveau social* ou plus rigoureusement *carrefour temporo-pariétal* (regroupant les zones 4, 5 et 6 de la figure I.3 et de la table I.1), est fortement sollicité pour interpréter les expressions corporelles, faciales et verbales d'un individu réel ou fictif (Saxe, Kanwisher, 2003).

Ces aires cérébrales, chez l'individu normalement développé, sont impliquées dans un processus cognitif appelé *théorie de l'esprit*, permettant d'inférer les intentions, croyances et désirs d'une autre personne (Premack, Woodruff, 1978). Cette faculté permet de se placer mentalement « à la place de l'autre » afin de reproduire les processus mentaux à l'origine de son comportement et, dans une certaine mesure, d'en inférer les possibles

actions futures. Par exemple, ces aires sont impliquées dans l'observation et l'interprétation d'un comportement malhonnête ou menaçant (Parkinson *et al.*, 2011), nécessitant de comprendre le but d'une action, d'un mouvement ou d'une expression corporelle chez l'individu observé.

La théorie de l'esprit peut également entraîner la production d'émotions dues à cette mise en situation pour comprendre l'autre (Bzdok *et al.*, 2012). Ce phénomène est appelé empathie et peut activer les aires cérébrales responsables du ressenti émotionnel (Farrow *et al.*, 2001). L'observateur peut alors ressentir ce qu'il suppose que l'autre ressent, même si cet autre est fictif (je comprends la détresse de Bambi), qu'il s'agit d'un être incapable de telles intentions ou émotions (ce chat a compris qu'il m'est difficile de rédiger ma thèse et cherche à me réconforter), ou parfois d'un objet (ce robot-phoque en peluche m'aime (Marti *et al.*, 2005), ce robot démineur mérite notre reconnaissance pour son courage et son sens du sacrifice (Tisseron, 2015)).

Empathie et théorie de l'esprit sont ainsi indispensables au bon fonctionnement du raisonnement moral puisqu'ils sont au cœur des processus mentaux permettant de comprendre et se soucier d'autrui. Certaines pathologies touchant ces aires cérébrales peuvent supprimer la compréhension du mal causé à autrui (Blair, 1995), supprimant la souffrance causée par empathie et qui jouait le rôle d'inhibiteur contre les comportements violents, nuisibles ou gênants. Si l'éthique et la morale ont pour but de conduire au bonheur et l'absence de trouble selon les philosophes d'Athènes, l'empathie et la théorie de l'esprit peuvent alors justifier un comportement altruiste puisque faire souffrir l'autre fait aussi souffrir l'agent, et faire du bien à l'autre fait aussi du bien à l'agent.

## I.3 En sciences sociales

La section I.2 a montré comment les notions définies à la section I.1 sont explicables par des phénomènes psychologiques et neurologiques relativement identiques d'un individu, sain et normalement développé, à un autre. Cependant, la totalité de la population ne parvient pas au même résultat et, à ces capacités innées issues de l'anatomie de notre système nerveux, viennent se mêler des éléments acquis liés à notre vécu personnel, notre culture et notre éducation.

Cette section vient achever cet état de l'art de l'éthique et de la morale humaines en exposant des travaux de sciences sociales qui montrent la diversité des formes de raisonnement et les différences de valeurs entre individus.

### I.3.1 Théorie du développement moral

En sections I.2.2 et I.2.3, nous avons montré que la plupart des adultes emploient dans leur jugement à la fois des facultés de raisonnement, des intuitions issues de la perception d'émotions et une théorie de l'esprit permettant de comprendre et adopter le point de vue d'un autre. Cependant toutes ces facultés ne sont pas opérationnelles dès la naissance et ne se développent pas de la même manière.

La *théorie du développement moral* développée par Kohlberg (Kohlberg, Hersh, 1977)

propose une décomposition en niveaux successifs du développement du jugement chez l'Homme. La table I.2 récapitule ces étapes en mentionnant la tranche d'âge majoritairement concernée par ces stades de développement (J. Snarey *et al.*, 1983).

| Age | Stade | Niveau |
|---|---|---|
| Enfance | Préconventionnel | 1. *Punition et obéissance* : jugement fondé sur l'apprentissage des effets de l'action. Seules les conséquences physiques de l'action comptent (absence de notion de respect de l'autorité ou de son incarnation).<br>2. *Intérêt personnel* : la seule valeur des actions est instrumentale. Débuts de notions d'équité, de réciprocité et de partage. Donnant-donnant. |
| Adolescence | Conventionnel | 3. *Concordance interpersonnelle* : le bon comportement est celui approuvé par les autres. La personne est en recherche de conformité avec un modèle.<br>4. *Maintient de l'ordre social* : attirance pour l'autorité, les règles fixées, le devoir moral et le maintien de l'ordre établi par la société. |
| Age adulte | Postconventionnel | 5. *Contrat social* : moralité établie après un avis critique. Conscience du relativisme des valeurs personnelles, attirance pour la recherche de consensus. La dimension légale compte mais doit être accordée à la morale et à l'utilité pour la société.<br>6. *Recherche de principes éthiques universels* : recherche d'objectivisme, de principes s'appliquant à tout être humain dans le respect de sa dignité et de son individualité. |

TABLE I.2 – Récapitulatif de la théorie du développement moral de Kohlberg (Kohlberg, Hersh, 1977)

L'âge indiqué est en grande partie indicatif : un enfant précoce peut atteindre un niveau de développement moral supérieur rapidement et, à l'inverse, seul un adulte américain sur vingt atteindraient le sixième stade (Dien, 1982).

Ce travail, bien que largement reconnu par les spécialistes du développement mental, a fait l'objet de nombreuses critiques, l'accusant de considérer à tort un point de vue nord-américain comme universel (Dien, 1982 ; J. R. Snarey, 1985).

Pour concevoir un système dans lequel des agents autonomes peuvent être amenés à interagir avec des humains, il est nécessaire de prendre en compte la diversité des formes de raisonnement éthiques utilisés et compréhensibles par ces individus.

### I.3.2 Les valeurs morales en sociologie

Comme nous avons montré dans la section précédente la coexistence de diverses approches éthiques au sein d'une même population, nous nous attachons ici à montrer l'existence d'une diversité morale. Différents travaux proposent d'étudier l'importance d'ensembles de valeurs morales en fonction du temps, du genre, de la situation sociale, de l'éducation, de l'âge (Rokeach, 1974) ou de la nationalité (Hofstede, Bond, 1984 ; Schwartz, 1992).

Le sociologue Shalom Schwartz (Schwartz, 2006) attribue six caractéristiques au concept

de valeur :

- **Les valeurs sont des croyances** indissociables des affects. Lorsqu'elles sont activées, elles se combinent aux sentiments.
- **Les valeurs ont trait à des objectifs désirables** qui motivent l'action.
- **Les valeurs transcendent les actions et les situations spécifiques**. Ce sont des concepts moins restreints que des normes ou des attitudes.
- **Les valeurs servent d'étalons ou de critères** pour la sélection ou l'évaluation des actions, des politiques, des personnes et des événements. On décide de ce qui est bon ou mauvais, justifié ou illégitime, de ce qui vaut la peine d'être fait ou de ce qui doit être évité en fonction des conséquences possibles pour les valeurs que l'on affectionne. Mais l'impact des valeurs sur les décisions de tous les jours est rarement conscient, et le devient lorsque des actions ou des jugements nous conduisent à des conflits entre les valeurs que l'on affectionne.
- **Les valeurs sont d'importance variable** les unes par rapport aux autres d'un point de vue personnel.
- **L'importance *relative* de multiples valeurs guide l'action.** Une action supporte plusieurs valeurs.

Le nombre et le nom des valeurs varie selon les études mais elles sont supposée exister en nombre fini, et être présentes dans toute culture, en variant seulement en importance (Schwartz, 1992). Les valeurs peuvent être organisées au sein d'un *système de valeurs*, que nous définissons de la manière suivante :

**Définition 9** (Système de valeurs)**.** *Un système de valeurs est constitué d'un ensemble de valeurs, structuré par un ensemble de relations hiérarchiques et de relations d'opposition, auxquelles une population accorde des importances plus ou moins grandes.*

Le modèle proposé par (Schwartz, 1994) semble être le plus efficace pour représenter les écarts culturels entre populations puisque les distances calculées entre systèmes de valeurs sont plus fortement corrélées à des données observables par exemple en sciences économiques (Imm Ng *et al.*, 2007).

La figure I.5 est une représentation graphique de l'ensemble organisé de valeurs établi par (Schwartz, 1994) sur lequel sont représentées les dix *valeurs fondamentales* que sont l'accomplissement, le pouvoir, l'auto-détermination, la stimulation, l'hédonisme, la sécurité, la conformité, la tradition, la bienveillance et l'universalisme. Chacune de ces valeurs fondamentales se décline en une multitude de valeurs plus spécifiques : l'indulgence, l'amitié, la serviabilité sont autant de formes spécifiques de la valeur fondamentale de bienveillance.

Cette représentation rend aussi compte de relations entre valeurs fondamentales : ces valeurs sont en relation de conflit avec certaines autres. Ainsi, par exemple, promouvoir la sécurité ou la tradition est en conflit avec l'auto-détermination. Cela ne signifie pas que l'une est le contraire de l'autre, mais que chaque personne situe sa préférence d'un côté ou de l'autre.

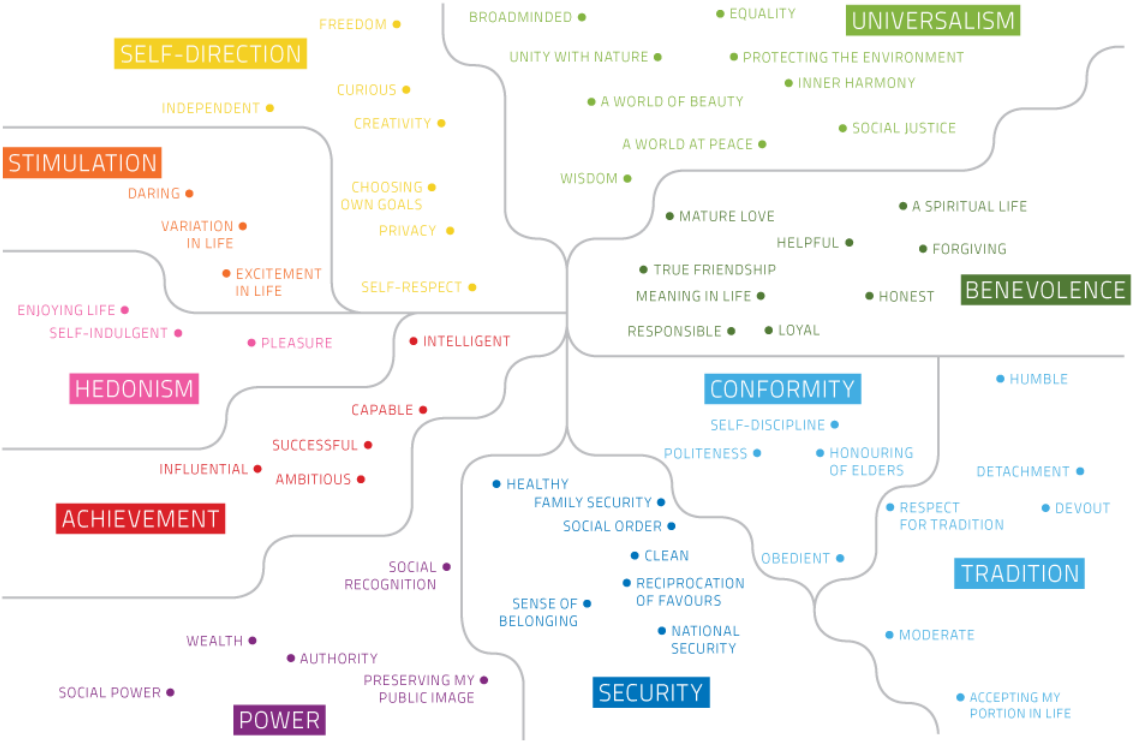


FIGURE I.5 – Représentation graphique du modèle de valeurs de Schwartz (Schwartz, 1994)

Des travaux employant ce modèle (Inglehart, Welzel, 2005) ont montré comment la culture de pays (ou d'ensembles de pays) peuvent être positionnés sur les axes que constituent ces valeurs en conflit afin de représenter une forme de mesure de proximité morale et culturelle.

## I.4 Synthèse

Nous avons commencé par présenter dans ce chapitre les principaux concepts employés en philosophie pour juger les actions. En premier lieu, nous avons montré comment est évaluée la moralité des actions, c'est-à-dire le fait qu'elles soient bonnes ou mauvaises au regard de règles morales, de valeurs morales et d'émotions. Ensuite, l'éthique permet de juger de la justesse des actions au regard d'un ou plusieurs principes éthiques, de désirs et de connaissances des capacités de l'agent afin de déterminer quelle action ou ensemble d'actions il est juste d'effectuer dans une situation.

Nous avons ensuite montré comment des travaux de neurologie et de psychologie nous permettent de comprendre les processus mentaux à l'œuvre lors du jugement, en distinguant des fonctions mentales sollicitées différemment selon l'implication ou non d'émotions, ou la nécessité de comprendre les intentions d'un autre à travers une interprétation de ses expressions corporelles et verbales (théorie de l'esprit, voir section I.2.3). La ré-

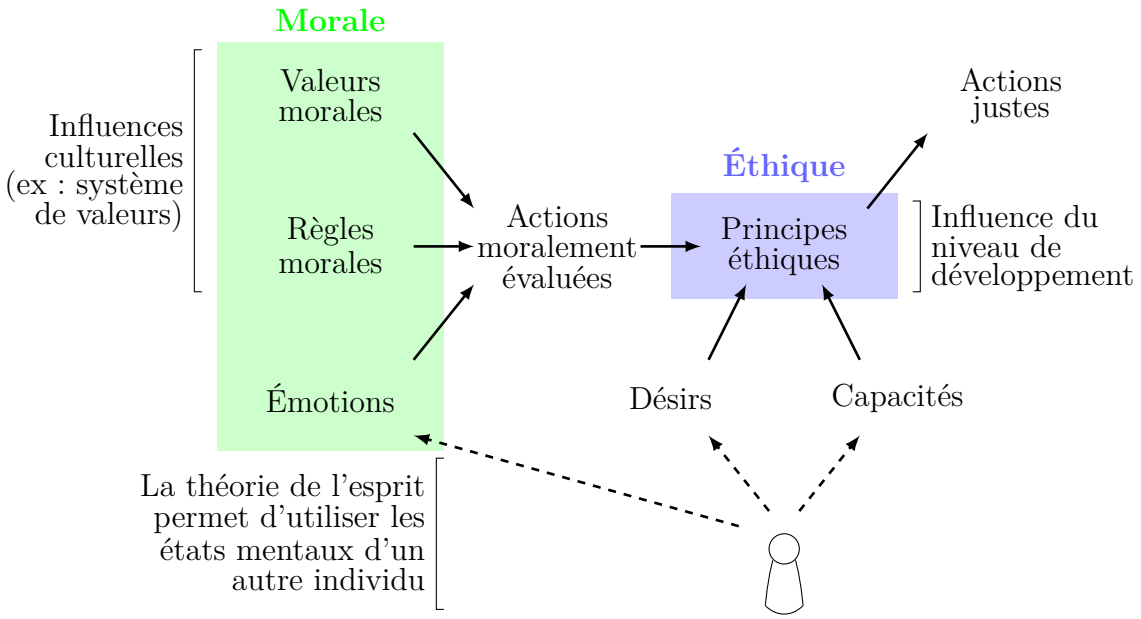


FIGURE I.6 – Représentation synthétique de la diversité des théories du bien et du juste.

ponse émotionnelle, plus rapide, semble offrir un premier jugement que l'humain tend à expliquer ou remettre en question par un raisonnement rationnel. La théorie de l'esprit permet à un individu d'introduire dans son processus de jugement ce qu'il pense être les états mentaux d'un autre individu qu'il observe afin de reproduire ce qu'il suppose être le raisonnement de l'autre.

Enfin nous avons montré comment les facultés mentales humaines sont utilisées de manières différentes par les individus (voir figure I.6) pour produire un jugement éthique reposant sur des éléments variés tels que des normes sociales ou des principes universels (voir section I.3.1). La morale employée dans ces jugements est également variable selon les populations et les influences culturelles, avec par exemple des variations significatives dans l'importance accordée à certaines valeurs morales (voir section I.3.2).

Dans le cadre de la réalisation de l'objectif fixé en introduction, il est intéressant de pouvoir montrer qu'il est possible de prendre en compte ces connaissances sur le bien et le juste pour configurer des agents autonomes et influencer leur processus de prise de décision. Une telle caractéristique pourra être justifiée soit par le fait qu'ils se trouvent en interaction avec des humains, soit qu'ils agissent au nom de personnes qui leur délèguent des prises de décision dans un système artificiel. Faire coexister dans un même système des agents capables de se juger respectivement en prenant en compte des états mentaux d'autres agents de manière analogue à la théorie de l'esprit (voir section I.2.3) pourrait être rendu possible par la conception d'un processus de raisonnement générique faisant abstraction de la diversité des morales et des éthiques qui leur sont données en paramètres.

Dans le cadre de la problématique de cette thèse, nous cherchons alors à permettre de représenter et utiliser au sein d'un agent artificiel les éléments du jugement éthique que

sont la théorie du bien, comprenant les valeurs morales et règles morales, et la théorie du juste, employant les principes éthiques (voir figure I.6), et utiliser ce jugement pour orienter son comportement et interagir avec les autres agents.

# CHAPITRE II
# Approches existantes

**II.1 Problèmes éthiques dans les systèmes multi-agents . . . . . 28**
II.1.1 Le jugement éthique pour la décision . . . . . 28
II.1.2 Influence de l'éthique sur les interactions entre agents . . . . 29
**II.2 Grille de lecture . . . . . 31**
II.2.1 Représentations implicites ou explicites des éthiques . . . . . 31
II.2.2 Approche rationaliste ou intuitionniste . . . . . 32
II.2.3 Généricité . . . . . 32
II.2.4 Caractère opérationnel . . . . . 33
**II.3 Approches procédurales . . . . . 33**
II.3.1 Ethical governor . . . . . 33
II.3.2 Value sensitive design . . . . . 35
**II.4 Approches numériques . . . . . 36**
II.4.1 Case Supported Principle-Based Paradigm . . . . . 36
II.4.2 Jugement par évaluation d'expressions . . . . . 38
**II.5 Approches déclaratives . . . . . 39**
II.5.1 Représentation logique de principes éthiques . . . . . 39
II.5.2 Morale et logique déontique . . . . . 40
II.5.3 Responsabilité morale, éthique et causale . . . . . 41
II.5.4 Jugement dans une architecture BDI . . . . . 42
II.5.5 Argumentation formelle avec des valeurs morales . . . . . 45
**II.6 Synthèse . . . . . 46**

Le chapitre I a présenté des définitions et des modèles concernant un ensemble de concepts philosophiques touchant à l'éthique et à la morale. Ce chapitre a pour but de comparer des travaux d'intelligence artificielle présentant des modèles ou des méthodologies permettant de concevoir des *agents autonomes éthiques*. Nous mettons ensuite en évidence les limites de ces approches afin de présenter les propriétés attendues de notre proposition.

La section II.1 met en évidence des problématiques d'éthique computationnelle soulevées par le fait de représenter dans le processus de décision des agents autonomes les concepts présentés au chapitre précédent. Puis nous proposons en section II.2 un ensemble de critères, permettant de positionner les travaux vis-à-vis de critères découlant de ces problématiques afin de mettre en lumière les différences et similitudes entre les réponses proposées. Les sections II.3, II.4 et II.5 présentent l'ensemble de ces travaux, en les positionnant par rapport aux critères précédemment présentés. Enfin la synthèse de ce chapitre propose une vision globale de cet état de l'art et expose un raffinement des objectifs de cette thèse en mettant en évidence les écarts entre les travaux de l'état de l'art d'une part et les propriétés attendues du jugement éthique d'un agent d'autre part. Cette synthèse s'appuiera sur la grille de lecture pour identifier les propriétés d'un modèle idéal.

## II.1 Problèmes éthiques dans les systèmes multi-agents

Nous identifions dans cette section deux problématiques distinctes : la première est l'utilisation du jugement comme un moyen de guider la décision de l'agent. L'agent doit, pour pouvoir effectuer ce jugement, être capable de manipuler des concepts de morale et d'éthique définis au chapitre I tels que des valeurs morales, règles morales et principes éthiques, spécifiés par un utilisateur ou le concepteur du système. Il doit pouvoir raisonner sur ces éléments dans les diverses situations auxquelles il est confronté, avec les spécificités liées au domaine applicatif dans lequel il est déployé. La seconde problématique est celle de l'emploi de ce même processus de jugement pour guider ses interactions avec les autres agents. Dans le cadre de tâches à déléguer ou à effectuer en coopération, il lui est nécessaire de juger le comportement des autres agents afin de déterminer lesquels ont un comportement compatible avec sa propre éthique. Cette seconde problématique est particulièrement importante dans le cadre de systèmes multi-agents ouverts et distribués dans lequel des agents peuvent être tenus de respecter des éthiques issues d'utilisateurs différents.

### II.1.1 Le jugement éthique pour la décision

Le chapitre précédent a montré les divers éléments qui, chez l'être humain, semblent intervenir dans le jugement. Donner à des agents autonomes la possibilité de manipuler des représentations de ces concepts leur permettrait de juger les actions à leur disposition lors de chaque prise de décision. L'un des objectifs de notre démarche est de permettre aux agents dotés d'un tel processus d'être déployés dans tout domaine applicatif et d'employer des théories morales et éthiques diverses décrites avec des approches différentes.

Afin de faciliter le déploiement de tels agents, il est important que le modèle de jugement ne nécessite pas d'être à nouveau implémenté lors de chaque déploiement. Pour cela, il est nécessaire de concevoir un modèle de jugement indépendant du domaine applicatif, et de considérer l'ensemble des connaissances spécifiques à ce domaine comme

un élément donné en paramètre du modèle sur lequel viendra s'appuyer le raisonnement. De même, la morale de l'agent, bien que dépendante du domaine applicatif (puisque les règles morales et valeurs morales qualifient des actions dans des contextes liés à ce domaine applicatif), doit pouvoir être changée sans modifier le modèle de jugement ou les connaissances du domaine. Ainsi, concevoir des agents avec des morales variées nécessite uniquement de changer les ensembles de connaissances sur les règles morales et valeurs morales, sans devoir modifier le modèle de jugement ou les connaissances du domaine. De même, l'éthique des agents devrait pouvoir être modifiée indépendamment des autres ensembles.

Enfin, une partie de la communauté académique s'intéressant à la propriété d'« explicabilité » des décisions, nous considérons nécessaire de rendre possible le fait d'expliquer un jugement éthique, d'autant plus si cet agent peut être amené à interagir avec des utilisateurs humains.

### II.1.2 Influence de l'éthique sur les interactions entre agents

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. Nous considérons alors les problématiques liées à un système où coexistent et évoluent des agents dotés d'éthiques individuelles variées.

Nous traitons en section II.1.2.1 des problématiques touchant au jugement du comportement d'un autre agent lors de son observation afin de construire une représentation mentale de sa conformité à une théorie du bien et du juste. Puis la section II.1.2.2 aborde les problématiques liées à l'usage de ces représentations dans les interactions entre agents afin de permettre des coopérations.

#### II.1.2.1 Jugement du comportement des autres

Afin de tenir compte du comportement des autres agents dans leurs interactions, les agents autonomes éthiques doivent être capables d'acquérir une représentation de l'éthique des autres, soit par construction à partir de l'observation de leur comportement, soit par transmission directe d'agent à agent.

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. 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, nous pouvons imaginer que les agents soient capables de communiquer aux autres leur éthique et leurs croyances. 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.

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 (voir section I.2.3) ne pourrait-il pas comparer le comportement de l'agent observé et sa propre conduite s'il s'était trouvé dans des conditions similaires ? 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.

L'action épistémique (qui a pour conséquence de mettre à jour les croyances de l'agent, qu'il faut distinguer des actions ontiques ayant pour conséquence un changement du monde (Van Dimarsch *et al.*, 2007)) 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'observations d'un comportement comme insuffisant pour se prononcer sur son caractère éthique.

#### II.1.2.2 Usage du jugement pour la coopération

Une fois un agent doté d'une représentation de l'éthique d'un autre agent du système, comment peut-il l'utiliser pour décider et agir ? L'agent doit pouvoir disposer d'une action interne lui permettant d'évaluer cette éthique. Notons que cette évaluation ne qualifie pas le comportement de l'autre agent de bon ou mauvais dans l'absolu, mais ne fournit qu'une comparaison de ce comportement observé et celui que son jugement aurait évalué comme juste. Cette évaluation peut 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, peut inférer les règles morales et le principe éthique de l'autre agent sans pour autant y adhérer et en déduire que, dans certaines situations, leur éthique diffère.

L'emploi d'une représentation de l'éthique d'un autre agent devrait permettre à l'agent juge de savoir si une action est non seulement éthique de son point de vue, mais également du point de vue de l'autre, et raisonner sur les conséquences de son exécution sur les relations de confiance et de collaboration établies avec les autres agents. Par exemple, un agent autonome éthique peut se trouver confronté à un choix entre une action juste de son point de vue individuel mais injuste pour les agents avec lesquels il coopère et une action injuste de son point de vue mais juste pour les autres. L'éthique dans un système composé de tels agents soulève alors des problématiques de dimension sociale

impliquant pour l'agent de prendre en compte cette dimension dans son raisonnement individuel.

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. Cela implique également qu'il est possible de définir une éthique de la coopération, définissant à partir de valeurs morales, règles morales et principes éthiques en quoi il est juste et bon, par exemple, d'accorder sa confiance, de faire preuve d'indulgence ou de coopérer.

## II.2 Grille de lecture

Nous proposons ici un ensemble de cinq critères découlant des problématiques liées au jugement éthique des agents et énoncées à la section précédente. Ces critères ont pour principal objet de permettre de positionner les divers travaux présentés dans ce chapitre les uns par rapport aux autres afin de montrer comment et dans quelle mesure ils permettent d'apporter une réponse aux problèmes de la représentation de l'éthique dans les agents autonomes évoqués dans la section précédente.

### II.2.1 Représentations implicites ou explicites des éthiques

Comme nous l'avons mentionné en section II.1, nous cherchons à concevoir des agents pouvant recevoir leur éthique d'un utilisateur ou d'un concepteur du système, capables de raisonner sur leur propre éthique et éventuellement de la transmettre à d'autres. De telles fonctionnalités nécessitent de pouvoir manipuler des représentations de l'éthique et de la morale et d'en distinguer les composants. De même, pour expliquer le résultat d'un jugement, l'agent doit être capable d'exhiber les éléments ayant conduit à ce résultat.

Lorsque le concepteur introduit dans l'implémentation du processus de décision de l'agent des contraintes et limitations pour produire un comportement conforme à une morale et une éthique, nous parlons de représentation implicite. Le principal inconvénient d'une telle proposition est l'impossibilité pour l'agent de raisonner sur la portée morale et éthique de son comportement pour le justifier. Il n'est alors juste qu'en raison de son incapacité à prendre des décisions injustes du point de vue du concepteur.

Le raisonnement sur des connaissances déclaratives grâce à un ensemble de déductions logiques offre la possibilité de concevoir des agents capables de mettre en évidence l'enchaînement causal produit par leur raisonnement et le présenter à un humain ou d'autres agents (Boella, Torre, 2003). Cela demande de représenter symboliquement la morale et l'éthique pour permettre à l'agent d'inclure ces connaissances dans le raisonnement, et de faire en sorte que le processus de décision tienne compte de ces éléments.

### II.2.2 Approche rationaliste ou intuitionniste

Nous distinguons ensuite les travaux *intuitionnistes* faisant reposer tout ou partie du jugement sur une simulation d'émotions au sein de l'agent, des travaux *rationalistes*, qui écartent toute notion émotionnelle du jugement pour ne reposer que sur des principes éthiques et des valeurs et règles morales.

Les modèles intuitionnistes offrent l'avantage de permettre de modéliser l'intervention des émotions dans le jugement humain (voir section I.2) et de concevoir des systèmes dans lesquels les agents peuvent effectuer des jugements irrationnels. Ces modèles de simulations d'émotions permettent par exemple d'enrichir la communication avec un utilisateur en suscitant chez lui de l'empathie pour l'agent (Marti *et al.*, 2005 ; Foster *et al.*, 2015).

Les modèles rationalistes, quant à eux, excluent l'influence d'un modèle émotionnel sur le jugement et ne tiennent compte que de connaissances sur le contexte, de valeurs morales, de règles morales et de principes éthiques. Une telle approche peut permettre de s'assurer que l'agent s'en tient à des règles de conduite définies. Cela favorise également l'explicabilité du système (voir section II.1.1) en permettant de résumer un jugement à un ensemble de déductions logiques à partir de connaissances.

Défendant l'idée de spécificités éthiques propres aux agents artificiels, (Malle *et al.*, 2015) apportent des éléments mettant en évidence le fait que les humains attendent des agents autonomes un jugement différent du jugement humain. Dans cette étude les auteurs montrent que, face au dilemme du trolley (voir section I.1.1.3), un observateur humain a tendance à blâmer un autre humain en cas d'action (sacrifier une vie pour en sauver cinq) et un agent autonome artificiel en cas d'inaction (laisser mourir cinq personne lorsqu'il a la possibilité de n'en tuer qu'une). Ces résultats sont interprétés comme une attente des humains envers les agents autonomes. Ces derniers étant dépourvus d'émotions, les humains semblent majoritairement préférer qu'ils adoptent une éthique utilitariste.

Notons toutefois qu'une approche rationnelle n'exclut pas la possibilité de prendre en compte dans le jugement des connaissances sur les émotions que peuvent susciter des actions ou des contextes sur d'autres agents ou des utilisateurs humains. Il s'agit alors de connaissances faisant partie du modèle du monde propre à l'agent.

### II.2.3 Généricité

Certaines propositions permettent de concevoir des agents éthiques indépendamment du domaine applicatif où ils seront déployés, bien qu'il puisse être nécessaire de donner à l'agent des connaissances en paramètre.

À l'inverse, les agents conçus selon certaines approches ne sont adaptés qu'à un domaine spécifique. Tout changement de domaine entraîne alors un nouveau travail de conception.

Nous distinguons ici les approches dites *génériques par rapport à la morale*, s'attachant

à décrire comment toute morale peut être prise en compte par un agent autonome, des approches *spécifiques à une morale* proposant la conception d'un agent conforme à une morale spécifique. Concevoir des agents dotés d'une architecture générique capable d'employer toute morale donnée en paramètre est justifiable d'un point de vue relativiste de la morale (voir section I.1.1) puisque cela permet de représenter dans un même système des agents dotés de mêmes facultés de raisonnement mais dont le jugement se fonde sur des morales différentes.

De manière analogue à la généricité par rapport à la morale, nous distinguons les approches dites *génériques par rapport à l'éthique*, s'attachant à décrire comment toute éthique peut être prise en compte par un agent autonome, des approches *spécifiques à une éthique* proposant la conception d'un agent conforme à une éthique qui est directement implémentée dans le processus de décision.

Notons qu'une approche peut être générique par rapport à l'éthique ou la morale et spécifique par rapport à l'autre.

### II.2.4 Caractère opérationnel

Plusieurs travaux présentés ici proposent une réflexion théorique et des modélisations de haut niveau de ce que devrait être un agent autonome éthique, ou bien ne détaillent qu'une partie de leur proposition. D'autres s'attachent à décrire une architecture opérationnelle et vont parfois jusqu'à démontrer son fonctionnement à l'aide d'une preuve de concept. Afin d'éprouver la modélisation du jugement, il est nécessaire de proposer une architecture d'agent opérationnelle et de confronter ce raisonnement éthique à des situations réalistes.

## II.3 Approches procédurales

L'ordre choisi ici pour ces ensembles d'approches est de présenter les approches manipulant des représentations implicites de l'éthique (appelées approches procédurales) en premier pour nous diriger vers des approches plus symboliques par la suite. Sans prétendre être exhaustif, cet état de l'art a pour but de montrer au lecteur comment divers cadres conceptuels d'intelligence artificielle ont jusqu'ici été proposés en éthique computationnelle. L'ensemble des approches procédurales regroupe ici des démarches ayant en commun de chercher à définir lors de la conception des agents l'éthique de leur comportement en toutes situations, puis d'implémenter ces agents de manière à ce qu'ils n'aient d'autre possibilité que de respecter cette éthique.

### II.3.1 Ethical governor

Un exemple typique de ces approches est détaillé dans le rapport technique *Governing lethal behavior : Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture* (Arkin, 2009), dans lequel l'auteur montre comment concevoir un drone militaire armé dont le comportement suit un grand nombre de règle issues du droit de la guerre et de règles d'engagement de l'armée américaine. Dans le cadre de ce travail, l'auteur

montre le vaste ensemble de règles à prendre en compte et en déduit les contraintes à implémenter dans un agent.

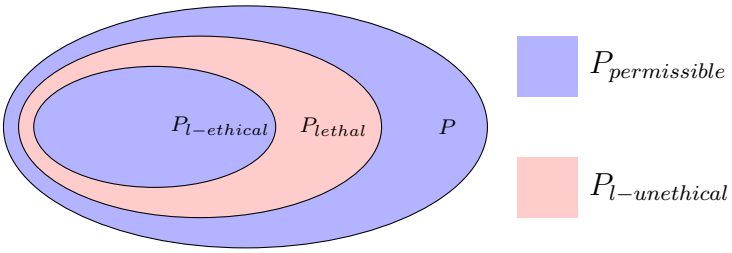


FIGURE II.1 – Espace des actions selon (Arkin, 2009)

La figure II.1 reprise de ce document illustre la démarche exposée par l'auteur, où $P$ est l'ensemble des actions possibles dans un certain contexte ; $P_{lethal} \subseteq P$ est l'ensemble des actions de $P$ ayant un caractère létal (ce sont les seules actions qui ne sont pas moralement neutres dans ce contexte applicatif selon l'auteur), et $P_{l-ethical} \subseteq P_{lethal}$ est le sous-ensemble des actions létales considérées comme éthiques (réciproquement, $P_{l-unethical} \subseteq P_{lethal}$ tel que $P_{lethal} = P_{l-ethical} \cup P_{l-unethical}$ est le sous-ensemble des actions létales non-éthiques). La proposition est alors une architecture qui, à l'aide d'un ensemble de contraintes $C$, empêche l'agent d'effectuer des actions appartenant à $P_{l-unethical}$.

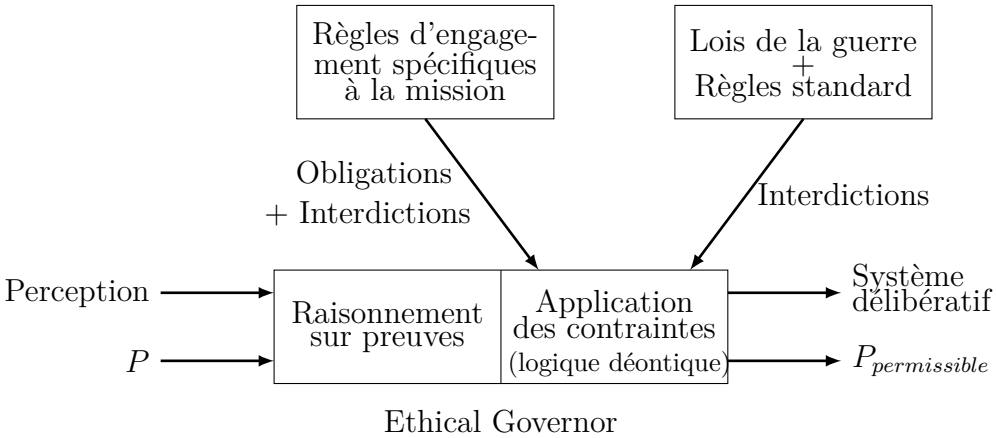


FIGURE II.2 – Architecture de l'Ethical Governor (Arkin, 2009)

Cette distinction entre actions éthiques et non-éthiques est opérée par un composant de l'architecture proposée nommé *Ethical Governor* (voir figure II.2) exécutant un algorithme mêlant des considérations spécifiques au domaine (par exemple l'optimisation de proportionnalité de la frappe ou la vérification de la dangerosité de l'objectif). L'ensemble des comportements possibles $P$ pris en entrée est fourni par un *contrôleur de comportement* évaluant la validité de leurs préconditions dans la situation perçue, et l'ensemble $P_{permissible}$ des comportements permissibles produit en sortie est envoyé aux actuateurs. Ce composant prend en compte des règles morales, exprimées en logique déontique, pouvant changer selon les missions (cas des règles d'engagement spécifiques à la mission sur la figure) ou rester pratiquement identiques d'une mission à l'autre (cas des lois et règles standard).

Ce modèle s'appuie sur une approche rationaliste et déontologique cherchant à appliquer au mieux un ensemble de règles, soit représentées symboliquement, soit implémentées directement dans le processus décisionnel. Cette proposition est spécifique au domaine et seule une partie de la morale peut être redéfinie (les règles d'engagement de la mission). L'éthique de l'agent est ici implémentée directement dans le processus de l'*Ethical Governor* permettant de pondérer et conférer un degré de priorité aux divers éléments de la morale et des impératifs opérationnels de l'agent. Cette proposition constitue bien une architecture d'agent opérationnelle, uniquement conçue dans le but de prendre des décisions éthiques et ne permet pas en l'état de fournir un jugement sur le comportement d'un autre agent.

### II.3.2 Value sensitive design

Le *Value sensitive design* (Friedman, 1996 ; Friedman *et al.*, 2013 ; Aldewereld *et al.*, 2015) constitue une approche que nous considérons comme similaire au sens où cela consiste également en une réflexion préalable, par des experts du domaines applicatif, pour définir l'éthique de l'agent afin de guider l'implémentation pour obtenir un logiciel dont le comportement est conforme à ces attentes. Il s'agit d'une méthodologie permettant de prendre en compte des valeurs morales (voir section I.3.2) dans la conception d'un logiciel. Cette démarche est constituée de trois parties :

— L'*investigation conceptuelle* a pour but d'analyser le besoin afin d'identifier les parties prenantes, les valeurs impliquées, les conflits entre ces valeurs et les compromis à faire entre ces valeurs (anonymisation vs. confiance, autonomie vs. sécurité, etc.).

— L'*investigation empirique* utilise des méthodes qualitatives et quantitatives pour observer, mesurer et documenter les activités des parties prenantes afin de découvrir d'éventuelles particularités et préciser les connaissances des concepteurs sur le système.

— L'*investigation technique* se focalise sur les technologies en elles-mêmes afin de déterminer comment elles permettent de concevoir un système répondant aux attentes énoncées lors de l'investigation conceptuelle.

La littérature fournit un grand nombre d'heuristiques, de précisions méthodologiques, de valeurs à examiner et de cas d'exemples (Friedman *et al.*, 2013 ; Aldewereld *et al.*, 2015). Cette approche jouit en outre d'un certain succès, comme en témoignent les nombreux workshops et publications liés à cette communauté (Friedman *et al.*, 2015).

Cette méthodologie permet de prendre en compte de manière explicite, dans le procédé de conception, des valeurs morales qui vont guider l'implémentation du logiciel en orientant le choix et la conception des fonctionnalités. Ces valeurs se retrouvent par la suite implicitement présentes dans l'exécution du logiciel, découlant non pas d'un raisonnement symbolique du programme final mais des choix pris en amont lors de sa création. Il est possible de prendre en compte à la fois une conception rationaliste de l'éthique, et des critères intuitionnistes (ressenti personnel des participants à la création ou des utilisateurs finaux). Une fois implémenté, le logiciel est spécifique à un domaine applicatif. La moralité et l'éthique de son comportement sont fixées et conformes aux spécifica-

tions définies lors des phases d'investigation. Cette méthodologie ne dit pas comment concevoir des logiciels amenés à juger le comportement d'autres logiciels. De plus, ces travaux étant des descriptions d'une démarche de conception, et non une proposition d'architecture, il n'est pas question ici d'en évaluer le caractère opérationnel.

## II.4 Approches numériques

Par *approches numériques*, nous désignons toutes les approches représentant les connaissances sous forme numérique. Par exemple, des propositions d'utilisation de techniques d'apprentissage ont été formulées afin de montrer comment des agents peuvent apprendre la morale et l'éthique à partir de données produite par des experts ou une population.

### II.4.1 Case Supported Principle-Based Paradigm

En ce sens, (Anderson, Anderson, 2014b) proposent le paradigme CPB (pour *Case-Supported Principle-Based Paradigm*), mêlant des techniques d'apprentissage et un mécanisme de raisonnement explicite sur les actions employant des représentations symboliques de préférences générées à partir de l'apprentissage. Les auteurs s'inspirent de la théorie du « devoir à première vue », ou *prima facie duties theory* (Ross, 1930 ; Canto-Sperber, Ogien, 2004).

La partie d'apprentissage, nommé *GenEth* (pour *General Ethical Dilemma Analyzer*) (Anderson, Anderson, 2014a), a pour but d'extraire des connaissances à partir d'évaluations d'actions en contexte par des experts (appelés *ethicists*). Pour décrire la moralité d'une action, ces experts affectent des pondérations de caractéristiques (*features*) reliant des actions à des devoirs moraux. Une pondération positive signifie que l'action supporte ce devoir et une pondération négative signifie qu'elle va à son encontre. Les experts évaluent également les différentes situations possibles afin d'affecter à chaque devoir moral de l'agent une pondération de devoir (*duties*) : une valeur positive signifie que l'accomplissement de ce devoir est primordial dans cette situation, une valeur négative signifie à l'inverse que ce devoir est sans importance.

Lorsque l'agent est confronté à une situation, son mécanisme de raisonnement compare les actions à disposition en examinant pour chacune le produit de leur pondérations de caractéristiques et de devoirs puis en calculant pour chaque couple d'actions la différences entre ces produits de pondérations. Le résultat permet alors d'établir une relation de préférence, accordant plus d'importance à celle qui obtient le meilleur score (c'est-à-dire celles qui satisfait le mieux les devoirs les plus importants dans la situation selon les experts). Les auteurs appellent « ethical principle » cette relation de préférence. Notons que, compte tenu de la distinction entre éthique et morale introduit en section I.1, l'appellation appropriée dans le cadre que nous proposons serait « préférence morale » puisqu'il s'agit du résultat de la comparaison de mesures d'adéquation entre les conséquences d'actions et des devoirs moraux. Au sens des définitions exposées au chapitre précédent, le principe éthique mis en œuvre par CPB considère comme action juste celle qui est préférée au sens de l'ensemble des relations de préférence calculées.

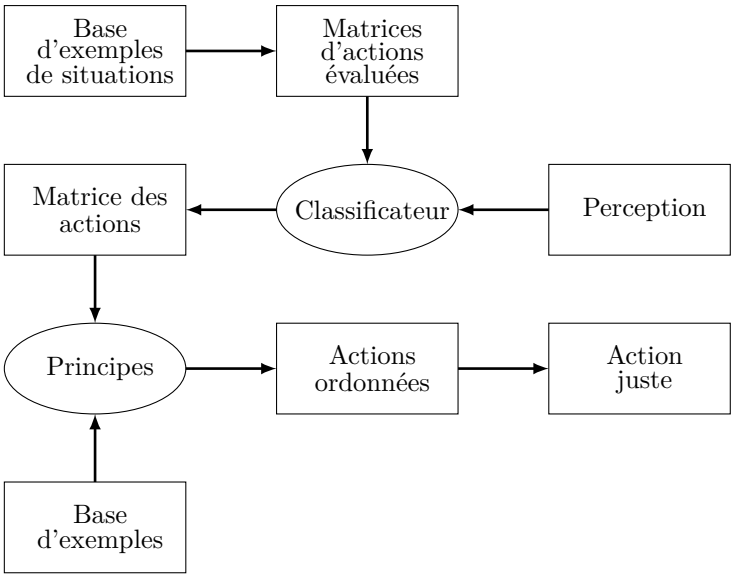


FIGURE II.3 – Architecture d'agent mettant en oeuvre CPB (Anderson *et al.*, 2017)

La figure II.3 présente l'architecture d'agent mettant en œuvre CPB (Anderson *et al.*, 2017) comme processus de décision. Le classificateur compare la situation perçue aux situations de la base d'exemples évalués par des experts. De cette comparaison résulte la matrice des actions dans laquelle une valeur entière est affectée à tout devoir pour toute action. Les préférences (nommées *principes* dans le cadre de CPB) sont alors générées en cherchant à maximiser l'adéquation entre les actions (dont on connait la pondération représentant leur capacité à satisfaire les devoirs grâce à la matrice des actions). L'ensemble des actions est ordonné par les préférences et l'action préférée, ou *action juste* au sens des définitions données en section I.1.2 est celle qui satisfait le mieux ces préférences.

Dans le cadre de CPB, la morale n'est pas explicitement représentée : seul le résultat du jugement des experts, sous forme d'une valeur numérique, est présente sans conserver d'information sur leur motivation. L'éthique de l'agent est fondée sur le choix de l'action répondant au mieux à la situation au regard des devoirs et de leur pondération. Cette approche permet de représenter des morales intuitionnistes : l'avis des experts peut tenir compte d'émotions suscitées par les situations envisagées. Enfin, l'architecture d'agent proposée est générique du point de vue du domaine applicatif (il faut seulement un moyen de décrire l'ensemble des états), générique du point de vue de la morale (les experts peuvent changer l'ensemble des critères d'évaluation, leur pondération dans l'ensemble des états, les caractéristiques des actions). Seul le principe éthique de calcul de l'action préférée reste invariable. Cette contribution répond au problème de la prise de décision éthique par la proposition d'une architecture opérationnelle accompagnée de résultats expérimentaux. Cependant, l'agent ne peut pas distinguer dans son environnement la présence d'autres agents, ni être capable de percevoir et juger leur comportement.

## II.4.2 Jugement par évaluation d'expressions

D'autres travaux proposent d'apprendre la morale par la compréhension de textes (Yamamoto, Hagiwara, 2014). Dans cette étude, les auteurs proposent un modèle étudiant la co-occurence de termes portants une connotation morale positive ou négative.

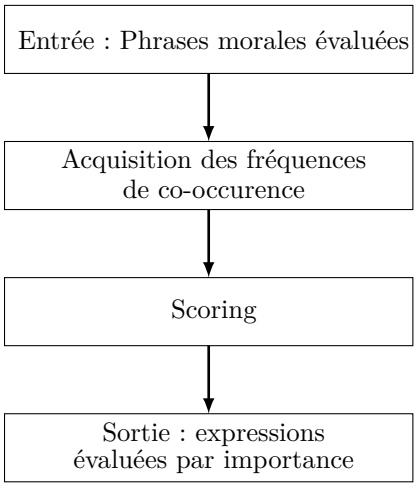


FIGURE II.4 – Phase d'apprentissage (Yamamoto, Hagiwara, 2014)

La figure II.4 illustre le comportement du système lors une phase d'apprentissage : ce processus prend en paramètre un texte dans lequel des phrases à connotation morales (par exemple : « ne triche pas, ne mens pas et vis dans l'obéissance » ) sont accompagnées d'une polarité donnée, c'est-à-dire une valeur numérique indiquant si cela décrit quelque chose de moral ou d'immoral. Ce processus produit un ensemble d'expressions évaluées, c'est-à-dire des couples de mots auxquels sont attribués des scores correspondant au nombre d'occurrences dans des phrases positivement évaluées moins le nombre d'occurrences dans des phrases négativement évaluées.

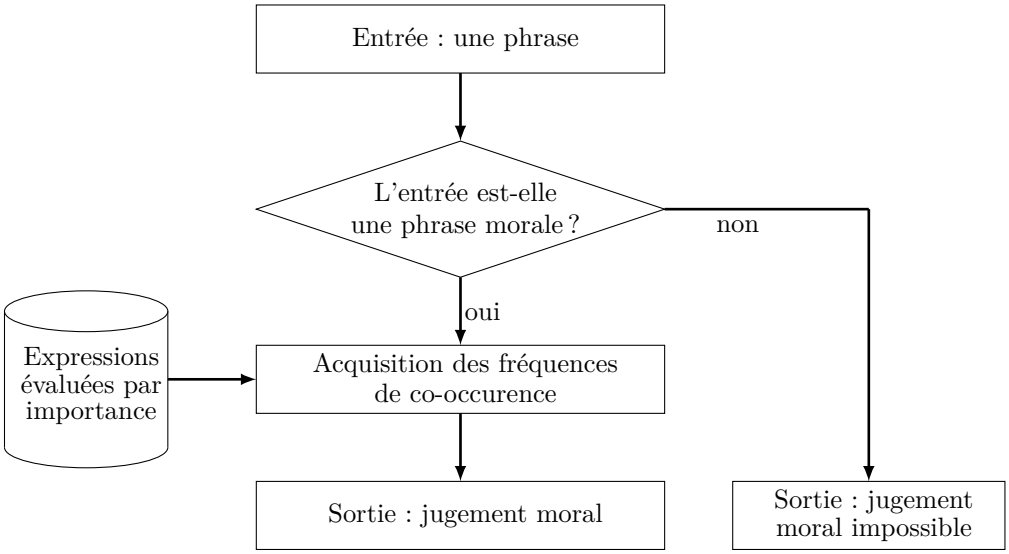


FIGURE II.5 – Phase de jugement (Yamamoto, Hagiwara, 2014)

La figure II.5 illustre le comportement de ce même système lors d'une phase de jugement. Premièrement, un test se fondant sur des règles grammaticales simples vérifie qu'il est pertinent de considérer que cette phrase peut avoir une quelconque portée morale (le sujet est un humain et le verbe de la phrase appartient à une certaine liste définie). Ensuite, la phrase se voit attribuer un score en employant les évaluations d'expressions obtenues lors de la phase d'apprentissage.

Dans ce système, la morale n'est pas explicitement représentée : elle n'apparaît que sous forme de valeurs numériques attribuées à des expressions, et est présente sans conserver d'information sur leur motivation. L'éthique du jugement repose dans le processus d'affectation de scores qui attribue un jugement fondé sur les connaissances morales apprises. Cette approche permet de représenter des morales intuitionnistes : l'avis des experts fournissant l'ensemble de phrases morales utilisées pour la phase d'apprentissage peut tenir compte d'émotions. Le système proposé est générique du point de vue du domaine applicatif (sur lequel porte le corpus de texte) et générique du point de vue de la morale (elle découle des évaluations des phrases du corpus). Seul le principe éthique de l'affectation de scores reste invariable. Le jugement produit n'est pas effectué dans le cadre d'une prise de décision ou du jugement d'un autre agent mais a pour but de restituer le sens d'un texte. L'architecture complète du logiciel a été implémentée et confrontée à des corpus pour démontrer son efficacité.

## II.5 Approches déclaratives

Cette section s'attache à décrire les travaux proposant une modélisation de tout ou partie du jugement sous forme d'un ensemble de règles logiques, tirant généralement partie de travaux philosophiques proposant des raisonnements rationnels.

### II.5.1 Représentation logique de principes éthiques

Pour représenter des théories philosophiques rationalistes, l'emploi de la programmation logique semble intuitivement adapté. Des travaux présentés en ce sens (Ganascia, 2007a ; 2007b ; Berreby *et al.*, 2015) proposent des implémentations de principes éthiques en langage de programmation logique. Les principes ainsi représentés tels que l'éthique d'Aristote, l'impératif catégorique de Kant (Ganascia, 2007b), l'objection de Benjamin Constant (Ganascia, 2007a) (voir exemple 5) et la doctrine du double effet (Berreby *et al.*, 2015).

**Exemple 6.** *Dans cet exemple tiré de (Ganascia, 2007b), l'éthique d'Aristote consiste à choisir parmi toutes les actions possibles, celle dont la pire conséquence est la moins immorale. Ce principe est formalisé en ASP de la manière suivante :*

```
act(P,A) :- action(A), not unjust(A).
unjust(A) :- action(A), action(AA),
  worst_consequence(A,C),
  worst_consequence(AA,CC),
```

```
  worse(C,CC).
worse(lie(P,Prop),A):-action(A),person(P),
  proposition(Prop), A!=lie(P,Prop), A!=murder.
worse(murder,A):-action(A),A!=murder.
```

*Au sens des définitions données en section I.1, les deux premiers prédicats décrivent un principe éthique : compte tenu de la moralité des actions, il décrit l'action qu'il est juste d'effectuer. Les deux prédicats suivants décrivent une morale, c'est-à-dire qu'ils positionnent les actions en décrivant ce qu'il est bon ou mauvais de faire. Dans cet exemple mentir est pire que toute autre action sauf le meurtre.*

Une telle approche permet de manipuler une représentation symbolique et explicite de l'éthique et de la morale sous forme de règles et de principes représentés dans un langage de haut niveau, assez proche du langage naturel. Les travaux évoqués ici ont tous été réalisés avec une approche rationaliste. Ces modèles de principes éthiques prennent en paramètre des connaissances sur la situation et sur la moralité des actions, interchangeables à volonté. En revanche chaque implémentation proposée ici ne modélise qu'un principe éthique unique. Il ne s'agit pas de propositions d'agents capable de manipuler plusieurs principes. Les principes implémentés jugent des actions, peu importe l'agent l'ayant ou envisageant de l'exécuter.

### II.5.2 Morale et logique déontique

La logique déontique (McNamara, 2014) offre également un formalisme adapté à la représentation de devoirs moraux (obligation de faire le bien). Des travaux montrent comment représenter l'éthique d'un agent avec ce type de représentation logique (Bringsjord *et al.*, 2006).

La proposition des auteurs s'appuie sur la logique déontique standard, ou SDL, dont la notation est introduite par (Chellas, 1980) et dont l'axiomatisation est proposée par les travaux de Yuko Murakami (Murakami, 2004). La logique déontique axiomatisée de Murakami (ou MADL) propose des notations pratiques pour décrire les devoirs d'agents :

1. $\ominus_\alpha P$ signifie que « l'agent $\alpha$ doit veiller à ce que $P$ » ;
2. $\Delta_\alpha P$ signifie que « l'agent $\alpha$ veille à ce que $P$ »

Dans (Bringsjord *et al.*, 2006 ; Bringsjord, Taylor, 2012), les auteurs appellent code d'éthique ce qui, au regard des définitions données en section I.1, correspond à ce que nous désignons sous le terme de *règle morale*. Si un code d'éthique $J$ est valable pour l'agent $\alpha$ (c'est-à-dire que $J$ est vrai), et que ce code implique qu'il est tenu de réaliser $\varphi$, nous pouvons représenter ce code d'éthique par $J \rightarrow \ominus_\alpha \varphi$.

Dans (Bringsjord *et al.*, 2016), une implémentation d'agents autonomes dotés de tels codes d'éthique montre que des agents dotés de codes différents peuvent les utiliser pour raisonner non seulement sur leur propres devoirs, mais également pour évaluer ce que leur devoir leur dicterait s'ils étaient confrontés à un choix face auquel un autre agent se trouve.

Ces travaux proposent une représentation explicite de l'éthique et de la morale sous forme de règles nommées et représentées dans un langage de haut niveau. L'intérêt d'une telle représentation est montrée dans (Bringsjord *et al.*, 2016), où l'utilisateur peut interagir en langage naturel avec l'agent et celui-ci peut identifier le ou les éléments de son éthique et de sa morale qui permettent ou vont à l'encontre de la demande de l'utilisateur. L'approche présentée ici est rationaliste, se reposant entièrement sur la formulation de devoirs et d'interdits. Les exemples présentés montrent que cette approche peut s'appliquer à divers domaines applicatifs et qu'il est facile de donner aux agents des morales différentes en leur attribuant des devoirs. L'éthique en revanche, au sens d'un ou plusieurs principes permettant de choisir l'action à effectuer au regard des devoirs de l'agent, est toujours le même : l'agent effectue en priorité les actions dictées par ses codes d'éthique et cette approche ne tolère pas de contradictions dans la morale (voir section I.1.1.3).

### II.5.3 Responsabilité morale, éthique et causale

Des travaux reprenant les idées présentées en section II.5.1 proposent une architecture d'agent capable de combiner un ensemble de principes éthiques dont la description est indépendante de la morale employée (Berreby *et al.*, 2017a ; 2017b).

Cette architecture est conçue spécialement dans le but de permettre à des principes éthiques conséquentialistes (voir section I.1.2.4) d'employer une version modifiée de l'Event Calculus (Shanahan, 1999) afin de raisonner sur la moralité de conséquences d'actions. L'agent raisonne indépendamment sur la moralité de conséquences possibles d'action d'une part et sur les conséquences d'autre part. La théorie du juste permet ensuite de déterminer l'action juste au regard des conséquences connues et de leur moralité. La proposition est illustrée par une implémentation en answer set programming organisée en modules indépendants prenant les connaissances spécifiques au domaine, à la morale et à l'éthique, comme des spécifications passées en paramètres (voir figure II.6).

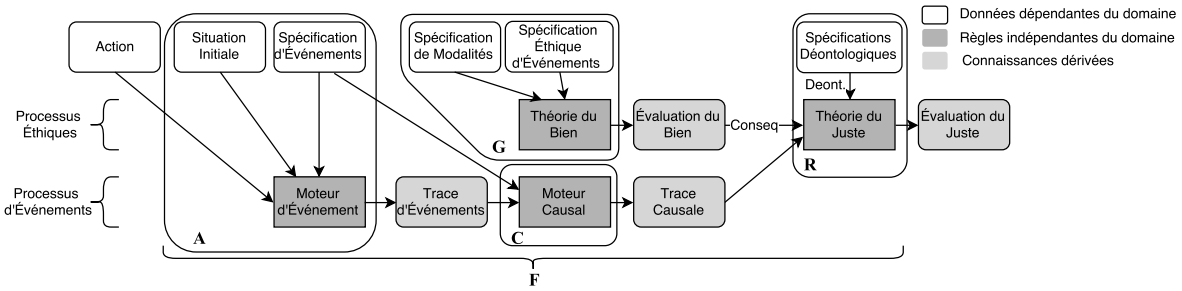


FIGURE II.6 – Cadre déclaratif modulaire de (Berreby *et al.*, 2017a ; 2017b)

Ce cadre d'évaluation éthique $\mathbb{F}$, est défini comme : $\mathbb{F} = \langle \mathbb{A}, \mathbb{C}, \mathbb{G}, \mathbb{R} \rangle$, où $\mathbb{A}$ est un *modèle d'action* permettant à l'agent de se représenter ses actions et son environnement ; $\mathbb{C}$ est un *modèle de causalité* permettant à l'agent de raisonner sur les conséquences des

actions et les notions de responsabilité ; $\mathbb{G}$ est un *modèle du bien*, permettant à l'agent d'évaluer la moralité de ses actions et $\mathbb{R}$ est un *modèle du juste*, permettant d'en évaluer la dimension éthique.

De manière similaire, (Pereira, Saptawijaya, 2017) propose un système de raisonnement logique sur les *situations hypothétiques* (*counterfactuals*) pour les agents autonomes. Dans cette proposition, l'agent peut envisager diverses situations et évaluer la justesse d'actions en fonction de leurs conséquences.

Cette approche déclarative propose un ensemble de règles permettant de construire un raisonnement entièrement rationnel fondé sur l'évaluation des conséquences des actions. Les spécificités du domaine, la morale et l'éthique de l'agent sont interchangeables à volonté dans (Berreby *et al.*, 2017a ; 2017b) et les auteurs illustrent cette généricité à l'aide de plusieurs exemples. Pour (Pereira, Saptawijaya, 2017), le raisonnement sur les situations hypothétiques est illustré dans divers cas applicatifs et des morales associées. Il montre l'utilité de ce modèle dans la représentation de principes éthiques conséquentialistes (la Doctrine du Double Effet et la Doctrine du Triple Effet). Ces modèles, décrivant un processus décisionnel, ne permettent pas en l'état de juger le comportement d'un autre agent sans disposer de la totalité de ses états mentaux nécessaires au fonctionnement de ces raisonnements.

### II.5.4 Jugement dans une architecture BDI

Cette section regroupe des propositions d'extensions du modèle *Belief Desire Intention* (ou BDI) afin de prendre en compte dans le raisonnement des éléments évoqués au chapitre I. Nous rappelons brièvement ce qu'est le modèle BDI puis nous exposons deux approches intégrant des systèmes moraux, émotionnels, et esthétiques.

#### Modèle BDI

L'architecture *BDI* (Bratman, 1987 ; A. S. Rao, Georgeff, 1991) propose d'organiser le raisonnement d'un agent autonome rationnel en distinguant les éléments suivants :

— l'ensemble $B$ des *croyances* (*Beliefs*) de l'agent constitue la représentation mentale des informations dont il dispose sur l'état du monde courant ;
— l'ensemble $D$ des *désirs* (*Desires*, parfois noté $G$ pour *goals*) de l'agent détermine les objectifs à atteindre ;
— l'ensemble $I$ des *intentions* de l'agent contient les actions que l'agent compte réaliser afin d'atteindre ses objectifs.

Le processus décisionnel de l'agent fait généralement appel à un ensemble $P$ de plans connus, c'est-à-dire de séquences d'actions qui, dans certaines situations (décrites par des conjonctions de croyances) permettent de satisfaire les désirs. Cette architecture permet de prendre en compte des événements grâce à des fonctions de perception et de communication venant actualiser les croyances et désirs de l'agent.

Cette architecture emploie volontairement une terminologie empruntée à la psychologie et la philosophie pour désigner ses composants et permettre une compréhension

globale du fonctionnement de l'agent à l'aide de définitions intuitives. De manière classique, un agent est dit *rationnel* lorsqu'il n'effectue des actions qu'en vue de satisfaire ses désirs. Le comportement d'un agent résulte de la succession des actions effectuées.

### Implication de valeurs dans le jugement

Pour chercher à répondre aux problèmes posés par la définition d'agents moraux, (Lorini, 2012) propose d'ajouter à l'architecture BDI un ensemble de valeurs morales définissant les devoirs et interdits moraux de l'agent. Les problèmes qui peuvent se présenter dans la prise en compte de ces valeurs dans la décision peuvent alors être de deux natures différentes : ils résultent soit de conflits entre les désirs et la morale, soit de conflits entre valeurs morales. Un certain paramétrage de l'éthique employée est proposé : des agents peuvent avoir une définition de l'utilité d'une action sous un angle purement hédoniste ou au contraire définir l'utilité exclusivement en fonction de la satisfaction morale apportée par l'action.

Dans ce travail, les auteurs s'attachent à montrer l'expressivité de cadre logiques existants (STIT et DL-MA) pour représenter une morale rationaliste sous la forme de règles. Cette approche ne s'applique pas à un domaine particulier et peut, dans la limite de ce qu'il est possible d'exprimer avec le langage proposé, permettre de représenter une grande variété de devoirs et interdits moraux. L'éthique proposée dans ces travaux est un principe reposant sur un calcul d'utilité, et d'une préférence à définir entre désir et morale. En l'état, cette proposition ne propose pas de jugement des comportements des autres et se focalise sur un point précis du raisonnement de l'agent qu'est celui du raisonnement sur la conciliation des désirs et des valeurs dans un cadre BDI opérationnel.

### Implication d'un modèle émotionnel dans le jugement

Certains travaux viennent ajouter à cette architecture des états émotionnels et décrivent leur intervention dans la décision éthique (Battaglino *et al.*, 2013). Les auteurs de cette proposition s'appuient sur l'idée d'un rôle indispensable des émotions dans le jugement, à la fois pour prendre conscience des enjeux moraux et éthiques, et ensuite pour justifier dans le raisonnement de l'agent la mise en avant de valeurs.

L'architecture proposée, illustrée en figure II.7, s'appuie sur un cycle de raisonnement de l'agent dans lequel viennent se mêler des étapes de raisonnement classiques des modèles BDI et des modifications d'un état émotionnel de l'agent. Les plans sont ainsi non seulement évalués pour les possibilités qu'ils offrent d'assouvir les désirs de l'agent, mais également pour les émotions qu'ils évoquent. De même, le changement d'état du monde perçu après exécution de l'action peut entraîner des modifications de l'état émotionnel. Les auteurs emploient à ces occasions des règles qui, selon l'ensemble des valeurs actives dans l'état courant et des désirs assouvis, peuvent changer l'état émotionnel.

Cette approche déclarative cherche à représenter de manière explicite les aspects émotionnels du jugement éthique. L'apport d'éléments intuitionnistes au modèle BDI a pour but, selon les auteurs, d'introduire une certaine empathie dans l'interaction avec d'autres agents, artificiels ou humains. La généricité de cette approche vis-à-vis du domaine ap-

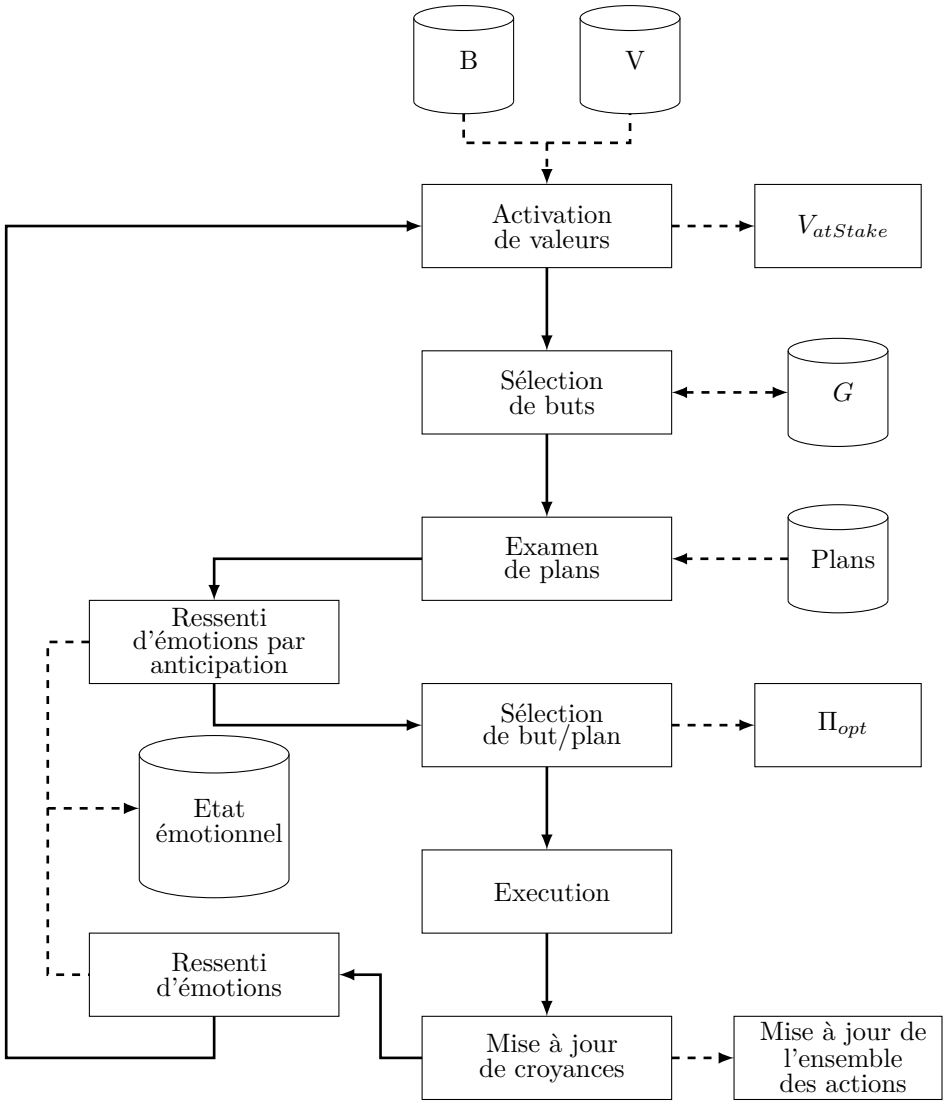


FIGURE II.7 – Architecture d'agent BDI émotionnel (Battaglino *et al.*, 2013)

plicatif et de la morale confiée à l'agent est assurée par la séparation entre l'architecture du processus décisionnel d'une part et les connaissances (valeurs morales, plans, buts, croyances) données en paramètre. L'éthique de l'agent, au sens des définitions que nous avons proposées, réside ici dans la manière dont les plans sont sélectionnés en fonction des émotions suscitées et de leur rentabilité en matière d'assouvissement des désirs. Ce mécanisme de décision ne semble pas paramétrable. Cette architecture offre la possibilité d'évaluer le comportement des autres au regard des émotions qu'il suscite.

## Considérations éthiques et esthétiques

Afin de représenter l'ensemble des éléments de la décision éthique, certains travaux tels que (Coelho, Rocha Costa, Trigo, 2010) proposent des architectures d'agent intégrant des connaissances sur la morale et l'éthique à un modèle classique BDI.

Dans un premier travail (Coelho, Rocha Costa, 2009), les auteurs avancent l'idée de

faire cohabiter au sein d'un agent autonome les éléments émotionnels et rationnels du jugement. Le jugement est, pour ces auteurs, un problème nécessitant plusieurs niveaux de raisonnement et résultant de la prise en compte d'éléments variés : émotions, valeurs et culture. Cette suggestion se concrétise dans (Coelho, Rocha Costa, Trigo, 2010) par la proposition d'une architecture d'agent intégrant une représentation de systèmes moraux, émotionnels et esthétiques au sein d'agents BDI (voir figure II.8).

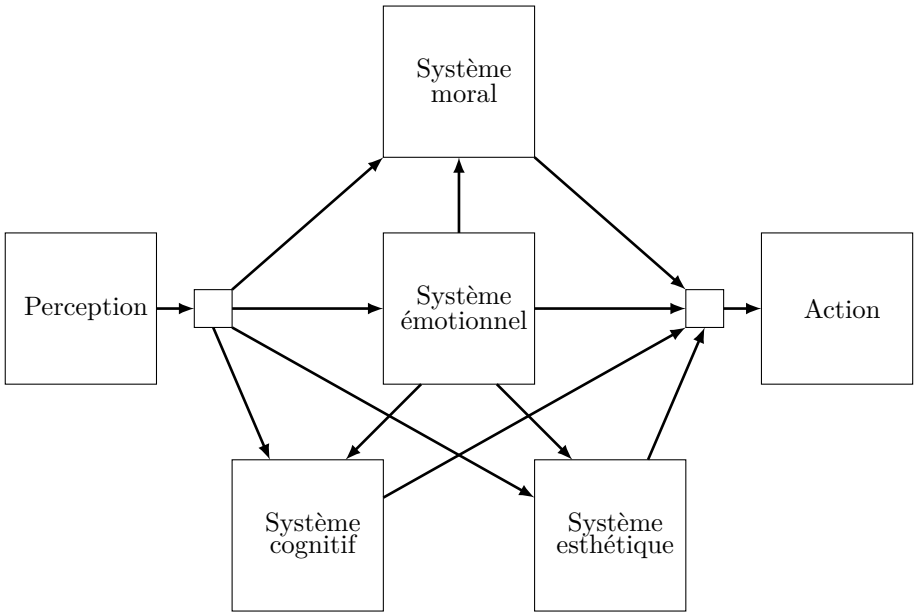


FIGURE II.8 – Proposition d'architecture d'agent moral (Coelho, Rocha Costa, Trigo, 2010)

Cette proposition cherche à capturer, de manière essentiellement explicite et depuis un niveau élevé d'abstraction, les différents éléments rationnels et émotionnels du jugement humain. Ce travail fait abstraction du domaine, de la morale et de l'éthique considérés et se concentre sur la mise en évidence des liens entre les systèmes décrits. L'éthique et la faculté de juger les autres sont présentées comme impactant le comportement de l'agent à l'échelle individuelle et l'ensemble du système à l'échelle collective. Ces travaux devraient être complétés par une proposition de mise en œuvre afin d'aboutir à une proposition de modèle pleinement opérationnel.

### II.5.5 Argumentation formelle avec des valeurs morales

Les travaux de (Atkinson, Bench-Capon, 2006) montrent que l'éthique pose des problèmes qui sont sujets à débat lorsque chaque action est simultanément soutenue et attaquée par des justification morales (au sens de la définition de la morale donnée en section I.1.1). Cette approche propose d'employer l'argumentation formelle comme une solution pour, non pas chercher à faire appliquer des normes par des agents, mais montrer comment des normes peuvent émerger d'interactions entre agents.

Dans une telle approche, les règles morales permettent de construire des relations d'attaque et de défense afin de déterminer quels choix sont indéfendables au regard des arguments proposés, et lesquels sont au contraire validés par la construction de l'ensemble

des relations. Les modèles d'argumentation permettent ensuite de représenter l'ensemble des arguments présentés et les relations d'attaques et de défenses entre ces arguments sous la forme d'un graphe orienté. Il est ensuite possible de déterminer selon divers critères quels sous-ensembles de ces arguments constituent un ensemble cohérent. Cette approche propose ainsi une manière d'évaluer les actions au regard d'une morale.

Cette approche rationaliste s'appuie sur une conception objectiviste de la morale (voir section I.1.1) et sur les mécanismes classiques d'argumentation pour, à partir de la représentation explicite de règles morales, trouver les actions défendables. Ce travail peut être appliqué à tout domaine applicatif et peut manipuler toute morale dont il est possible de tirer des arguments. Ce mécanisme représente en lui-même une forme spécifique de raisonnement éthique puisqu'il désigne comme juste l'action soutenue par les arguments formulés.

## II.6 Synthèse

Le tableau II.1 récapitule les principaux travaux cités dans les sections II.3, II.4 et II.5. Leur positionnement selon les critères définis en section II.2 et employés en conclusion de la présentation de chaque approche permet de mettre en évidence certaines de leurs caractéristiques. Ce tableau montre que les approches procédurales, proposant généralement d'introduire des contraintes dans l'implémentation de l'agent afin de produire un comportement, ne proposent pas (ou de manière partielle) d'architectures permettant de redéfinir la morale ou l'éthique de l'agent sans modifier le processus décisionnel. Les approches par représentation numérique permettent en revanche de confier aux agents des morales variées, représentées sous la forme de pondérations, même si l'éthique des agents reste figée dans le processus de raisonnement. Ces modèles numériques présentent généralement l'avantage de pouvoir construire cette représentation à partir d'un ensemble d'exemples évalués par des experts, les rendant de fait génériques par rapport à la morale. Enfin de nombreux travaux s'attachent à proposer des approches déclaratives permettant de définir divers modèles de jugement. Ces approches permettent de définir séparément le modèle de décision des connaissances employées, tant celles concernant l'environnement et les spécificités du domaine applicatif que celles qui décrivent la théorie du bien et la théorie du juste employées. Certains travaux conçoivent également le jugement comme un moyen non seulement de guider la décision de l'agent juge, mais également d'évaluer le comportement des autres agents.

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

| Approche | Travaux | Type de représentation | Rationalisme vs. intuitionnisme | Généricité | | | Jugement des autres | Opérationnel |
|---|---|---|---|---|---|---|---|---|
| | | | | domaine | morale | éthique | | |
| Procédurale | (Arkin, 2009) | implicite | rationaliste | non | partielle | non | non | oui |
| | (Friedman, 1996) | implicite | les deux | non* | non* | non* | non | - |
| | (Aldewereld *et al.*, 2015) | implicite | les deux | non* | non* | non* | non | - |
| Numérique | (Anderson *et al.*, 2017) | implicite | les deux | oui | oui | non | non | oui |
| | (Yamamoto, Hagiwara, 2014) | implicite | les deux | oui | oui | non | - | oui |
| Déclarative | (Ganascia, 2007a) | explicite | rationaliste | oui | oui | non | - | oui |
| | (Berreby *et al.*, 2015) | explicite | rationaliste | oui | oui | non | - | oui |
| | (Bringsjord *et al.*, 2016) | explicite | rationaliste | oui | oui | oui | non | oui |
| | (Berreby *et al.*, 2017a) | explicite | rationaliste | oui | oui | oui | non | oui |
| | (Pereira, Saptawijaya, 2017) | explicite | rationaliste | oui | oui | oui | non | non |
| | (Lorini, 2012) | explicite | rationaliste | oui | oui | non | non | oui |
| | (Battaglino *et al.*, 2013) | explicite | intuitionniste | oui | oui | non | oui | oui |
| | (Coelho,Rocha Costa, Trigo, 2010) | explicite | les deux | oui | oui | oui | oui | non |
| | (Atkinson, Bench-Capon, 2006) | explicite | - | oui | oui | non | - | non |

*Dans le cadre du *value sensitive design*, seule l'approche est générique. Le logiciel produit est en revanche incapable de mettre en œuvre une autre éthique ou morale dans un autre domaine applicatif.

Table II.1 – Tableau récapitulatif des travaux existants

Les agents ont besoin d'une représentation générique et explicite de la morale et de l'éthique afin de pouvoir manipuler les états mentaux d'autres agents lors du jugement de leurs actions, comme suggéré par la théorie de l'esprit en psychologie (voir section I.2.3 ). L'éthique des autres ne peut être comprise que par une représentation au sein de l'agent de leur éthique individuelle (Kim, Lipson, 2009). Afin d'exprimer et concilier un maximum de théories du bien et du juste, nous proposons de les représenter au sein de composants clairement définis reprenant les concepts définis au chapitre I : des *valeurs morales* caractérisant des actions dans certains contextes, des *règles morales* positionnant en bien ou mal des actions en fonction du contexte et des valeurs morales, des *principes éthiques* caractérisant en termes de juste ou d'injuste des actions et des *préférences éthique* permettant de définir l'importance de ces principes éthiques dans le jugement.

La généricité de ces représentations permettrait de dissocier la problématique de la définition du processus de jugement d'une part, et la description de la morale et de l'éthique de l'agent d'autre part. Une telle distinction permettrait de confier la configuration des agents par des non-spécialistes de l'intelligence artificielle tandis que l'implémentation des processus manipulant ces états mentaux pourrait être confiée à des non-spécialistes de l'éthique et de la morale. Cela simplifierait également les communications avec d'autres agents, y compris des humains.

Les agents ont besoin d'un modèle de jugement produisant un jugement explicable, c'est-à-dire pouvant être justifié à l'aide de déductions logiques reposant sur des connaissances sur le contexte, le domaine applicatif, et les représentations de la morale et de l'éthique conférées à l'agent. En accord avec les précédentes définitions, nous considérons le jugement comme une évaluation de la conformité d'un ensemble d'actions au regard d'un ensemble de valeurs et règles morales, que nous appelons *connaissance morale*, ainsi que de principes et préférences éthiques, que nous appelons *connaissance éthique*. Ainsi, des agents capables de se construire une représentation des connaissances morales ou éthiques des autres agents peuvent être en mesure de les juger de manière informée. Enfin, ce jugement peut être utilisé dans les procédures de décision, de coopération et de confiance (Mao, Gratch, 2013).

Enfin, nous remarquons que certains travaux cherchent à proposer un processus entièrement rationnel du jugement tandis que d'autres introduisent un système simulant des émotions ou faisant intervenir des motivations issues de réactions émotionnelles. Nous relevons que l'emploi de modèles émotionnels permet de favoriser une certaine empathie dans les relations de l'utilisateur humain avec l'agent (voir sections I.2.3 et II.5.4) ou permet de reproduire des réactions humaines utiles dans le cadre de problématiques de modélisation de comportement humain (par exemple dans le cadre de simulations sociales). Toutefois, lorsque les agents se voient confier des responsabilités, les humains semblent attendre d'eux un jugement rationnel (voir la présentation de (Malle *et al.*, 2015) en introduction de ce chapitre). Nous exclurons donc la représentation des émotions au sein du raisonnement de l'agent pour fonder son jugement sur un modèle entièrement rationnel.

Aucun modèle présenté dans ce chapitre ne propose de modèle opérationnel permettant de concevoir des agents autonomes répondant à l'ensemble de ces besoins. Bien que certaines approches fournissent des pistes de réflexion pour répondre aux problématiques de l'éthique à l'échelle collective, il reste nécessaire de concevoir un modèle d'agent capable de les mettre en œuvre et d'éprouver ensuite ce modèle dans une démarche expérimentale.

Le chapitre III propose en ce sens un modèle de jugement des actions permettant d'introduire les concepts de morale et d'éthique définis précédemment dans le cadre d'une architecture BDI. Le chapitre IV présente ensuite l'emploi de ce modèle de jugement dans un cadre de coopération entre agents fondé sur l'éthique.

# CHAPITRE III

# Modèle de jugement des actions

**III.1 Préambule** . . . . . **52**
III.1.1 Fondements . . . . . 52
III.1.2 Scénario illustratif . . . . . 55
**III.2 Présentation globale du modèle de jugement** . . . . . **57**
**III.3 Reconnaissance de situation** . . . . . **59**
**III.4 Évaluation de la possibilité et de la désirabilité des actions** . **60**
III.4.1 Évaluation de la possibilité. . . . . 60
III.4.2 Évaluation de la désirabilité . . . . . 62
**III.5 Évaluation de la moralité des actions** . . . . . **63**
III.5.1 Système de valeurs morales . . . . . 64
III.5.2 Supports de valeurs . . . . . 65
III.5.3 Évaluation des supports de valeurs. . . . . 66
III.5.4 Règles morales . . . . . 67
III.5.5 Évaluation de la moralité . . . . . 69
**III.6 Évaluation de l'éthique des actions** . . . . . **70**
III.6.1 Principes éthiques . . . . . 71
III.6.2 Fonction d'évaluation de l'éthique . . . . . 72
III.6.3 Préférences éthiques . . . . . 73
III.6.4 Fonction de jugement . . . . . 73
**III.7 Exemple récapitulatif** . . . . . **74**
**III.8 Synthèse** . . . . . **76**

Après avoir défini les notions de morale et d'éthique au chapitre I puis montré comment des travaux d'Intelligence Artificielle proposent de modéliser ces notions au chapitre II, ce chapitre introduit un modèle de jugement d'un agent autonome. Ce modèle est une proposition répondant aux principes énoncés et motivés au chapitre précédent, c'est-à-dire un modèle utilisant une représentation explicite des connaissances de l'agent sur le

domaine applicatif, la morale et l'éthique, excluant l'intervention d'émotions simulées et s'attachant à définir l'ensemble des éléments nécessaires à sa mise en œuvre.

Les définitions formelles des éléments du modèle sont illustrées par un exemple que nous présentons en section III.1. La section III.2 présente le modèle de jugement dans son ensemble afin d'en décrire le fonctionnement à haut niveau. Les sections III.3, III.4, III.5 et III.6 décrivent les éléments du modèle de jugement des actions que sont respectivement les modèles de reconnaissance de situation et d'évaluation de la désirabilité d'une part et d'évaluation de la moralité et d'évaluation de l'éthique d'autre part. Enfin la section III.7 expose à l'aide d'un exemple récapitulatif l'ensemble du fonctionnement de ce modèle.

# III.1 Préambule

Avant de définir comment un agent autonome peut juger du caractère éthique d'une action, nous précisons ici le cadre conceptuel dans lequel s'inscrit ce modèle de jugement. Le scénario utilisé pour illustrer l'ensemble des notions introduites dans ce chapitre est ensuite présenté.

## III.1.1 Fondements

Comme mentionné dans le chapitre précédent, notre approche de l'éthique des agents autonomes vise à proposer un modèle pour que les agents raisonnent sur une représentation symbolique de l'éthique et de la morale. La représentation de nombreuses notions employées dans la littérature philosophique (désirs, actions, intentions, etc.) a déjà fait l'objet de travaux et sont des éléments classiques des cadres conceptuels de raisonnement des agents autonomes. L'un des plus connus d'entre eux est le cadre « Belief-Desire-Intention » (ou BDI, voir section II.5.2)(Bratman, 1987 ; A. Rao, Georgeff, 1995) dans lequel les problèmes de décision d'un agent sont ramenés au choix d'une intention par la sélection d'une action au regard de croyances (*Beliefs*) afin de satisfaire un ou plusieurs désirs (*Desires*).

L'approche présentée dans ce chapitre est pensée comme une extension du cadre BDI, avec la prise en compte de concepts liés à la morale et à l'éthique en plus des croyances et désirs. L'objectif est de proposer un modèle du jugement permettant de qualifier les actions de juste ou non dans un contexte et au regard de connaissances.

Le modèle est décrit à l'aide de la logique du premier ordre afin de définir formellement et sans ambiguïté les éléments employés. Enfin ce modèle est présenté avec l'hypothèse d'un monde ouvert afin de pouvoir représenter ce que l'agent ignore et prendre en compte dans le raisonnement l'absence de connaissance. Nous considérons un langage $\mathcal{L}$ permettant de décrire la totalité des mondes possibles, ainsi que des expressions portant sur les raisonnements des agents, où la négation forte sera notée $\neg$ et la négation faible *not*. Le *monde* désigne ici l'environnement de l'agent, c'est-à-dire tout ce qui n'est pas l'agent lui-même dans le système. Ainsi, les informations communiquées par les autres agents font partie des croyances de l'agent.

**Définition 10** (Croyance)**.** *Les* croyances *de l'agent sont des informations portant sur*

*l'état du monde ou sur ses propres raisonnements. Ces croyances sont des expressions exprimées dans $\mathcal{L}$, et la totalité des expressions possibles dans ce langage est notée $B$, que nous appelons l'*ensemble des croyances possibles*. L'ensemble courant des croyances de l'agent est noté $\mathcal{B}$.*

**Exemple 7.** *Un agent peut, par exemple disposer d'informations sur l'état du monde telles que le fait qu'une porte $p$ soit ouverte, représenté par le prédicat ouverte($p$).*

Comme mentionné en section 1, la manière dont est jugée une action peut fortement dépendre du contexte, c'est-à-dire de l'état du monde décrit par $\mathcal{B}$ et interprété par l'agent.

**Définition 11** (Désir)**.** *Les* désirs *d'un agent représentent un ensemble d'états à atteindre. Dans le modèle présenté ici, les désirs, exprimés dans $\mathcal{L}$, décrivent un ensemble d'états que l'agent souhaite atteindre. L'ensemble des désirs d'un agent est noté $\mathcal{D}$. La totalité des désirs possibles est notée $D$.*

Dans un modèle BDI classique, l'agent cherche en permanence une solution pour satisfaire ses désirs en sélectionnant des actions supposées le mener dans un état désiré. Dans le cadre de ce modèle, $\psi \in \mathcal{D}$ est représenté sous la forme d'un prédicat $desire(\psi)$ où $\psi$ est une expression (une conjonction ou disjonction d'expressions appartenant à $D$). L'état décrit par les croyances et désirs courants est appelé *situation courante* :

**Définition 12** (Situation courante)**.** *Une* situation courante $s = \mathcal{B} \cup \mathcal{D}$ *désigne l'ensemble des croyances et désirs de l'agent au moment du raisonnement.*

Nous emploierons également le terme de *situation hypothétique* tel que :

**Définition 13** (Situation hypothétique)**.** *Une* situation hypothétique *notée $w \subseteq B \cup D$ désigne un ensemble d'états représentés par une conjonction de croyances et désirs.*

L'agent est capable de raisonner sur des *situations hypothétiques*, différentes de la situation courante, afin d'envisager par exemple des états dans lesquels peuvent le mener des actions.

**Définition 14** (Action)**.** *Les* actions *à disposition de l'agent sont modélisées par un nom permettant d'identifier l'action, des conditions, c'est-à-dire des croyances nécessaires pour qu'il soit possible d'effectuer l'action, des conséquences, c'est-à-dire des croyances que l'agent s'attend à voir devenir vraies suite à l'action, et des arguments exprimés dans $\mathcal{L}$ désignant des éléments du monde pouvant être employés dans la description des conditions et conséquences. L'ensemble des actions connues de l'agent est noté $A$.*

La représentation employée ici, par certains aspects inspirée de PDDL (Ghallab *et al.*, 1998) (Planning Domain Definition Language), définit les actions de l'agent comme des opérateurs disponibles sur les états du monde dans lesquels les préconditions de l'action

sont vraies et menant à des états du monde dans lesquels les conséquences de l'action sont vraies. Dans le cadre de ce modèle, toute action a au moins pour conséquence la production de la croyance $done(a)$ signifiant que l'action a été effectuée.

Dans le cadre de ce modèle, lorsque l'ensemble des conditions d'une action $a$ est satisfait dans une situation courante $s$, constituée de la conjonction d'expressions de $\mathcal{B}$ et $\mathcal{D}$, le prédicat $condition(a, s)$ est vrai.

De manière analogue, le prédicat indiquant que $\psi$, décrivant un ensemble d'états atteignables grâce à l'action $a$ à partir de la situation courante $s$, est une conséquence de $a$ notée $postcond(a, s, \psi)$.

**Exemple 8.** *Considérons une action $ouvre(p)$ consistant à ouvrir une porte $p$ dans une situation $s$ et décrite de la manière suivante :*

$$\begin{aligned} condition(ouvre(p), s) \equiv & s \rightarrow \neg\ ouverte(p) \\ postcond(ouvre(p), s, \psi) \equiv & \psi \rightarrow ouverte(p) \wedge done(ouvre(p)) \end{aligned}$$

*Afin de permettre à l'agent de raisonner sur la faisabilité de cette action, nous décrivons les conditions de sa réalisation : ici, la porte doit être fermée (c'est-à-dire que la croyance $\neg\ ouverte(p)$, représentant l'information dont dispose l'agent sur l'état de la porte, est vraie dans la situation courante $s$).*

*De même, les postconditions de l'action $ouvre(p)$ dans la situation $s$ montrent que, dans l'ensemble d'états $\psi$ atteint par la réalisation de l'action, $ouverte(p)$ est vrai.*

Comme dans PDDL, des préconditions secondaires sont employées pour décrire des conséquences dépendantes de la situation dans laquelle est réalisée l'action. Nous les représentons sous la forme d'implications dans les postconditions.

**Exemple 9.** *Pour reprendre l'exemple ci-dessus, nous ajoutons une postcondition secondaire indiquant que si un système d'alarme est installé sur la porte (connaissance représentée par la croyance $alarme_sur(p)$), celui-ci se mettra à sonner lorsque l'action sera effectuée (la croyance $sonne_alarme(p)$ sera présente dans $\psi$) :*

$$\begin{aligned} condition(ouvre(p), s) \equiv & s \rightarrow \neg\ ouverte(p) \\ postcond(ouvre(p), s, \psi) \equiv & \psi \rightarrow ouverte(p) \wedge done(ouvre(p)) \\ & \wedge (s \rightarrow alarme_sur(p)) \rightarrow (\psi \rightarrow sonne_alarme(p)) \end{aligned}$$

Il est courant de regrouper des séquences d'actions permettant d'atteindre un état désiré au sein de plans et d'évaluer ces plans. Cependant, ce chapitre ne traite ici que de l'évaluation d'actions seules. L'évaluation de séquences d'actions est abordée au chapitre IV.

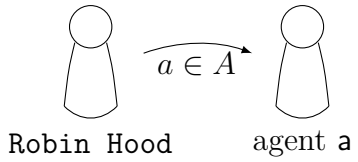


FIGURE III.1 – Illustration du scénario pris en exemple : Robin Hood

### III.1.2 Scénario illustratif

Dans le but de faciliter la compréhension du modèle de jugement, ce chapitre est illustré par un exemple basé sur un scénario fictif simple qu'est l'histoire de « Robin des bois », démontrant l'expressivité du modèle sans les restrictions qu'imposerait un scénario issu d'un cas d'usage du monde réel.

Dans ce scénario (voir figure III.1), un agent nommé Robin Hood est situé dans un environnement où tout agent a est doté de propriétés observables permettant de l'identifier comme pauvre ($poor(\mathsf{a})$), riche ($rich(\mathsf{a})$) ou ni pauvre ni riche ($\neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})$), et éventuellement noble ($noble(\mathsf{a})$).

L'ensemble $A$ des actions connues de Robin Hood est $A = \{wait, steal, give, tax, court\}$. Ces actions sont décrites de la manière suivante :

$a_0$ : $wait()$ pour rester dans l'état courant, avec

$$\begin{aligned} condition(wait(), s) \equiv & \top \\ postcond(wait(), s, \psi) \equiv & \psi \rightarrow done(wait()) \end{aligned}$$

signifiant que cette action est possible dans tout état du monde et qu'aucun changement d'état n'est causé par cette action ;

$a_1$ : $steal(\mathsf{a})$ pour voler à un agent a non pauvre une part de ses richesses afin de s'enrichir, avec :

$$\begin{aligned} condition(steal(\mathsf{a}), s) \equiv & s \rightarrow not\ poor(\mathsf{a}) \wedge not\ rich(\text{Robin Hood}) \\ postcond(steal(\mathsf{a}), s, \psi) \equiv & (s \rightarrow poor(\text{Robin Hood})) \rightarrow (\psi \rightarrow \neg\ poor(\text{Robin Hood})) \\ & \wedge (s \rightarrow \neg poor(\text{Robin Hood}) \wedge \neg rich(\text{Robin Hood})) \\ & \quad \rightarrow (\psi \rightarrow rich(\text{Robin Hood})) \\ & \wedge (s \rightarrow rich(\mathsf{a})) \rightarrow (\psi \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \\ & \wedge (s \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \rightarrow (\psi \rightarrow poor(\mathsf{a})) \\ & \wedge \psi \rightarrow done(steal(\mathsf{a})) \end{aligned}$$

$a_2$ : $give(\mathsf{a})$ pour donner de l'argent à l'agent $\mathsf{a}$ non riche en s'appauvrissant, avec :

$$\begin{aligned}
condition(give(\mathsf{a}), s) \equiv & s \rightarrow not\ rich(\mathsf{a}) \wedge not\ poor(\text{Robin Hood}) \\
postcond(give(\mathsf{a}), s, \psi) \equiv & (s \rightarrow rich(\text{Robin Hood})) \rightarrow (\psi \rightarrow \neg rich(\text{Robin Hood})) \\
& \wedge (s \rightarrow \neg poor(\text{Robin Hood}) \wedge \neg rich(\text{Robin Hood})) \\
& \quad \rightarrow (\psi \rightarrow poor(\text{Robin Hood})) \\
& \wedge (s \rightarrow poor(\mathsf{a})) \rightarrow (\psi \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \\
& \wedge (s \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \\
& \quad \rightarrow (\psi \rightarrow rich(\mathsf{a})) \\
& \wedge \psi \rightarrow done(give(\mathsf{a}))
\end{aligned}$$

$a_3$ : $tax(\mathsf{a})$ pour réclamer à l'agent $\mathsf{a}$ une part de ses richesses avec :

$$\begin{aligned}
condition(tax(\mathsf{a}), s) \equiv & s \rightarrow not\ poor(\mathsf{a}) \wedge not\ rich(\text{Robin Hood}) \wedge noble(\text{Robin Hood}) \\
postcond(tax(\mathsf{a}), s, \psi) \equiv & (s \rightarrow poor(\text{Robin Hood})) \rightarrow (\psi \rightarrow \neg\ poor(\text{Robin Hood})) \\
& \wedge (s \rightarrow \neg poor(\text{Robin Hood}) \wedge \neg rich(\text{Robin Hood})) \\
& \quad \rightarrow (\psi \rightarrow rich(\text{Robin Hood})) \\
& \wedge (s \rightarrow rich(\mathsf{a})) \rightarrow (\psi \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \\
& \wedge (s \rightarrow \neg poor(\mathsf{a}) \wedge \neg rich(\mathsf{a})) \\
& \quad \rightarrow (\psi \rightarrow poor(\mathsf{a})) \\
& \wedge \psi \rightarrow done(tax(\mathsf{a}))
\end{aligned}$$

$a_4$ : $court(\mathsf{a})$ pour tenter de s'attirer les faveurs de l'agent $\mathsf{a}$ avec

$$\begin{aligned}
condition(court(\mathsf{a}), s) \equiv & \top \\
postcond(court(\mathsf{a}), s, \psi) \equiv & \psi \rightarrow done(court(\mathsf{a}))
\end{aligned}$$

signifiant comme pour $a_0$ que cette action est possible dans tout état du monde ;

Nous considérons que les actions $a_1$ et $a_3$ ont toutes deux des conséquences identiques : pour l'agent effectuant l'action, il devient riche s'il n'était ni pauvre ni riche, et ni pauvre ni riche s'il était pauvre. Pour l'agent $\mathsf{a}$ ciblé par cette action, l'inverse se produit : il devient pauvre s'il n'était ni riche ni pauvre et ni riche ni pauvre s'il était riche. Pour que ces actions soient possibles, l'agent $\mathsf{a}$ ne doit pas être pauvre. La seule condition supplémentaire pour effectuer l'action $a_3$ est d'être noble. L'action $a_2$, à l'inverse des deux précédentes, fait passer l'agent effectuant l'action de l'état riche à l'état ni riche ni pauvre ou de ni riche ni pauvre à pauvre. Il est nécessaire que l'agent effectuant l'action ne soit pas pauvre. L'agent $\mathsf{a}$ ciblé par l'action passe de pauvre à ni pauvre ni riche ou de ni pauvre ni riche à riche.

Les actions $a_0$ et $a_4$ sont exécutables dans toute situation et n'ont aucune conséquence connue sur les prédicats *poor*, *rich* et *noble*.

Au cours des prochaines sections, nous allons voir comment définir la morale et l'éthique de Robin Hood et comment le jugement montre que son comportement est conforme à son éthique de « voler aux riches pour donner aux pauvres ».

## III.2 Présentation globale du modèle de jugement

Le modèle de jugement prend en entrée l'état du monde et produit en sortie un ensemble $\mathcal{A}_r$ d'actions justes. La figure III.2 illustre ce modèle en montrant comment les différents éléments qui le composent viennent étendre le modèle BDI classique afin de fournir à la fonction de sélection des intentions une information sur la justesse des actions.

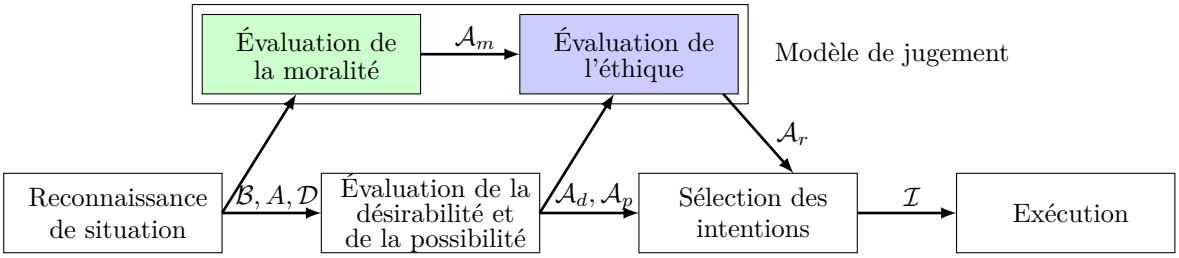


FIGURE III.2 – Intégration du modèle de jugement dans l'architecture BDI

Le but du modèle de jugement présenté dans ce chapitre est de produire l'ensemble $\mathcal{A}_r$ des actions justes en ajoutant à l'architecture classique des agents BDI un ensemble de connaissances et de fonctions. Le chapitre suivant montrera divers usages de ce modèle de jugement. Comme évoqué précédemment, ce modèle a pour objectif de qualifier des actions. Par souci de concision nous nommons ces modèles « modèle d'évaluation de la désirabilité » (respectivement « de la possibilité », « de la moralité » ou « de l'éthique ») en lieu et place de « modèle d'évaluation de la désirabilité des actions » (de même pour les autres).

Notons que, dans le cadre de ce chapitre, nous considérons les autres agents comme faisant partie de l'environnement. L'agent peut ici raisonner sur des connaissances décrivant la situation des autres (autrement dit, cela fait entièrement partie de la situation courante, c'est-à-dire l'état de l'environnement tel que perçu par l'agent) mais ne peut, à ce stade, employer les connaissances des autres dans son raisonnement.

Le modèle de jugement présenté ici suppose l'existence de deux modèles dits « classiques » au sens de leur présence dans la grande majorité des travaux sur l'architecture BDI. Nous redéfinissons ici les caractéristiques de ces éléments :

- le modèle de reconnaissance de situation (ou $SAM$ pour *Situation Assesment Model*), récupérant des informations sur l'état du monde pour créer la *situation courante*, c'est-à-dire l'ensemble des croyances et désirs de l'agent, sur laquelle s'appuient les modèles qui suivent ;
- le modèle d'évaluation de la désirabilité et de la possibilité (ou $DPEM$ pour *Desirability and Possibility Evaluation Model*), déduisant la faisabilité et la désirabilité

des actions de l'agent en fonction de la situation obtenue du modèle précédent.

À ces modèles classiques, nous ajoutons deux modèles :

- le modèle d'évaluation de la moralité (ou $MEM$ pour *Morality Evaluation Model*), évaluant la moralité des actions dans la situation courante au regard de la théorie du bien (voir section I.1.1) ;
- Le modèle d'évaluation de l'éthique (ou $EEM$ pour *Ethics Evaluation Model*), évaluant le caractère éthique des actions au regard des évaluations fournies par les modèles précédents et la théorie du juste (voir section I.1.2).

Plus formellement, nous pouvons donner la définition suivante :

**Définition 15** (Modèle de jugement)**.** *Un* modèle de jugement *(ou $JM$ pour Judgment Model) est défini comme une composition d'un modèle d'évaluation de la moralité ($MEM$) et d'un modèle d'évaluation de l'éthique ($EEM$). Il s'appuie sur une ontologie du jugement $\mathcal{O}$. Ce modèle de jugement produit une évaluation des actions pour l'état courant du monde $W$ en tenant compte de considérations morales et éthiques.*

$$JM = \langle MEM, EEM, \mathcal{O} \rangle$$

**Définition 16** (Ontologie du jugement)**.** *L'ontologie du jugement $\mathcal{O}$ permet de définir les éléments de langage employés pour évaluer les actions dans le modèle de jugement.*

$$\mathcal{O} = \mathcal{O}_v \cup \mathcal{O}_{sv} \cup \mathcal{O}_{hv} \cup \mathcal{O}_m \cup \mathcal{O}_d \cup \mathcal{O}_p$$

L'ontologie $\mathcal{O}$ contient :

$\mathcal{O}_v$ la description d'un ensemble $V$ de valeurs morales dont l'existence et le nom sont connus de tous les agents. Bien que chaque agent puisse disposer de son propre ensemble de valeurs morales issues de $\mathcal{O}_v$, cette partie de l'ontologie permet de modéliser le caractère universel des valeurs reconnues (voir section I.3.2).

$\mathcal{O}_{sv}$ l'ensemble $SV$ des *valuations de support* et la relation d'ordre total sur celui-ci, représentant une échelle de trahison ou promotion d'une valeur par une action dans une situation.

$\mathcal{O}_{hv}$ un ensemble de relations hiérarchiques entre valeurs permettant de définir un *système de valeur*.

$\mathcal{O}_m$ l'ensemble $MV$ de *valuations morales* et la relation d'ordre total sur celui-ci, représentant par un nombre fini et ordonné de valeurs discrètes, une échelle permettant d'exprimer un niveau de moralité.

$\mathcal{O}_d$ un ensemble $DV$ ordonné de *valuations de désirabilité* et la relation d'ordre total sur celui-ci, commun aux agents et leur permettant de représenter et exprimer le caractère désirable ou indésirable d'une action.

$\mathcal{O}_p$ l'ensemble $PV$ ordonné des *valuations de possibilité* et la relation d'ordre total sur celui-ci, permettant d'exprimer la faisabilité d'une action.

**Exemple 10.** *Dans l'exemple du scénario illustratif, nous prendrons l'ensemble de valuations de support* $SV = \{defeat, promote\}$*, ordonné selon cet ordre croissant, l'ensemble de valuations morales* $MV = \{very\ immoral, immoral, amoral, moral, very\ moral\}$*, ordonné selon cet ordre croissant de moralité, l'ensemble de valuations de désirabilité* $DV = \{undersirable, neutral, desirable\}$*, ordonné selon cet ordre croissant de désirabilité et l'ensemble de valuations de possibilité* $PV = \{impossible, possible\}$*, ordonné selon cet ordre croissant de possibilité. Les valuations* $amoral$ *et* $neutral$ *sont ici des* valeurs de référence *servant de seuil entre les valuations dites « positives » et « négatives » de leurs ensembles respectifs.*

Les quatre sections suivantes ont pour but de définir les modèles de reconnaissance de situation et d'évaluation sur lesquels s'appuie le modèle de jugement, puis les modèles d'évaluation de la moralité et d'évaluation de l'éthique qui le constituent. Ces modèles s'appuient tous sur le système de valuation défini par l'ontologie $\mathcal{O}$ pour caractériser les actions.

## III.3 Reconnaissance de situation

Le modèle de reconnaissance de situation a pour but de permettre à l'agent d'évaluer l'état du monde au travers de ses perceptions, ses fonctions de communication et ses processus internes ayant pour effet la mise à jour de l'ensemble de ses croyances et désirs (voir figure III.3). Ce modèle est nécessaire pour acquérir la situation dans laquelle les actions vont être évaluées. Nous n'aborderons pas les questions liées à la fiabilité des informations acquises qui, bien que pouvant soulever des problématiques éthiques, sortent du cadre présenté ici visant à proposer un modèle permettant de raisonner sur les informations acquises.

**Définition 17** (Modèle de reconnaissance de situation)**.** *Le* modèle de reconnaissance de situation *(ou Assessment Model)* $SAM$ *génère l'ensemble des croyances qui décrivent l'état courant du monde* $W$ *et l'ensemble des désirs qui décrivent les buts de l'agent. Il est défini comme :*

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

*où* $\mathcal{B}$ *est l'ensemble des croyances que l'agent a sur* $W$ *parmi* $B$ *l'ensemble des croyances possibles, et* $\mathcal{D}$ *ses désirs parmi* $D$ *l'ensemble des désirs possibles. Ces deux ensembles sont produits par la fonction de reconnaissance de situation* $SA$ *à partir du monde* $W$ *:*

$$SA : W \to 2^B \cup 2^D$$

**Exemple 11.** *Dans notre exemple, le modèle de reconnaissance de situation est employé par* Robin Hood *pour construire sa situation courante. Ainsi,* Robin Hood *peut obtenir des croyances sur sa propre situation et celle des autres (tel que* $poor$(Robin Hood)*,* $noble$(Prince John)*,* $poor$(Friar Tuck) *et not* $poor$(Prince John)*), et ses propres désirs tel que* $desire(done(court(\text{Marian})))$*.*

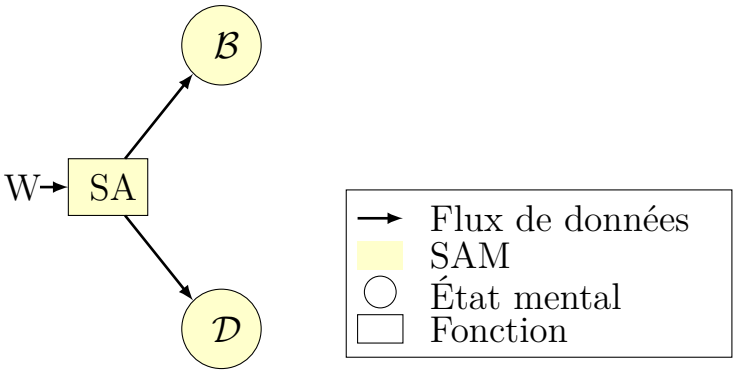


FIGURE III.3 – Modèle de reconnaissance de situation

## III.4 Évaluation de la possibilité et de la désirabilité des actions

À partir de l'ensemble des croyances et désirs courants, un agent utilise le modèle d'évaluation (voir figure III.4) pour calculer l'ensemble des actions évaluées en fonction de leur caractère désirable $\mathcal{A}_d$ d'une part (c'est-à-dire des associations d'actions et de valuations de désirabilité de $\mathcal{O}_d$ en fonction des désirs de l'agent et des postconditions des actions décrites dans $A$) et les actions possibles $\mathcal{A}_p$ d'autre part (c'est-à-dire les actions pouvant être effectuées dans la situation courante $s$ au regard de leurs préconditions décrites dans $A$).

Par la suite, nous appelons connaissances contextuelles notées $CK$ (pour *Contextual Knowledge*) avec $CK = A \cup \mathcal{O}_d \cup \mathcal{O}_p \cup s$ l'ensemble des connaissances sur les actions et ontologies employées dans ce modèle, ainsi que la situation courante perçue par l'agent.

**Définition 18** (Modèle d'évaluation de la désirabilité et de la possibilité)**.** *Le* modèle d'évaluation de la désirabilité et de la possibilité *(ou Desirability and Possibility Evaluation Model) $DPEM$ produit les ensembles des actions évaluées par rapport à leur désirabilité $\mathcal{A}_d$ et leur possibilité $\mathcal{A}_p$ à partir des ensembles de désirs, croyances et connaissances sur les actions :*

$$DPEM = \langle A, \mathcal{A}_d, \mathcal{A}_p, DE, PE, \mathcal{O} \rangle$$

*où $\mathcal{A}_d \subseteq A \times DV$ et $\mathcal{A}_p \subseteq A \times PV$ sont respectivement l'ensemble des couples d'actions et de valuation de désirabilité $\mathcal{A}_d$ produit par la fonction d'évaluation de la désirabilité $DE$ (pour Desirability Evaluation), et l'ensemble des couples d'actions et de valuations de possibilité $\mathcal{A}_p$, produit par la fonction d'évaluation de la possibilité $PE$ (pour Possibility Evaluation). $\mathcal{O}$ est l'ontologie employée.*

### III.4.1 Évaluation de la possibilité

L'évaluation des conditions d'une action $a$ permettent d'évaluer si sa réalisation est possible dans l'état courant de $\mathcal{B}$. Cette information est nécessaire pour comparer les actions connues de l'agent.

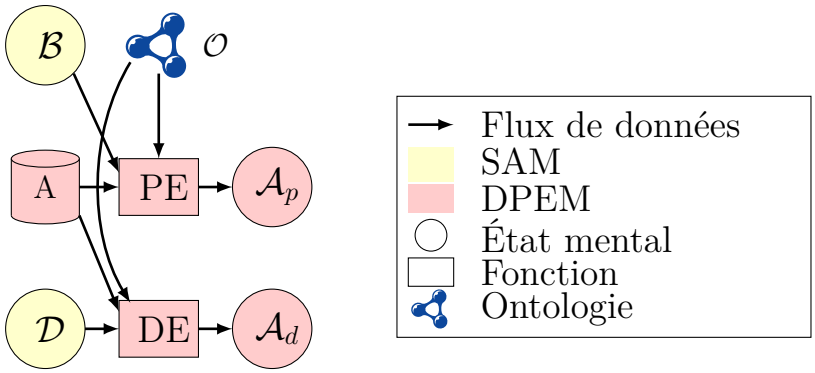


FIGURE III.4 – Modèle d'évaluation de la désirabilité et de la possibilité

**Définition 19** (Évaluation de la possibilité)**.** *La fonction d'*évaluation de la possibilité *PE (pour Possibility Evaluation) produit l'ensemble $\mathcal{A}_p$ des actions possibles pour lesquelles l'agent a affecté une valuation de possibilité en fonction de l'état courant décrit par $\mathcal{B}$ et des conditions de l'action décrites dans $A$.*

$$PE : 2^{\mathcal{B}} \times A \times \mathcal{O} \rightarrow \mathcal{A}_p$$

$$\mathcal{A}_p \subseteq A \times PV$$

$$(a, pv) \in \mathcal{A}_p : a \in A, pv \in PV$$

Ainsi, l'ensemble $\mathcal{A}_p$ des actions possibles est obtenu par déduction à partir des conditions des actions en affectant une valuation de possibilité $pv \in PV$ à chaque action $a \in A$ grâce au calcul du prédicat $possible(a, s, pv)$ :

$$\mathcal{A}_p = \{(a, pv) \in A \times PV \ s.t. \ possible(a, s, pv) \in \mathcal{B}\}$$

**Exemple 12.** *Dans le cadre de notre exemple illustratif, avec $PV = \{impossible, possible\}$, l'évaluation se fait de la manière suivante :*

$$possible(a, s, possible) \equiv condition(a, s)$$
$$possible(a, s, impossible) \equiv \neg condition(a, s)$$

*À partir des connaissances sur les actions décrites en section III.1.2 et de la situation courante décrite en section III.3,* Robin Hood *peut, par exemple, évaluer que :*

— $(wait(), possible) \in \mathcal{A}_p$ *puisque la condition de cette action est vérifiée dans toutes les situations possibles et donc particulièrement dans la situation courante ;*
— $(steal($Prince John$), possible) \in \mathcal{A}_p$ *puisque, dans la situation courante $s$,* Robin Hood *est pauvre et* Prince John *n'est pas pauvre ;*
— $(give($Marian$), impossible) \in \mathcal{A}_p$ *puisque, dans la situation courante $s$,* Robin Hood *est pauvre.*

### III.4.2 Évaluation de la désirabilité

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

Ainsi, une action $a$ est évaluée comme étant désirable si l'agent désire la réalisation de $a$ ou la réalisation de ses conséquences (c'est-à-dire désirer $\lambda$ dans une situation courante $s$ tel que $\exists\psi : postcond(a, s, \psi)$ avec $\psi \rightarrow \lambda$). À l'inverse, l'action peut être indésirable s'il désire que ses conséquences ne se réalisent pas (c'est-à-dire désirer $\neg\lambda$ tel que $\exists\psi : postcond(a, s, \psi)$ avec $\psi \rightarrow \neg\lambda$).

**Exemple 13.** *Nous instancions la fonction d'évaluation de la désirabilité telle qu'une action puisse être désirable et indésirable simultanément, par exemple si l'agent désire $\lambda_1$ et $\lambda_2$ dans la situation $s$ et que $\exists\psi : postcond(a, s, \psi)$ avec $\psi \rightarrow \lambda_1, \psi \rightarrow \neg\lambda_2$.*

*Notons également que les conséquences d'une action $a_1$ peuvent être désirables dans une situation courante $s$ en raison de conséquences $\lambda_2$ d'une autre action $a_2$, désirable, et dont les conditions sont des conséquences de $a_1$, c'est-à-dire que l'agent désire $\lambda_2$, et puisque $\exists\psi_2 : postcond(a_2, s_2, \psi_2)$ avec $\psi_2 \rightarrow \lambda_2$, $\exists s_2 : precond(a_2, s_2)$ avec $s_2 \rightarrow \lambda_1$ et $\exists\psi_1 : postcond(a_1, s, \psi_1)$ avec $\psi_1 \rightarrow \lambda_1$, alors $a_1$ serait désirable.*

**Définition 20** (Évaluation de la désirabilité)**.** *La fonction d'évaluation de la désirabilité $DE$ (pour Desirability Evaluation) produit l'ensemble $\mathcal{A}_d$ des actions évaluées en fonction de leur désirabilité dans la situation courante à partir de $\mathcal{D}$ l'ensemble des désirs et $A$ l'ensemble des actions.*

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

$$\mathcal{A}_d \subseteq A \times DV$$

$$(a, dv) \in \mathcal{A}_d : a \in A, vd \in DV$$

*Ainsi l'ensemble des actions évaluées en fonction de la désirabilité associe aux actions des valuations de désirabilité.*

Le calcul de l'ensemble $\mathcal{A}_d$ est effectué par le prédicat $desired(a, s, d, dv)$ affectant à une action $a$ une valuation de désirabilité $dv$ décrite dans $\mathcal{O}_d$ dans la situation $s$ en raison du désir $d \in D$ :

$$\mathcal{A}_d = \{(a, dv) : A \times DV \ s.t. \ \exists d \in D \ desired(a, s, d, dv)\}$$

**Exemple 14.** *Dans le cadre de notre exemple, avec $DV = \{undesirable, neutral, desirable\}$,*

*l'évaluation se fait de la manière suivante :*

$$
\begin{aligned}
desired(a, s, desire(\psi), desirable) \equiv & postcond(a, s, w) \wedge w \rightarrow \psi \wedge desire(\psi) \\
desired(a, s, \emptyset, neutral) \equiv & postcond(a, s, w) \wedge w \rightarrow \psi \\
& \wedge not\ desire(\psi) \wedge not\ desire(\neg\psi) \\
desired(a, s, desire(\neg\psi), undesirable) \equiv & postcond(a, s, w) \wedge w \rightarrow \psi \wedge desire(\neg\psi) \\
desired(a, s, desire(\psi), desirable) \equiv & \exists\psi', \exists w, \exists a' : desire(\psi) \\
& \wedge postcond(a, s, w) \\
& \wedge \neg condition(a', s) \\
& \wedge condition(a', w') \\
& \wedge postcond(a', w', w'') \\
& \wedge w \subseteq w' \\
& \wedge w'' \rightarrow \psi
\end{aligned}
$$

*Par exemple, supposons un agent* Prince John *ayant dans l'état courant $s$ le désir $d_1 = desire(rich(\text{Prince John}))$ d'être riche dans la situation courante $s$. Sachant que $\exists\psi : postcond(tax(\text{Robin Hood}), s, \psi)$ et $\psi \rightarrow rich(\text{Prince John})$, cet agent va produire $desired(tax(\text{Robin Hood}), s, desire(rich(\text{Prince John})), desirable)$ signifiant qu'il affecte la valuation de désirabilité desirable à l'action $tax(\text{Robin Hood})$ dans la situation courante $s$ en raison de $d_1$.*

*Pour l'agent* Robin Hood*, seul le désir $d_2 = desire(done(court(\text{Marian})))$ signifiant que l'agent souhaite se trouver dans un état où il a courtisé l'agent* Marian*, est présent dans l'ensemble $\mathcal{D}$ des désirs courants de l'agent. Ainsi* Robin Hood *va produire $desired(court(\text{Marian}), s, desire(done(court(\text{Marian}))), desirable)$ signifiant qu'il affecte la valuation de désirabilité desirable à l'action $court(\text{Marian})$ dans la situation courante $s$ en raison de $d_2$.*

*Cette manière d'évaluer la désirabilité est évidement d'une très grande simplicité en comparaison de ce qui est utilisé de façon courante dans les travaux traitant des problématiques de prise de décision. Toute fonction d'évaluation plus complexe peut être employée si elle satisfait la définition 20.*

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

## III.5 Évaluation de la moralité des actions

Comme cela a été montré en section 1, avant d'évaluer le caractère éthique d'une action dans une situation donnée, il est nécessaire d'en évaluer la dimension morale, c'est-à-dire associer à cette action une ou plusieurs valuations au regard d'une théorie du bien. Cette

section présente dans cette optique la partie du modèle (voir figure III.5) permettant de représenter et employer dans le raisonnement les différents constituants de la théorie du bien, ou connaissance du bien (voir section 1) notée $GK$ (pour *Goodness Knowledge*), constituée des supports de valeurs morales et des règles morales, soit $GK = VS \cup MR$.

**Définition 21** (Modèle d'évaluation morale)**.** *Un* modèle d'évaluation morale $MEM$ *identifie les actions morales étant donné les informations dont l'agent dispose sur le contexte $s$, et sa théorie du bien $GK$.*

$$MEM = \langle VS, MR, \mathcal{A}_v, \mathcal{A}_m, ME, VE, \mathcal{O} \rangle$$

*où $VS$ est l'ensemble des supports de valeurs connus, $MR$ est l'ensemble des règles morales connues, $\mathcal{A}_v$ est l'ensemble des actions évaluées par les valeurs produit par une fonction $VE$ d'évaluation des supports de valeur et $\mathcal{A}_m$ est l'ensemble des actions moralement évaluées par la fonction d'évaluation morale $ME$, c'est-à-dire associées à une ou plusieurs valuations morales au regard d'un ensemble de règles morales. $\mathcal{O}$ est l'ontologie employée.*

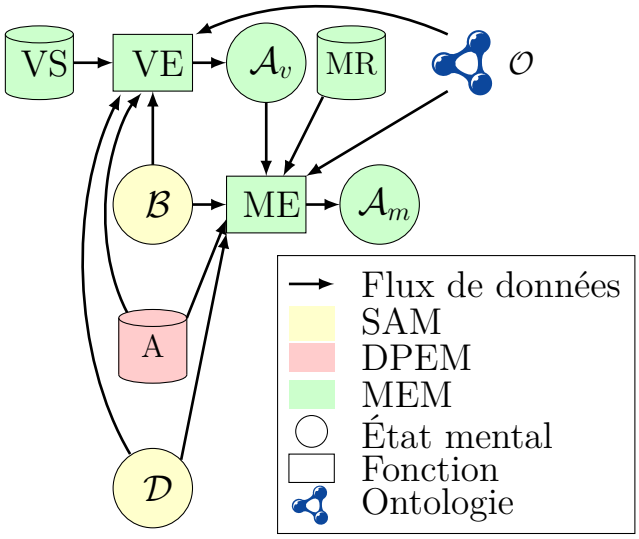


FIGURE III.5 – Modèle d'évaluation de la moralité

## III.5.1 Système de valeurs morales

La représentation et le raisonnement sur les valeurs morales nécessite la définition de plusieurs éléments. Premièrement, les supports de valeurs décrivent la trahison ou promotion de valeurs $v$, appartenant à l'ensemble $V$ des valeurs morales connues mentionnées dans $\mathcal{O}_v$, par des actions dans certaines situations. Ensuite, une relation de hiérarchie permet de définir quelles valeurs constituent des sous-valeurs de valeurs plus générales.

Les valeurs morales de $V$ sont organisées au sein d'un système de valeurs que nous définissons de la manière suivante :

**Définition 22** (Système de valeurs morales)**.** *Un* système de valeurs morales *est constitué d'un ensemble de valeurs morales $V$ décrit dans $\mathcal{O}_v$ et de* relations de hiérarchies *décrites dans $\mathcal{O}_{hv}$.*

Une relation hiérarchique entre les valeurs se définit grâce au prédicat $subvalue(v_1, v_2)$, signifiant que la valeur $v_1 \in V$ est une sous-valeur de $v_2 \in V$. Il est ainsi possible, par exemple, de définir que l'indulgence est une sous-valeur de la bienveillance (voir I.3.2).

**Exemple 15.** *Pour* Robin Hood*, la valeur generosity est une sous-valeur de benevolence. Cela signifie qu'agir généreusement, pour* Robin Hood*, est également bienveillant.*

### III.5.2 Supports de valeurs

L'ensemble $VS$ des supports de valeurs décrit comment les actions, dans des situations, peuvent promouvoir ou trahir une valeur $v$ de l'ensemble $V$ des valeurs morales de $\mathcal{O}_v$. Les supports de valeurs permettent ainsi d'exprimer comment les valeurs peuvent qualifier le comportement d'un agent.

**Définition 23.** *Un* support de valeur *est un quintuplet $\langle a, w, w', v, sv\rangle \in VS$ associant à une action $a \in A \cup any$ une valuation de support $sv \in SV$ vis-à-vis de la valeur $v \in V$ dans le cas où la condition $w$ est vraie dans la situation courante $s$ et dans le cas où la condition $w'$ est satisfaite par l'exécution de l'action $a$. Cette valuation de support $sv$ indique à quel point la réalisation de l'action promeut ou trahit la valeur.*

La valeur particulière *any* pour $a$ signifie que toute action de $A$ peut-être concernée par ce support. Pour définir précisément le sens d'une valeur en pratique dans le système, il est nécessaire de décrire un ensemble $VS$ de supports de valeurs. Nous supposons ici qu'un tel ensemble peut être défini par l'utilisateur ou le concepteur de l'application, donnant ainsi à l'agent une part de sa théorie du bien (voir I.1.1).

Notons qu'à ce stade, les supports de valeurs ne permettent pas de savoir si une action est morale ou non. Un agent peut ainsi détenir des connaissances sur des valeurs dans le seul but d'évaluer le support des actions à celles-ci sans y attacher une quelconque connotation morale.

**Exemple 16.** *Dans la morale de* Robin Hood *prise en exemple, quelques-uns des supports de valeurs que nous emploierons sont les exemples $vs_1$, $vs_2$, $vs_3$ et $vs_4$ suivants :*

$vs_1$ : *« Effectuer toute action qui rend un agent initialement pauvre non-pauvre promeut la valeur de générosité » s'écrit*

$$vs_1 = \langle any, poor(\mathsf{a}), \neg poor(\mathsf{a}), generosity, promote\rangle$$

$vs_2$ : *« Rendre un agent pauvre trahit la valeur de générosité » s'écrit*

$$vs_2 = \langle any, \neg poor(\mathsf{a}), poor(\mathsf{a}), generosity, defeat\rangle$$

*$vs_3$ : « Taxer un agent non riche trahit la valeur de générosité », s'écrit*

$$vs_3 = \langle tax(\mathsf{a}), \neg rich(\mathsf{a}), \top, generosity, defeat \rangle$$

*$vs_4$ : « Voler trahit l'honnêteté », s'écrit*

$$vs_4 = \langle steal(\mathsf{a}), \top, \top, honesty, defeat \rangle$$

*$vs_3$ et $vs_4$ s'appliquent ici dans le cas ou les conséquences de l'action impliquent $\top$, c'est-à-dire pour toutes conséquences connues. De même, $vs_4$ décrit la trahison de l'honnêteté dans toute situation.*

### III.5.3 Évaluation des supports de valeurs

Le support ou la trahison de valeurs par les actions étant dépendant de la situation, il est nécessaire de produire l'ensemble $A_v$ des actions évaluées par les supports de valeurs à partir de la connaissance de l'agent sur les valeurs et des états mentaux représentant la situation perçue par l'agent. À cet effet nous définissons la fonction $VE$ d'évaluation des actions par les supports de valeurs comme suit :

**Définition 24** (Évaluation par les supports de valeurs)**.** *La fonction d'*évaluation par les supports de valeurs *$VE$ produit l'ensemble des actions évaluées par les supports de valeurs $\mathcal{A}_v$ en fonction de l'ensemble des supports de valeurs $VS$, des connaissances sur les actions $A$ et de la situation courante représentée par les croyances $\mathcal{B}$ et les désirs $\mathcal{D}$.*

$$VE : A \times 2^{\mathcal{B} \times \mathcal{D}} \times VS \rightarrow \mathcal{A}_v$$

$$\mathcal{A}_v \subseteq A \times V \times SV$$

$$(a, v, sv) \in \mathcal{A}_v : a \in A, v \in V, sv \in SV$$

Pour qu'un support de valeur $vs = \langle a, s, w, w', sv \rangle$ intervienne dans l'évaluation d'une action, il est nécessaire que la proposition $w$ soit vérifiée dans la situation courante $s$. La production de l'ensemble $\mathcal{A}_v$ par l'évaluation des supports de valeurs est calculée à l'aide du prédicat $support_value(a, s, v, sv)$ signifiant que l'action $a \in A$ dans le contexte $s$ est associée à une valuation $sv$ indiquant à quel point cela va trahir ou promouvoir la valeur $v \in V$ :

$$\mathcal{A}_v = \{(a, v, sv) : support_value(a, s, v, sv)\}$$

$$\begin{aligned} support_value(s, a, v, sv) \equiv \exists vs : vs = \langle a, w, w', v, sv \rangle \\ \wedge\, s \rightarrow w \wedge postcond(a, s, \psi) \wedge \psi \rightarrow w' \end{aligned}$$

**Exemple 17.** *Dans la situation courante perçue précédemment dans notre exemple, puisque* Robin Hood *sait que* Friar Tuck *est pauvre, il peut en déduire que give*(Friar Tuck) *est une action qui promeut la valeur generosity tandis que les actions steal*(Friar Tuck) *et tax*(Friar Tuck) *trahissent cette même valeur. Notons qu'il est cohérent dans ce modèle d'évaluer par les supports de valeurs des actions qui sont pourtant évaluées comme impossibles par la fonction d'évaluation de la possibilité $PE$. C'est le cas dans cet exemple pour steal*(Friar Tuck).

La relation définie précédemment entre les valeurs est employée lors de l'évolution des supports de valeurs de la manière suivante :

$$subvalue(v_1, v_2) \equiv support_value(a, s, v_1, sv) \rightarrow support_value(a, s, v_2, sv)$$

### III.5.4 Règles morales

Afin d'évaluer le caractère moral d'une action, l'agent dispose d'un ensemble de règles associant une valuation morale $mv \in MV$ à des ensembles d'actions dans des contextes définis.

**Définition 25.** *Une* règle morale *est un tuple $\langle a, w, w', vc, mv\rangle \in MR$ où $a \subset A \cup \{any\}$ est un ensemble d'actions pour lequel la règle morale associe une valuation morale $mv \in MV$, $w$ est une situation hypothétique dans laquelle la règle est valide, $w'$ est une situation hypothétique décrivant des conséquences de $a$ nécessaires pour que la règle s'applique, $vc$ est l'ensemble de* conditions de valeurs *permettant de définir des valuations de support minimales ou maximales nécessaires pour que la règle soit appliquée, avec $vc = \{\langle v, sv, m\rangle\}$ où $v \in V$ est une valeur, $sv \in SV$ une valuation de support servant de valeur seuil et $m \in \{min, max\}$ un sens de comparaison. Ici, $a = \{any\}$ signifie que la règle est valable pour toute action de $A$.*

Ces règles morales associent à une action ou un ensemble d'actions, une valuation morale. Ces règles peuvent être décrites de diverses manières afin de capturer les approches présentées en section 1 :

— Une approche *déontologique* est généralement décrite par des règles spécifiques décrivant des devoirs ou des interdits (e.g. « Les journalistes doivent refuser toute faveur aux publicitaires, donateurs ou groupes d'intérêt et résister aux pressions internes ou externes qui tenteraient de les influencer » [1]). Dans ce modèle, cela est représenté en désignant directement l'ensemble d'actions auquel la valuation morale est affectée dans le cas où la situation courante satisfait l'ensemble d'états décrit par $w$ (voir $mr'_2$ dans l'exemple 18) ;

— une approche *conséquentialiste* utilise à la fois des règles générales et spécifiques concernant les états et les conséquences (e.g. « Tout médecin doit s'abstenir, même

1. Extrait de (Professional Journalists, 2014), section « Act Independently ».

en dehors de l'exercice de sa profession, de tout acte de nature à déconsidérer celle-ci. » [2]). L'évaluation de la moralité des actions dépend alors des connaissances de l'agent sur les conséquences de ses actions. Une telle approche peut être représentée par la description dans $w'$ de l'ensemble d'actions concernées par la règle en désignant ici une partie de leurs postconditions (voir $mr_2''$ dans l'exemple 18) ;

— Une approche *vertueuse* utilise des règles générales s'exprimant sur des valeurs morales (e.g. « Il est moral d'être généreux »). L'évaluation de la moralité d'une action dépend alors de l'ensemble des supports des valeurs et sous-valeurs employées dans les règles morales traitant de cette action. Dans ce modèle, une telle approche nécessite la description de cet ensemble d'actions en désignant celles qui, dans l'ensemble $\mathcal{A}_v$ des actions évaluées par les valeurs, se voient attribuer une certaine valuation de support vis-à-vis d'un ensemble de valeurs (voir $mr_2$ dans l'exemple 18).

Une règle peut s'appliquer de manière plus ou moins spécifique à un ensemble d'action $a$, une situation $w$ des conséquences $w'$ et des contraintes de valeurs $vc$. Par exemple « Il est immoral d'agir injustement » est plus général (s'applique à un plus grand nombre d'actions et de situations) que « Juger un meurtrier en tenant compte de sa religion, sa couleur de peau, son origine ethnique ou ses opinions politiques est immoral ».

**Exemple 18.** *Dans la section précédente, nous avons décrit les valeurs connues de* Robin Hood. *Pour lui permettre d'évaluer la moralité des actions à sa disposition, nous pouvons le doter de règles morales telles que :*

$mr_1$ : *« Trahir la valeur honnêteté est modérément immoral » qui est une règle morale d'approche vertueuse affectant la valuation morale immoral à toute action pour laquelle il existe au moins un support de valeur indiquant que dans ce contexte une telle action trahit la valeur d'honnêteté ;*

$$mr_1 = \langle \{any\}, \top, \top, \{\langle honesty, defeat, max\rangle\}, immoral\rangle$$

$mr_2$ : *« Effectuer une action généreuse est très moral » qui est une règle générale faisant référence à une valeur comme la règle précédente ;*

$$mr_2 = \langle \{any\}, \top, \top, \{\langle generosity, promote, min\rangle\}, very\ moral\rangle$$

$mr_2'$ : *« Donner à un agent pauvre est très moral » est une alternative déontologique à la règle précédente, désignant directement une action sans exprimer sa moralité à l'aide de valeurs ;*

$$mr_2' = \langle \{give(\mathsf{a})\}, poor(\mathsf{a}), \top, \emptyset, very\ moral\rangle$$

$mr_2''$ : *« Effectuer une action envers un agent pauvre ayant pour conséquence de le rendre non-pauvre est très moral » est une alternative conséquentialiste des deux règles*

2. Code de déontologie médicale, article 31.

*précédentes portant sur les conséquences connues des actions pour en évaluer leur moralité ;*

$$mr_2'' = \langle \{any\}, poor(\mathsf{a}), \neg poor(\mathsf{a}), \emptyset, very\ moral \rangle$$

$mr_3$ : *« Effectuer une action dont au moins l'une des conséquences rend possible une action très morale est moral » est un exemple de règle permettant de raisonner à plus long terme ;*

$$mr_3 = \langle \{any\}, \top, \exists mr \exists a : possible(a, possible) \wedge goodness(a, w', mr, very\ moral), \emptyset, moral \rangle$$

$mr_4$ : *« Trahir la valeur de générosité est immoral » qui, comme pour la règle $mr_1$, fait de la générosité une vertu contre laquelle il est immoral d'agir.*

$$mr_4 = \langle \{any\}, \top, \top, \{\langle generosity, defeat, max \rangle\}, immoral \rangle$$

### III.5.5 Évaluation de la moralité

Les règles morales affectant des valuations morales aux actions dans certains contextes sont employées par la fonction d'évaluation morale $ME$ (pour *morality evaluation*) afin de produire l'ensemble des actions moralement évaluées :

**Définition 26** (Fonction d'évaluation de la moralité)**.** *La* fonction d'évaluation de la moralité $ME$ *évalue les actions de $A$ au regard d'une théorie du bien $GK$ pour produire l'ensemble des actions moralement évaluées $\mathcal{A}_m$. Elle est définie comme :*

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

L'attribution d'une valuation à une action par une règle est représentée par le prédicat $goodness(a, s, mr, mv)$ où $a \in A$ est l'action évaluée, $s$ la situation dans laquelle est effectuée l'évaluation et $mr \in MR$ la règle morale lui affectant la valuation morale $mv \in MV$ :

$$\begin{aligned} goodness(a, s, mr, mv) \equiv & mr = \langle a', w, w', vc, mv \rangle \\ & \wedge (a \in a' \vee a' = \{any\}) \\ & \wedge s \rightarrow w \\ & \wedge postcond(a, s, \psi) \\ & \wedge \psi \rightarrow w' \\ & \wedge (\forall \langle v, sv, m \rangle \in vc, \\ & \quad support_value(a, s, v, sv') \\ & \quad \wedge m = max \rightarrow sv' \leqslant sv \\ & \quad \wedge m = min \rightarrow sv' \geqslant sv) \end{aligned}$$

$$\mathcal{A}_m = \{(a, mr, mv) | a \in A : goodness(a, s, mr, mv)\}$$

**Exemple 19.** *Dans le cadre de notre exemple,* Robin Hood *est ainsi capable d'associer à l'action $a_2$ (celle consistant à donner ses richesses) la valuation morale "très moral" au regard d'une règle ($mr_2$, $mr'_2$ ou $mr''_2$ selon l'approche adoptée) si l'agent ciblé par l'action est pauvre au moment de l'évaluation.*

*Ainsi, dans la situation décrite précédemment, il obtient l'évaluation suivante :*

$$\langle give(\text{Friar Tuck}), mr_2, very\ moral\rangle \in \mathcal{A}_m$$

*De même, puisque le support de valeur $vs_4$ et la règle morale $mr_1$ sont valables dans tout état du monde, il obtient nécessairement, par exemple, l'évaluation suivante :*

$$\langle steal(\text{Prince John}), mr_1, immoral\rangle \in \mathcal{A}_m$$

*Notons ici que dans ces évaluations il n'est nulle part fait mention du caractère désirable ou même possible de l'action de donner à* Friar Tuck *:* Robin Hood *sait qu'elle est conforme à sa théorie du bien, même si elle est impossible à réaliser ou contraire à ses objectifs.*

Ce modèle d'évaluation morale vient ainsi fournir le dernier élément nécessaire au raisonnement éthique qui va permettre de comparer toutes les actions évaluées pour distinguer l'action juste.

## III.6 Évaluation de l'éthique des actions

Les sections III.4 et III.5 ont montré comment évaluer le caractère possible, désirable, moral ou non des actions de $A$ et comment leur affecter des valuations de désirabilité, de possibilité et de moralité. Nous allons montrer ici comment l'agent peut raisonner sur ces informations afin de juger du caractère éthique ou non des actions au regard d'une théorie du juste dans une situation. Cette section présente dans cette optique la partie du modèle (voir figure III.6) permettant de représenter et employer dans le raisonnement les différents constituants de la théorie du juste, ou connaissance du juste (voir section 5) notée $RK$ (pour *Rightness Knowledge*) constituée des principes éthiques et des préférences éthiques, soit $RK = P \cup \succ_e$.

Ce modèle est défini de la manière suivante :

**Définition 27** (Modèle d'évaluation de l'éthique)**.** *Un* modèle d'évaluation de l'éthique *(ou Ethics Evaluation Model) EEM produit l'ensemble des actions justes $\mathcal{A}_r$ étant donné les ensembles d'actions évaluées par les modèles d'évaluation de la possibilité $\mathcal{A}_p$, de la désirabilité $\mathcal{A}_d$, de la moralité $\mathcal{A}_m$ et sa théorie du juste $RK$. Nous définissons ce modèle comme :*

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

*où $P$ est la base de connaissance sur les principes éthiques de l'agent, $\succ_e \subseteq P \times P$ est un ensemble de relations de préférences représentant un ordre partiel sur ces principes, $\mathcal{A}_r$ est l'ensemble des actions justes. $EE$ une fonction d'évaluation éthique et $J$ une fonction de jugement.*

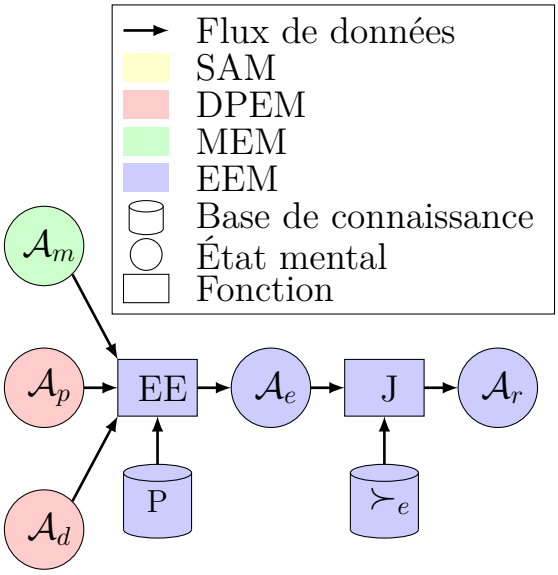


FIGURE III.6 – Modèle d'évaluation de l'éthique

## III.6.1 Principes éthiques

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

**Définition 28** (Principe éthique)**.** *Un principe éthique (ou Ethical Principle) $p \in P$ est une fonction décrivant la justesse d'une action évaluée en termes de possibilité, désirabilité et moralité dans une situation donnée. Elle est définie comme :*

$$p : A \times 2^{\mathcal{B}\times\mathcal{D}} \times 2^{\mathcal{A}_d} \times 2^{\mathcal{A}_m} \times 2^{\mathcal{A}_p} \to \{\top, \bot\}$$

*avec $a \in A$ l'action évaluée, $s$ la situation dans laquelle l'action est évaluée et $p \in P$ le principe éthique.*

Ce modèle de jugement permet à un agent d'employer plusieurs principes présents au sein de l'ensemble $P$ des principes connus de l'agent.

Bon nombre de principes éthiques sont présentés dans la littérature philosophique et quelques-uns ont déjà fait l'objet de travaux visant à les représenter formellement et les implémenter dans des langages de programmation logique.

**Exemple 20.** *Notre agent* Robin Hood *pourrait ainsi connaître divers principes tels que la doctrine du double effet ou l'éthique d'Aristote (voir section I.1.2). Il est également*

*possible de formuler des principes éthiques plus simples permettant de distinguer l'action juste à effectuer dans des situations courantes. Par exemple, considérons que* Robin Hood *possède les principes suivants : "Si une action est désirable sans être en même temps indésirable, morale ou très morale sans être en même temps immorale ou très immorale et possible, alors elle est juste" que nous appellerons $p_1$, "Si une action est possible et évaluée comme morale ou très morale par au moins une règle morale, alors elle est juste" que nous appellerons $p_2$, l'éthique d'Aristote $p_3$ (voir I.1.2.3) et la Doctrine du Double Effet $p_4$ (voir I.1.2.4).*

## III.6.2 Fonction d'évaluation de l'éthique

Les principes éthiques sont employés par la fonction d'évaluation éthique $EE$ qui évalue la conformité des actions au regard de $P$, compte tenu de leurs évaluations au regard de la moralité, de la désirabilité et de la possibilité. La fonction d'évaluation éthique $EE$ renvoie l'évaluation de toutes les actions désirables ($\mathcal{A}_d$), réalisables ($\mathcal{A}_c$) ou morales ($\mathcal{A}_m$) étant donné l'ensemble $P$ des principes éthiques connus. La conformité d'une action au regard d'un principe éthique est représentée par le prédicat $ethical_conform(a, s, p, ev)$.

**Définition 29** (Fonction d'évaluation éthique)**.** *La fonction d'évaluation éthique $EE$ renvoie l'évaluation de toutes les actions désirables ($\mathcal{A}_d$), réalisables ($\mathcal{A}_c$) ou morales ($\mathcal{A}_m$) étant donné l'ensemble $P$ des principes éthiques connus. Elle est définie comme :*

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

*L'ensemble des actions associées à un principe éthique et l'évaluation de cette action par celui-ci forme l'ensemble $\mathcal{A}_e$ des actions éthiquement évaluées.*

$$\mathcal{A}_e = \{(a, p, ev) \in A \times P \times \{\top, \bot\},\ s.t.\ ethical_conform(a, s, p, ev)\}$$

**Exemple 21.** *Prenons l'exemple d'une situation dans laquelle* Robin Hood*, pauvre, serait dans l'incapacité de donner à d'autres agents pauvres. Dans une telle situation, la fonction d'évaluation de la possibilité montre que l'action $a_1$ (voler) sur un agent riche est possible, que l'action $a_2$ (donner) est impossible puisque* Robin Hood *est pauvre, que l'action $a_3$ est impossible puisque* Robin Hood *n'est pas noble et $a_4$ est possible. La fonction d'évaluation de la désirabilité montre que seule l'action $a_4$ sur* Marian *est désirable. La fonction d'évaluation morale indique que $a_1$ envers un agent riche est immorale selon $vs_3$ et $mr_1$, mais aussi qu'elle est morale au regard de $vs_4$, $mr_2$ et $mr_3$, que $a_2$ est morale au regard de $vs_4$ et $mr_2$ et que $a_3$ sur un agent ni pauvre ni riche est immorale au regard de $vs_6$ et $mr_4$. La table III.1 récapitule ces résultats en précisant les éléments du modèle à l'origine de l'obtention des valuations morales*

*Aucune action n'étant à la fois morale, désirable, possible, non-immorale et non-indésirable, le principe $p_1$ ne permet pas de considérer l'une de ces actions comme juste*

| Action | Évaluations | | | Évaluations éthiques | | | |
|---|---|---|---|---|---|---|---|
| | $dv$ | $pv$ | $mv$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ |
| $a_1$ : *steal*(a) | *neutral* | *possible* | *immoral* et *moral* | $\bot$ | $\top$ | $\bot$ | $\bot$ |
| $a_2$ : *give*(a) | *neutral* | *impossible* | *very moral* ssi *poor*(a) | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| | | | *neutral* sinon | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| $a_3$ : *tax*(a) | *neutral* | *impossible* | *immoral* | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| $a_4$ : *court*(a) | *desirable* si a=Marian *neutral* sinon | *possible* | *neutral* | $\bot$ | $\bot$ | $\top$ | $\top$ |

TABLE III.1 – Tableau récapitulatif des évaluations lorsque Robin Hood est pauvre.

*dans ce contexte. Pour $p_2$ en revanche, $a_1$ est la seule action juste, puisqu'elle est possible et motivée par au moins une règle morale. Du point de vue de $p_3$ et $p_4$, $a_4$ est la seule action juste puisqu'elle est l'alternative la moins condamnable du point de vue moral. Ainsi* Robin Hood *est capable de discerner quelle action est conforme à l'éthique ou non du point de vue de diverses doctrines.*

### III.6.3 Préférences éthiques

Maintenant qu'il est possible de connaître la conformité des actions à un ensemble de principes, nous cherchons à définir l'action ou les actions justes, du point de vue de l'ensemble $P$ des principes éthiques, au regard d'un ensemble de préférences noté $\succ_e$ défini de la manière suivante :

**Définition 30** (Préférence éthique)**.** *Une relation de* préférence éthique *notée $\succ_e$ est une relation binaire interne sur $P$, transitive et asymétrique. $p_1 \succ_e p_2$ signifie que « $p_1$ est préféré à $p_2$ ».*

L'ensemble des préférences éthiques d'un agent permet de définir l'importance de chaque principe dans son éthique.

**Exemple 22.** *Nous définissons les préférences de* Robin Hood *suivantes : $p_1 \succ_e p_2$, $p_2 \succ_e p_3$ et $p_3 \succ_e p_4$. Cela signifie que* Robin Hood *considère comme juste les actions satisfaisant $p_1$, qu'en cas d'égalité il sélectionne celles qui satisfont $p_2$, puis $p_3$, puis $p_4$.*

### III.6.4 Fonction de jugement

Considérant les divers éléments de l'éthique d'un agent, la fonction de jugement $J$ permet de calculer l'ensemble des actions justes $\mathcal{A}_r$ (pour *Rightful Actions*). Elle est définie de la manière suivante :

**Définition 31** (Fonction de jugement)**.** *La* fonction de jugement *$J$ retourne l'ensemble des actions justes $\mathcal{A}_r$ en fonction de l'ensemble des actions éthiquement évaluées et l'ensemble des préférences éthiques $\succ_e$.*

$$J : 2^{\mathcal{A}_e} \times 2^{\succ_e} \to 2^{\mathcal{A}_r}$$

$$rightness(a, s, p) \equiv \nexists a' \in A, \nexists p' \in P \\ ethical_conform(a', s, p', \top) \wedge p' \succ_e p$$

$$\mathcal{A}_r = \{a \in A : \exists p \ s.t. \ rightness(a, s, p)\}$$

En pratique, l'ensemble des actions justes $\mathcal{A}_r$ produit par la fonction de jugement est calculé à l'aide du prédicat $rightness(a, s, p)$ où $a \in A$ est l'action considérée juste et $p \in P$, le principe le plus élevé dans l'ordre de préférence satisfait dans la situation $s$ par cette action.

**Exemple 23.** *Par exemple, une fonction de jugement intuitive peut considérer qu'une action est juste si elle est celle qui satisfait le principe éthique le plus élevé dans l'ordre des préférences parmi l'ensemble des principes satisfaits par au moins une action.*

*Dans le cas de* Robin Hood*, dans la situation évoquée en III.6.2 et avec les préférences définies en III.6.3, la fonction de jugement fondée sur l'ordre lexicographique produirait* $\mathcal{A}_r = \{a_1\}$ *puisqu'aucune autre action n'est conforme à un principe préféré à* $p_2$ *ou à* $p_2$ *lui-même.*

La fonction de Jugement $J$ est le dernier élément de l'ensemble formant le *modèle d'évaluation de l'éthique* d'un agent permettant, à partir des informations fournies par le modèle d'évaluation de la possibilité de de la désirabilité, et le modèle d'évaluation de la moralité, de déterminer l'ensemble des actions justes. Ce modèle étant le dernier du modèle de jugement (voir III.7), les sections suivantes visent à illustrer son fonctionnement au travers d'un exemple récapitulatif, puis de montrer comment il peut être employé au sein d'un processus de décision.

Les rectangles représentent des fonctions encapsulant des sous-parties du raisonnement de l'agent. Les cylindres représentent les connaissances prises en paramètres du modèle. Ces connaissances sont données à l'initialisation et représentent l'éthique et la morale de l'agent et ses connaissances sur ses actions. Les cercles représentent des états mentaux et, à l'inverse des connaissances, sont produits et mis à jour par les fonctions.

## III.7 Exemple récapitulatif

Nous proposons d'illustrer ici le fonctionnement de l'intégralité du modèle à travers l'exemple d'un nouvel agent Sheriff of Nottingham qui n'est pas noble. L'objectif est de montrer comment cet agent doté de règles morales différentes peut employer le même modèle de jugement et la même ontologie $\mathcal{O}$ pour prendre ses décisions.

Premièrement, nous considérons que Sheriff of Nottingham est doté de l'ensemble $VS$ comprenant les supports $vs_1$, $vs_2$, $vs_3$ et $vs_4$ présentés en section III.5 et qu'il partage avec Robin Hood, auxquels vient s'ajouter $vs_5$ :

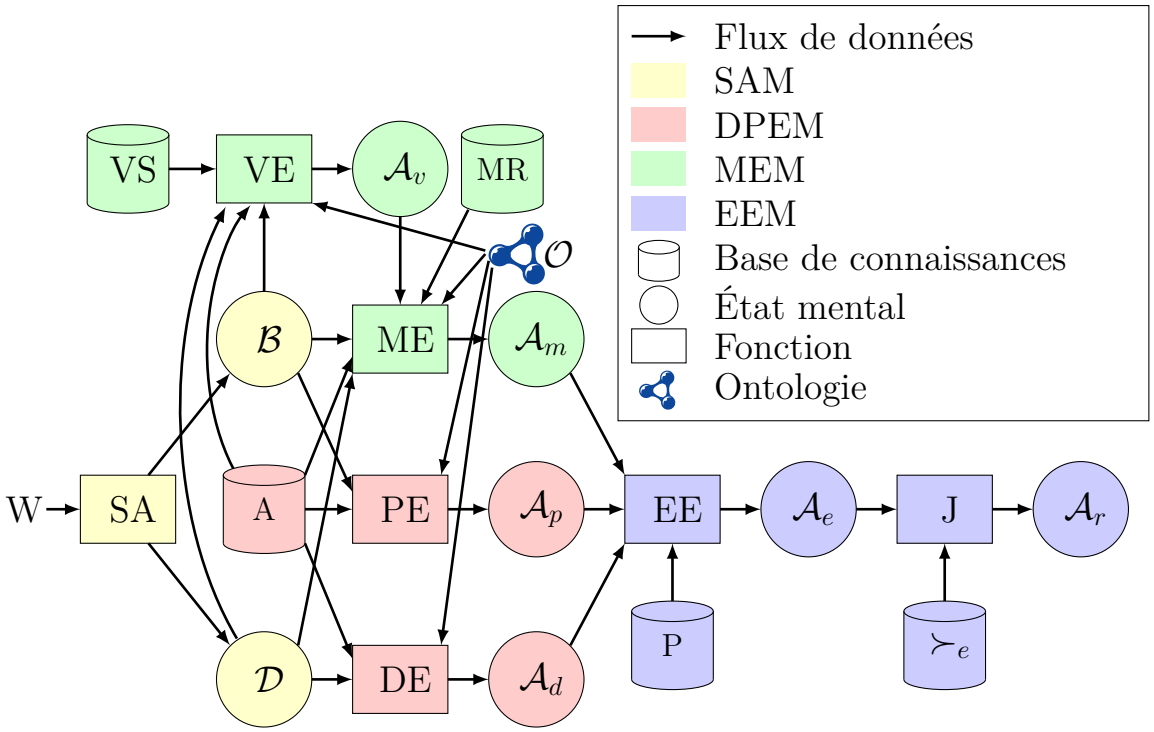


FIGURE III.7 – Représentation globale du modèle de jugement

$vs_5$ : « Voler un noble trahit la valeur d'obéissance », est un support spécifique à une action restreint aux situations dans lesquelles la cible est noble

$$vs_5 = \langle steal(\mathsf{a}), noble(\mathsf{a}), \top, obedience, defeat \rangle$$

Autrement dit, Sheriff of Nottingham a en commun avec Robin Hood ses connaissances sur les valeurs de générosité (grâce à $vs_1$, $vs_2$ et $vs_3$) et d'honnêteté (grâce à $vs_4$), auxquelles vient s'ajouter sa connaissance de l'obéissance.

La morale de Sheriff of Nottingham est exprimée à l'aide d'un ensemble de règles $MR$ comprenant $mr_3$ qu'il partage avec Robin Hood et la règle $mr_5$ :

$mr_5$ : « Toute action trahissant la valeur obéissance est très immorale » qui est une règle morale d'approche vertueuse valable en toute situation et pour toute action.

$$mr_5 = \langle \{any\}, \top, \top, \{\langle obedience, defeat, max \rangle\}, very\ immoral \rangle$$

Cet agent est également doté d'un ensemble $D$ de désirs composé du désir $d_1 = desire(rich(\text{Sheriff of Nottingham}))$ similaire à celui de Prince John et du désir $d_2 = desire(done(court(\text{Marian})))$ identique à celui de Robin Hood.

Dans une situation où Sheriff of Nottingham est ni pauvre ni riche, les actions évaluées comme désirables sont $court(\text{Marian})$, qui est désirable en toute situation, ainsi que $steal(\mathsf{a})$ et $tax(\mathsf{a})$ qui permettraient à Sheriff of Nottingham de devenir riche.

L'évaluation de la possibilité génère l'ensemble des actions $\mathcal{A}_p$ dans lequel sont pos-

sibles les actions $steal(\mathsf{a})$ pour tout agent a du système qui n'est pas pauvre, $give(\mathsf{a})$ pour tout agent a du système qui n'est pas riche et $court(\mathsf{Marian})$, qui est possible en toutes circonstances. L'action $tax(\mathsf{a})$ est évaluée comme impossible puisque Sheriff of Nottingham n'est pas noble.

L'évaluation de la moralité révèle qu'il serait très immoral d'effectuer l'action $steal(\mathsf{a})$ si cet agent est noble, en raison de la règle $mr_5$ et du support de valeur $vs_5$.

En employant les mêmes principes éthiques que pour Robin Hood, nous obtenons les résultats résumés par le tableau III.2.

| Action | Évaluations | | | Évaluations éthiques | | | |
|---|---|---|---|---|---|---|---|
| | $dv$ | $pv$ | $mv$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ |
| $a_1 : steal(\mathsf{a})$ | *desirable* | *possible* ssi $\neg poor(\mathsf{a})$ | *very immoral* si $noble(\mathsf{a})$, *neutral* sinon | $\bot$* <br> $\bot$* | $\bot$* <br> $\bot$* | $\bot$* <br> $\bot$* | $\bot$* <br> $\bot$* |
| $a_2 : give(\mathsf{a})$ | *neutral* | *possible* ssi $\neg rich(\mathsf{a})$ | *neutral* | $\bot$* | $\bot$* | $\bot$* | $\bot$* |
| $a_3 : tax(\mathsf{a})$ | *desirable* | *impossible* | *neutral* | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| $a_4 : court(x)$ | *desirable* ssi a=Marian | *possible* | *neutral* | $\bot$ | $\bot$ | $\top$ | $\top$ |

* Évaluations indiquées dans le cas où l'action est possible

TABLE III.2 – Tableau récapitulatif des évaluations lorsque Sheriff of Nottingham n'est ni pauvre ni riche.

Compte tenu des ensembles d'actions évaluées par les modèles d'évaluation de la possibilité $\mathcal{A}_p$, de la désirabilité $\mathcal{A}_d$, et de la moralité $\mathcal{A}_p$ et des principes éthiques ordonnés par les préférences, l'action juste dans cette situation pour Sheriff of Nottingham est de courtiser Marian pusqu'il s'agit de la seule action satisfaisant le principe éthique $p_3$, qui est ici le principe le plus élevé dans l'ordre de préférences de l'ensemble des principes satisfaits par au moins une action.

## III.8 Synthèse

Ce chapitre a présenté un modèle de jugement et montré comment il permet d'employer les éléments classiques du modèle BDI pour juger du caractère éthique des actions connues de l'agent en fonction d'une théorie du bien et d'une théorie du juste passées en argument.

Ce modèle prend en paramètres un ensemble de connaissances sur les valeurs morales, les règles morales, les principes éthiques et les relations de préférences entre ces principes. Il évalue, de manière classique, les actions qui sont possibles et les actions qui sont désirables dans la situation courante et ajoute une évaluation de ces actions par rapport à des valeurs morales, puis par rapport à des règles morales pour fournir un ensemble $\mathcal{A}_m$ des actions moralement évaluées. Le modèle d'évaluation de l'éthique évalue ensuite la conformité des actions à des principes éthiques et en déduit l'ensemble des actions justes, c'est-à-dire celles qui satisfont le mieux l'ensemble des principes éthiques, compte

tenu des préférences éthiques.

Cette connaissance des actions justes, c'est-à-dire de l'ensemble $\mathcal{A}_r$ dans un contexte donné, peut être directement employé comme processus de sélection des intentions. Nous désignons par la suite sous le terme d'*agent guidé par l'éthique* tout agent dont le processus de décision ne sélectionne des intentions que parmi les actions présentes dans $\mathcal{A}_r$ au regard de son modèle de jugement.

Nous avons illustré le fonctionnement de ce modèle de jugement à l'aide d'exemples faisant intervenir deux agents aux morales et éthiques légèrement différentes et connaissant peu d'actions. En montrant comment le modèle de jugement pouvait employer ces connaissances, nous avons montré que les jugements des agents produisaient des résultats différents. En employant leur modèle de jugement comme processus de décision, ils devraient donc avoir des comportements différents et il serait intéressant de pouvoir constater cette différence lors de l'observation du comportement de l'autre. De plus, pour envisager de collaborer avec d'autres agents ayant des comportements différents, il semble intuitivement intéressant de permettre aux agents de se construire une représentation mentale de la conformité de ces comportements à une théorie du bien et une théorie du juste. Le chapitre suivant s'attache donc à étendre le modèle présenté ici pour permettre le jugement de comportements observés.

# CHAPITRE IV

# Jugement des autres

**IV.1 Préambule** . . . . . **80**
IV.1.1 Notations . . . . . 80
IV.1.2 Scénario illustratif . . . . . 80
**IV.2 Typologie des jugements** . . . . . **81**
IV.2.1 Niveau d'information du jugement . . . . . 81
IV.2.2 Temporalité du jugement . . . . . 83
**IV.3 Images de la moralité et de l'éthique d'un agent** . . . . . **84**
IV.3.1 Image de la moralité des actions d'un agent . . . . . 84
IV.3.2 Image de l'éthique des actions d'un agent . . . . . 88
**IV.4 Coopération fondée sur le jugement des autres** . . . . . **92**
IV.4.1 Construction de la confiance . . . . . 92
IV.4.2 Éthique de la confiance . . . . . 93
IV.4.3 Utilisation de la confiance pour la coopération éthique. . . . 94
**IV.5 Synthèse** . . . . . **96**

Le chapitre précédent propose un modèle de jugement des actions d'un agent avec ses propres connaissances. Ce chapitre étend le modèle de jugement pour permettre à un agent de raisonner sur le caractère moral et éthique du *comportement* des autres agents du système, non seulement avec ses propres théories du bien et du juste, mais également avec toute connaissance contextuelle du bien ou du juste qu'il serait en mesure d'obtenir. Notre objectif est de montrer comment le processus de jugement peut être employé, cette fois-ci non pas comme tout ou partie d'un processus de décision mais pour évaluer le comportement des autres et devenir un élément central dans un modèle de coopération fondée sur l'éthique. Un tel système de coopération permet de faire un premier pas vers la formation de collectifs d'agents dont les éthiques conduisent à des comportements acceptables du point de vue des autres. La coopération éthique conduit donc un agent à utiliser le jugement des actions décrit précédemment selon deux optiques : un jugement du comportement des autres agents, c'est-à-dire le jugement des

actions conduites par les autres agents avec lesquels il coopère, et un jugement de ses propres actions en direction des autres agents, reposant sur une description, dans ses règles morales, de bonnes conduites à adopter envers les autres agents en prenant en compte dans le contexte de ces règles des connaissances acquises sur leur comportement.

# IV.1 Préambule

L'objet de ce chapitre étant de montrer comment employer le modèle de jugement défini au chapitre précédent afin de juger les actions des autres agents et prendre en compte ce jugement dans la décision, nous introduisons ici de nouvelles notations et un exemple illustratif afin de clarifier notre propos.

## IV.1.1 Notations

Chaque agent disposant de son propre processus de reconnaissance de situation, ses propres états mentaux, sa propre théorie du bien et sa propre théorie du juste, nous abordons dans ce chapitre la possibilité de raisonner sur les états mentaux et connaissances des autres dans un système multi-agent. Que ces croyances soient obtenues par communication, observation ou tout autre procédé, nous les attribuerons de la même façon à un agent $\mathsf{a} \in Ag$, avec $Ag$ l'ensemble des agents du système.

Afin d'éviter toute confusion entre les éléments du modèle de l'agent qui effectue le raisonnement et les divers agents qu'il observe, nous indicerons, dans ce chapitre et dans les suivants, tout élément décrit au chapitre précédent à l'aide du nom de l'agent auquel il appartient. Ainsi $VS_{\mathsf{a}}$ est l'ensemble des supports de valeur de l'agent $\mathsf{a}$, $\mathcal{D}_{\mathsf{a}}$ l'ensemble de ses désirs, etc. De manière analogue pour chaque prédicat défini au chapitre précédent, nous augmentons son arité afin de donner en paramètre l'identité de l'agent jugé.

De même, dans le but de prendre en compte des successions de jugements d'actions, nous introduisons une dimension temporelle aux croyances et états mentaux du modèle. Chaque état mental ou croyance comporte ainsi l'instant indiquant le moment de sa production. Par exemple, la croyance en l'exécution par un agent $\mathsf{a}$' d'une action $a_k$ à un instant $t$ est noté $done(\mathsf{a}', a_k, t)$.

## IV.1.2 Scénario illustratif

En reprenant le scénario illustratif employé au chapitre précédent, nous allons montrer comment, en employant son modèle de jugement, l'agent Robin Hood peut évaluer le comportement de Friar Tuck et de Prince John et comment les informations issues de ces évaluations permettent de définir dans les connaissances du bien et du juste de Robin hood une éthique de la coopération.

Par souci de simplicité, nous considérons dans le cadre de cet exemple que lorsqu'une action est effectuée, tous les agents sont informés par l'ajout de la croyance $done(\mathsf{a}, a, t)$ signifiant que l'agent $\mathsf{a}$ a effectué l'action $a$ à un instant $t$.

## IV.2 Typologie des jugements

Juger un autre agent, c'est-à-dire évaluer le caractère éthique de son comportement (voir section IV.2.2) dans un contexte au regard d'une théorie du bien et d'une théorie du juste peut se faire de diverses manières, selon son *caractère informé*, c'est-à-dire les connaissances du bien et du juste employées par ce jugement et le *caractère temporel*, c'est-à-dire instantané ou incrémental de ce jugement. Cette section explore l'espace défini par ces deux dimensions et explicite les différences de sens entre chaque type de jugement.

### IV.2.1 Niveau d'information du jugement

Le jugement présenté au chapitre III emploie des connaissances contextuelles $CK$, connaissances du bien $GK$ et connaissances du juste $RK$ pour juger les actions. Nous définissons ici plusieurs formes de jugements substituant les connaissances de l'agent jugé à celles de l'agent juge.

De manière naïve, un agent $\mathsf{a}$ peut juger l'action $a_{\mathsf{a}'}$ effectuée par l'agent $\mathsf{a}'$ dans une situation $s$ du monde en vérifiant si, dans cette situation $s$, l'action $a_{\mathsf{a}'}$ est bien une action juste, c'est-à-dire que $a_{\mathsf{a}'} \in \mathcal{A}_{r,\mathsf{a}}$ dans la situation $s$. Notons tout d'abord que pour cela, l'agent $\mathsf{a}$ doit connaitre $a_{\mathsf{a}'}$ (c'est-à-dire connaître les préconditions et postconditions de $a_{\mathsf{a}'}$). Nous supposons par la suite que lors du jugement de l'action $a_{\mathsf{a}'}$, si celle-ci n'est pas connue de l'agent juge (c'est-à-dire $a_{\mathsf{a}'} \notin A_{\mathsf{a}}$), l'agent juge $\mathsf{a}$ emploie dans son jugement l'ensemble $\{a_{\mathsf{a}'}\} \cup A_{\mathsf{a}}$ en considérant $a_{\mathsf{a}'}$ comme possible dans l'état $s$ (ce qui est manifestement vrai pour l'agent $\mathsf{a}'$).

Il est utile de rappeler ici que si l'action est jugée juste dans la situation $s$ (c'est-à-dire $a_{\mathsf{a}'} \in \mathcal{A}_{r,\mathrm{a}}$), cela ne fait pas de $a_{\mathsf{a}'}$ une action éthique de manière absolue, mais seulement relative à l'ensemble des croyances, connaissances du contexte, connaissances du bien, et connaissances du juste employé lors du jugement ayant produit $\mathcal{A}_{r,\mathrm{a}}$. Un jugement employant ainsi le modèle défini au chapitre précédent pour évaluer la conformité d'une action au regard des connaissances de l'agent juge est appelé *jugement éthique aveugle* et est défini de la manière suivante :

**Définition 32** (Jugement éthique aveugle)**.** *Un* jugement éthique aveugle *est un jugement dans lequel l'agent juge* $\mathsf{a}$ *emploie uniquement ses propres connaissances contextuelles* $CK_{\mathsf{a}}$*, sa propre connaissance du bien* $GK_{\mathsf{a}}$ *et sa propre connaissance du juste* $RK_{\mathsf{a}}$ *pour juger une action* $a_{\mathsf{a}'}$ *d'un agent jugé* $\mathsf{a}'$*.*

Un tel jugement donne une information sur la compatibilité d'une action d'un agent observé avec la morale et l'éthique de l'agent juge.

**Exemple 24.** *Dans notre exemple, nous considérons l'agent pauvre* Robin Hood*, en présence d'un agent riche* Prince John *et d'autres agents pauvres* Friar Tuck *et* Little John*. En observant* Friar Tuck *effectuer l'action wait*(), Robin Hood *peut utiliser son modèle de jugement dans la situation courante afin de produire l'ensemble des actions justes*

$\mathcal{A}_{r,\text{Robin Hood}} = \{steal(\text{Prince John})\}$ *et constater que* $wait() \notin \mathcal{A}_{r,\text{Robin Hood}}$*, signifiant que cette action exécutée par* Friar Tuck *n'est pas juste pour* Robin Hood *dans cette situation.*

Ce jugement étant relatif aux ensembles de connaissance $CK$, $GK$ et $RK$ employés, tout usage d'informations autres que celles représentant les croyances et paramètres éthiques de l'agent juge dans son jugement peut en changer l'issue. Ainsi, en obtenant une partie des croyances ou paramètres de l'agent jugé, l'agent peut évaluer la conformité de $a_{\mathsf{a'}}$ à tout ou partie des éléments supposés du jugement de a'. Employer la connaissance sur le contexte, le bien ou le juste d'un autre agent permet de produire un jugement plus informé, c'est-à-dire fondé sur des informations qui sont supposées avoir guidé le jugement de l'autre, de manière analogue à la théorie de l'esprit (voir section I.2.3). Ce type de jugement est appelé *jugement partiellement informé*, il est défini de la manière suivante :

**Définition 33** (Jugement éthique partiellement informé)**.** *Un* jugement éthique partiellement informé *est un jugement éthique dans lequel l'agent juge* a *emploie en partie les connaissances contextuelles* $CK_{\mathsf{a'}}$*, connaissances du bien* $GK_{\mathsf{a'}}$ *et connaissance du juste* $RK_{\mathsf{a'}}$ *d'un autre agent* a'*.*

Un tel jugement peut être produit par l'emploi de connaissances sur la situation ou de règles morales d'un autre agent, soit $s_{t,\mathsf{a'}}$ ou $mr_{\mathsf{a'}}$, pour évaluer la moralité d'une action. Ainsi, le prédicat $goodness(a, s_{t,\mathsf{a'}}, mr_{\mathsf{a'}}, mv)$ prenant en paramètres la connaissance $s_{t,\mathsf{a'}}$ d'un autre agent a' sur la situation et l'une de ses règles morales $mr_{\mathsf{a'}}$, permet d'affecter une valuation morale $mv$ à l'action $a$ en employant des connaissances qui ne sont pas propres à l'agent effectuant le jugement.

De même, un tel jugement peut également être produit en employant des principes éthiques d'autres agents. Le prédicat $rightness(a, s_{\mathsf{a'},t}, p_{\mathsf{a'}})$ prenant en paramètres la connaissance $s_{t,\mathsf{a'}}$ d'un autre agent a' sur la situation et l'un de ses principes éthiques $p_{\mathsf{a'}}$ permet d'évaluer, dans un contexte qui n'est pas celui perçu par l'agent juge mais donné par un autre agent, la conformité d'une action au regard d'un principe éthique également donné par un autre agent.

**Exemple 25.** *Reprenons l'exemple précédent. Si l'agent* Friar Tuck *annonce avoir pour seules règles morales "Voler est immoral" et "Donner est très moral",* Robin Hood *peut juger* Friar Tuck *en employant* $CK_{\text{Robin Hood}}$*,* $GK_{\text{Friar Tuck}}$ *et* $RK_{\text{Robin Hood}}$ *pour savoir si ce seul changement permet de considérer l'action de* Friar Tuck *comme juste.*

Enfin, un agent peut, s'il dispose de toutes les informations nécessaires, employer les connaissances d'un autre à la place des siennes pour juger une action. Cette substitution de tous les éléments du jugement est censée reproduire à l'identique le processus de l'autre, produire les mêmes états mentaux intermédiaires et le même ensemble final des actions justes ($\mathcal{A}_{r_{\mathsf{a'}}}$). Un tel jugement est appelé *jugement éthique pleinement informé* et défini de la manière suivante :

**Définition 34** (Jugement éthique pleinement informé)**.** *Un* jugement éthique pleinement informé *est un jugement dans lequel l'agent juge emploie uniquement d'autres connaissances que les siennes pour juger une action.*

À l'inverse du jugement aveugle qui permet à un agent d'évaluer ses actions ou celles des autres au regard de sa propre éthique, un jugement partiellement ou pleinement informé permet de l'évaluer par rapport à une éthique autre que celle du juge. Cette éthique employée pour le jugement peut avoir été transmise par un autre agent ou lui être attribuée. Notons que juger une action conforme à une éthique ne signifie pas que l'agent auteur de cette action ait effectivement employé cette éthique dans son processus de décision. Par exemple, un agent n'ayant aucune éthique mais n'ayant pas d'autres actions possibles que celles considérées éthiques par l'éthique employée pour le juger verra nécessairement son action évaluée comme conforme à cette éthique.

**Exemple 26.** Friar Tuck *peut transmettre à* Robin Hood *la totalité de ses connaissances (soit* $CK_{\text{Friar Tuck}}$*,* $GK_{\text{Friar Tuck}}$ *et* $RK_{\text{Friar Tuck}}$*) pour permettre à* Robin Hood *de les employer dans son jugement et ainsi reproduire le jugement qui a guidé la décision de* Friar Tuck*. Il est également capable de juger sa propre action steal*(Prince John) *en employant les connaissances de* Friar Tuck*.*

## IV.2.2 Temporalité du jugement

Le jugement tel que présenté dans la section précédente permet d'effectuer un *jugement ponctuel* d'une action dans une situation à un instant donné. Ce chapitre traite de l'information obtenue lors de jugements d'actions successives constituant le *comportement* d'un même agent en ajoutant une dimension temporelle.

**Définition 35** (Comportement)**.** *Le comportement* $b_{\mathsf{a}',[t_0,t]}$ *(pour behavior) d'un agent* a' *sur l'intervalle temporel* $[t_0, t]$ *est l'ensemble des actions* $a_{\mathsf{a}'}$ *exécutées par l'agent* a' *entre* $t_0$ *et* $t$ *tel que* $0 \leqslant t_0 \leqslant t$*.*

$$b_{\mathsf{a}',[t_0,t]} = \{a_{\mathsf{a}'} \in A : \exists t' \in [t_0, t] \ s.t. \ done(\mathsf{a}', a_k, t')\}$$

Par agrégation de jugements ponctuels successifs, l'agent peut construire de manière incrémentale et cumulative une image de la conformité du comportement de l'agent jugé vis-à-vis d'un ensemble de connaissances (CK,GK,RK) employé lors des jugements. Une image peut être calculée avec divers types de jugements agrégés de différentes manières. La période temporelle sur laquelle le comportement de l'agent est jugé pour construire cette image est l'un des paramètres du jugement.

Cependant, le niveau d'information pouvant varier d'un jugement à l'autre, le jugement de comportement dépend des jugements d'actions agrégés.

**Exemple 27.** *En observant les actions de l'agent* Friar Tuck*, l'agent juge* Robin Hood *peut les juger successivement et construire une image de l'éthique du comportement de*

Friar Tuck. *Cette image est une croyance permettant de qualifier le comportement observé de* Friar Tuck *vis-à-vis des connaissances employées pour le juger. Cette image évolue au cours de l'observation et permet à* Robin Hood *de la considérer comme un élément de la situation.*

## IV.3 Images de la moralité et de l'éthique d'un agent

La section précédente a présenté la notion de niveau d'information du jugement et de temporalité. Cette section décrit le mécanisme d'agrégation d'informations sur la moralité des actions, puis de jugement éthique progressif de comportements. L'objectif est de fournir à l'agent juge une information sur la conformité du comportement de l'autre vis-à-vis de tout ou partie de ses connaissances sur le contexte, la morale et l'éthique, en agrégeant les jugements ponctuels pour constituer un jugement progressif.

Un agent peut disposer de plusieurs images d'un même comportement construites par jugements progressifs portant sur divers éléments ou ensembles d'éléments du modèle de jugement, de manière aveugle, partiellement ou totalement informés.

### IV.3.1 Image de la moralité des actions d'un agent

Afin de construire une image de la moralité d'un autre agent, l'agent juge utilise le modèle d'évaluation morale (voir III.5) pour évaluer la conformité d'un comportement observé à un ensemble de règles morales *ms* (pour *moral set*) et classer ainsi chaque action $a_k$ d'un comportement d'agent $b_{\mathsf{a}',[t_0,t]}$ au regard de sa conformité à un ensemble de connaissances. La définition d'un tel ensemble de règles morales permet au concepteur de définir une sous-partie de la théorie du bien au regard de laquelle il est pertinent d'évaluer la conformité du comportement d'un autre. La conformité morale est définie ainsi :

**Définition 36** (Conformité morale d'une action)**.** *Une action $a_k$ est dite* moralement conforme *au regard des connaissances du contexte ($CK_{\mathsf{a}}$) et d'une règle des connaissances du bien ($GK_{\mathsf{a}}$) d'un agent* a *à un instant $t'$ si la valuation morale associée par la fonction d'évaluation morale au regard d'une règle morale $mr \in MR$ est supérieure à un seuil moral $mt \in MV$. Cette conformité est notée :*

$$moral_conformity(a_k, mr, mt, t')$$

*si et seulement si $a_k$ se trouve dans $\mathcal{A}_m$ affecté, en raison de la règle $mr$, d'une valuation morale supérieure ou égale à $mt$ au sens de l'ordre sur $MV$, à l'instant $t'$.*

*Soit :*

$$moral_conformity(a_k, mr, mt, t') \equiv goodness(a_k, s_{a,t'}, mr, mv) \wedge mv \geqslant mt$$

Remarquons que le prédicat $moral_conformity$ ne permet que d'évaluer une seule action au regard d'une seule règle. Pourtant, afin de permettre à un agent de se construire une image de la moralité du comportement d'un autre, il est nécessaire d'évaluer la

conformité de ce comportement en évaluant la conformité morale des actions successives qui le composent.

Cependant, définir la conformité morale d'un comportement à un ensemble de règles comme la conjonction de la conformité morale de toute action de ce comportement au regard de toute règle morale de cet ensemble serait problématique puisqu'une seule action mauvaise suffirait à condamner un comportement, peu importerait le nombre de bonnes actions observées. En effet, comme des contradictions peuvent être présentes dans la morale (voir I.1.1), toute action effectuée dans le cadre d'un dilemme moral (voir I.1.1.3) se voyant simultanément affectée de valuations supérieures et inférieures au seuil $mt$ rendrait le comportement de l'agent moralement non-conforme à l'ensemble de règles.

Afin de permettre à l'agent de raisonner sur la proportion d'actions évaluées, la conformité morale est utilisée pour calculer l'ensemble $MC^+$ des actions moralement conformes au regard de $ms$ et l'ensemble $MC^-$ des actions moralement non conformes au regard de $ms$ du comportement observé $b_{\mathsf{a'},[t_0,t]}$ de l'agent jugé $\mathsf{a'}$ :

$$\begin{aligned} MC^+_{b_{\mathsf{a'},[t_0,t]},ms,mt} = \{ & a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathsf{a'}, a_k, t') \\ & \wedge moral_conformity(a_k, mr, mt, t') \wedge mr \in ms\} \\ MC^-_{b_{\mathsf{a'},[t_0,t]},ms,mt} = \{ & a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0,t] \text{ s.t. } done(\mathsf{a'}, a_k, t') \\ & \wedge \neg moral_conformity(a_k, mr, mt, t') \wedge mr \in ms\} \end{aligned}$$

Dans le cadre de ce modèle, $MC_{b_{\mathsf{a'},[t_0,t]},ms,mt}$ représente l'ensemble des actions moralement évaluées de l'agent $\mathsf{a'}$ au regard de $ms$ et du seuil $mt$ entre l'instant $t_0$ et $t$. $MC^+_{b_{\mathsf{a'},[t_0,t]},ms,mt}$ est l'ensemble des actions moralement évaluées de $MC_{b_{\mathsf{a'},[t_0,t]},ms,mt}$ dont la valuation est supérieure ou égale à $mt$, tandis que $MC^-_{b_{\mathsf{a'},[t_0,t]},ms,mt}$ est l'ensemble des actions moralement évaluées de $MC_{b_{\mathsf{a'},[t_0,t]},ms,mt}$ dont la valuation est inférieure à $mt$. Ainsi, $MC_{b_{\mathsf{a'},[t_0,t]},ms,mt} = MC^+_{b_{\mathsf{a'},[t_0,t]},ms,mt} \cup MC^-_{b_{\mathsf{a'},[t_0,t]},ms,mt}$.

Pour agréger les évaluations morales ponctuelles de chaque action du comportement, l'agent juge a besoin d'une fonction $moralAggregation$ appliquée aux actions moralement évaluées $MC_{\mathsf{a'},ms,mt,[t_0,t]}$. Lors de l'agrégation de ces évaluations, la fonction $weight()$ prend en paramètre une action et donne un nombre réel permettant d'affecter une pondération à certaines actions dans la construction de l'image. Ainsi l'agent peut, par exemple, accorder plus d'importance aux actions jugées de manière pleinement informée, ou plus récentes. Dans le cas contraire, cette fonction attribue le même nombre pour toute action.

**Définition 37** (Fonction d'agrégation morale)**.** *Une* fonction d'agrégation morale *attribue une valeur quantitative représentant le ratio pondéré des actions d'un comportement évaluées moralement conformes par rapport à l'ensemble des actions de ce comportement.*

$moralAggregation : 2^{\mathcal{A}} \rightarrow [0,1]$ *tel que*

$$moralAggregation(MC_{b_{\mathsf{a}',[t_0,t]},ms,mt}) = \frac{\sum_{a_k \in MC^{+}_{b_{\mathsf{a}',[t_0,t]},ms,mt}} weight(a_k)}{\sum_{a_k \in MC_{b_{\mathsf{a}',[t_0,t]},ms,mt}} weight(a_k)}$$

L'agrégation des évaluations de conformité morale permet de construire un ensemble de croyances qualifiant la conformité du comportement d'un agent à un ensemble de règles morales. Le produit de cette agrégation est une image du caractère respectueux du comportement de l'agent observé vis-à-vis d'un ensemble de règles. Cette image est définie de la manière suivante :

**Définition 38** (Image morale)**.** *Une* image morale *d'un agent* a' *est une croyance construite par agrégation d'évaluations morales du comportement* $b_{\mathsf{a}',[t_0,t]}$ *de cet agent au regard d'un ensemble de règles morales* $ms$*, d'une connaissance du contexte* $CK$ *et d'une connaissance du bien* $GK$*. Cette image associe à ce comportement une valuation de conformité* $cv \in CV$*, où* $CV$ *est un ensemble ordonné de* valuations de conformité *défini dans l'ontologie. L'image morale qu'un agent* a' *se construit par évaluation de la conformité morale du comportement d'un agent* a' *au regard de* $ms$ *et* $mt$ *entre* $t_0$ *et* $t$ *est notée* $moral_image(\mathsf{a}', \mathsf{a}, ms, mt, cv, t_0, t)$

Cette image qualifiant la conformité du comportement au regard d'un ensemble de règles permettra dans les sections suivantes de tenir compte de cette information dans les interactions entre les agents.

**Exemple 28.** Robin Hood *dispose de trois ensembles moraux : le premier,* $ms_1 = \{mr_2, mr_4\}$*, contient les deux règles permettant d'évaluer la moralité des actions en fonction de la promotion ou trahison de la valeur de générosité (voir l'exemple 18 page 68). Du point de vue sémantique, un comportement dont les actions se verraient en majorité évaluées positivement par des règles de* $ms_1$ *serait donc un comportement que l'on peut qualifier de généreux (au sens de « promouvant la générosité et s'abstenant de l'enfreindre » ). Deux autres ensembles moraux contiennent les règles morales restantes :* $ms_2 = \{mr_1\}$ *et* $ms_3 = \{mr_3\}$*. Nous dotons également* Robin Hood *d'une fonction de discrétisation des agrégations morales avec* $CV = \{improper, neutral, congruent\}$ *et*

$$\begin{aligned} moral_image(\mathsf{a}', \mathsf{a}, ms, mt, cv, t_0, t) \equiv & moralAggregation(MC_{b_{\mathsf{a}',[t_0,t]},ms,mt}) \in [0, 0.4[ \\ & \rightarrow cv = improper \\ & \wedge moralAggregation(MC_{b_{\mathsf{a}',[t_0,t]},ms,mt}) \in [0.4, 0.6[ \\ & \rightarrow cv = neutral \\ & \wedge moralAggregation(MC_{b_{\mathsf{a}',[t_0,t]},ms,mt}) \in [0.6, 1] \\ & \rightarrow cv = congruent \end{aligned}$$

*Le seuil moral* $mt \in MV$ *est fixé à neutral et, par simplicité, la fonction* $weight()$ *attribue un poids identique de 1 pour toute les actions.*

Robin Hood *observe le comportement de* Prince John *qui taxe un agent* a' *ni pauvre ni riche à* $t-1$ *puis donne à l'agent pauvre* Sheriff of Nottingham *à* $t$. Robin Hood, *qui n'avait à* $t-2$ *pas d'image du comportement de* Prince John, *se construit puis met à jour à chaque action une image de la moralité du comportement de* Prince John *par rapport à l'ensemble moral* $ms_1$.

- *l'action* $tax(\mathsf{a'})$ *est selon* $vs_2$ *une action qui trahit la valeur de générosité, donc une action immorale selon* $mr_4$. *Aucun autre support de valeur n'étant actif dans ce contexte, la règle* $mr_2$ *ne permet pas d'évaluer l'action.*
  *Ainsi,* $\neg moral_conformity(tax(\mathsf{a'}), mr_4, neutral, t-1) \wedge mr_4 \in ms_1$ *est vérifié, ce qui permet d'ajouter l'action* $tax(\mathsf{a'})$ *à l'ensemble* $MC^-_{b_{Prince\ John,[t_0,t-1]},ms_1,neutral}$ *des actions immorales de ce comportement au regard de l'ensemble moral* $ms_1$ *et du seuil moral* $neutral$.
  *Réévaluant son image de* Prince John, *l'agent* Robin Hood *calcule l'agrégation morale. Ici,*

$$\begin{aligned} MC_{b_{Prince\ John,[t_0,t-1]},ms_1,neutral} &= MC^-_{b_{Prince\ John,[t_0,t-1]},ms_1,neutral} \\ &= \{tax(\mathsf{a'})\} \end{aligned}$$

  *donc* $moralAggregation(MC_{b_{Prince\ John,[t_0,t-1]},ms_1,neutral}) = 0$.
  *L'image morale produite est donc :*

$$moral_image(\text{Prince John}, \text{Robin Hood}, ms_1, neutral, improper, t_0, t-1)$$

  *indiquant que le comportement observé n'est pas conforme à l'ensemble moral* $ms_1$.
- *l'action* $give(\text{Sheriff of Nottingham})$ *est selon* $vs_1$ *une action qui promeut la valeur de générosité, donc une action très morale selon* $mr_2$. *Aucun autre support de valeur n'étant actif dans ce contexte, la règle* $mr_4$ *ne permet pas d'évaluer l'action.*
  *Ainsi,* $moral_conformity(give(\text{Sheriff of Nottingham}), mr_2, neutral, t) \wedge mr_2 \in ms_1$ *est vérifié, ce qui permet d'ajouter l'action évaluée* $give(\text{Sheriff of Nottingham})$ *à l'ensemble* $MC^+_{b_{Prince\ John^t,[t_0,t]},ms_1,neutral}$ *des actions morales de ce comportement au regard de l'ensemble moral* $ms_1$ *et du seuil moral* $neutral$.
  *Réévaluant son image de* Prince John, *l'agent* Robin Hood *calcule l'agrégation morale.*
  *Ainsi à l'instant* $t$ *on obtient,*

$$\begin{aligned} MC^+_{b_{Prince\ John,[t_0,t]},ms_1,neutral} &= \{give(\text{Sheriff of Nottingham})\} \\ MC^-_{b_{Prince\ John,[t_0,t]},ms_1,neutral} &= \{tax(\mathsf{a'})\} \\ MC_{b_{Prince\ John,[t_0,t]},ms_1,neutral} &= \{give(\text{Prince John}), tax(\mathsf{a'})\} \end{aligned}$$

*donc* $moralAggregation(MC_{b_{Prince\ John,[t_0,t]},ms_1,neutral}) = 0.5$.
*L'image morale produite est donc :*

$$moral_image(\text{Prince John}, \text{Robin Hood}, ms_1, neutral, neutral, t_0, t)$$

*indiquant que le comportement observé est à présent neutre par rapport à l'ensemble* $ms_1$.

*Cet exemple a montré comment les évaluations successives de ces deux actions a conduit* Robin Hood *à considérer le comportement de* Prince John *comme non-conforme puis comme neutre du point de vue de l'ensemble moral* $ms_1$ *regroupant les règles morales faisant appel à la valeur de générosité pour évaluer la moralité des actions. Un agent peut maintenir simultanément plusieurs calculs d'images avec des ensembles moraux, des fonctions de calcul de l'image morale, des seuils moraux ou des périodes de temps différents. Cela permet à* Robin Hood *de disposer à l'instant t d'une image de la moralité du comportement de* Prince John *avec cet ensemble de critères d'évaluation.*

### IV.3.2 Image de l'éthique des actions d'un agent

Le jugement d'actions d'un comportement permet d'évaluer leur conformité éthique et classer ainsi chaque action $a_k$ d'un comportement d'agent $b_{\mathsf{a'},[t_0,t]}$ au regard du résultat de son jugement en employant un ensemble de connaissances. La conformité éthique est définie ainsi :

**Définition 39** (Conformité éthique)**.** *Une action* $a_k$ *est dite* éthiquement conforme *au regard des connaissances du contexte* ($CK_{\mathsf{a}}$)*, connaissances du bien* ($GK_{\mathsf{a}}$) *et connaissances du juste* ($RK_{\mathsf{a}}$) *d'un agent* a *à un instant* $t'$ *si elle est jugée éthique. Cette conformité est notée :*

$$ethical_conformity(a_k, t')$$

*si et seulement si* $a_k$ *est dans l'ensemble des actions justes, c'est-à-dire* $a_k \in \mathcal{A}_{r_{\mathsf{a}}}$ *construit par le jugement éthique* $J_{\mathsf{a}}$ *de l'agent juge* a*, au regard des connaissances* $[CK_{\mathsf{a}}, GK_{\mathsf{a}}, RK_{\mathsf{a}}]$ *à l'instant* $t'$.

*Soit :*

$$ethical_conformity(a_k, t') \equiv \exists p : rightness(a_k, s_{\mathsf{a},t'}, p)$$

De manière analogue à la construction de la conformité morale d'un comportement, la conformité éthique est utilisée pour calculer l'ensemble $EC^+$ des actions éthiquement conformes et l'ensemble $EC^-$ des actions éthiquement non conformes du comportement observé $b_{\mathsf{a},[t_0,t]}$ de l'agent jugé a' entre $t_0$ et $t$ :

$$EC^+_{b_{\mathsf{a'}},[t_0,t]} = \{a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0, t] \text{ s.t. } done(\mathsf{a'}, a_k, t') \wedge ethical_conformity(a_k, t')\}$$
$$EC^-_{b_{\mathsf{a'}},[t_0,t]} = \{a_k \in b_{\mathsf{a'},[t_0,t]} \wedge t' \in [t_0, t] \text{ s.t. } done(\mathsf{a'}, a_k, t') \wedge \neg ethical_conformity(a_k, t')\}$$

Dans le cadre de ce modèle, $EC_{b_{\mathsf{a'}},[t_0,t]}$ représente l'ensemble des actions jugées de l'agent a' entre l'instant $t_0$ et $t$. $EC^{+}_{\mathsf{a'},[t_0,t]}$ est l'ensemble des actions de $EC_{b_{\mathsf{a'}},[t_0,t]}$ jugées éthiques, tandis que $EC^{-}_{b_{\mathsf{a'}},[t_0,t]}$ est l'ensemble des actions de $EC_{b_{\mathsf{a'}},[t_0,t]}$ jugées non-éthiques. Ainsi, $EC_{b_{\mathsf{a'}},[t_0,t]} = EC^{+}_{b_{\mathsf{a'}},[t_0,t]} + EC^{-}_{b_{\mathsf{a'}},[t_0,t]}$.

Afin d'agréger les jugements ponctuels de chaque action du comportement, l'agent juge a besoin d'une fonction *ethicAggregation* appliquée aux actions éthiquement évaluées $EC_{b_{\mathsf{a'}},[t_0,t]}$. De manière analogue à la fonction d'agrégation morale, cette fonction est définie de la manière suivante :

**Définition 40** (Fonction d'agrégation éthique)**.** *Une* fonction d'agrégation éthique *attribue une valeur quantitative représentant le ratio pondéré des actions d'un comportement jugées éthiques par rapport à l'ensemble des actions de ce comportement.* $ethicAggregation$ : $2^{\mathcal{A}} \rightarrow [0,1]$ *tel que*

$$ethicAggregation(EC_{b_{\mathsf{a'},[t_0,t]}}) = \frac{\sum_{a_k \in EC^{+}_{b_{\mathsf{a'}},[t_0,t]}} weight(a_k)}{\sum_{a_k \in EC_{b_{\mathsf{a'}},[t_0,t]}} weight(a_k)}$$

L'agrégation des jugements éthiques permet de construire des croyances qualifiant la conformité du comportement d'un agent à un ensemble de connaissances $CK$, $GK$ et $RK$. Le produit de cette agrégation est une image de l'éthique du comportement de l'autre. Cette image est définie de la manière suivante :

**Définition 41** (Image éthique)**.** *Une* image éthique *d'un agent* a' *est un jugement agrégé du comportement* $b_{\mathsf{a'},[t_0,t]}$ *de cet agent au regard d'une éthique, d'une connaissance du contexte* $CK$*, d'une connaissance du bien* $GK$ *et d'une connaissance du juste* $RK$*. Cette image attribue une valuation de conformité* $cv \in CV$*, où* $CV$ *est un ensemble ordonné de* valuations de conformité*. L'image éthique qu'un agent* $x$ *se construit par observation du comportement d'un agent* a' *entre* $t_0$ *et* $t$ *est notée* $ethical_image(\mathsf{a'}, \mathsf{a}, cv, t_0, t)$

Cette image qualifiant le caractère éthique du comportement permettra dans les sections suivantes de tenir compte de cette information dans les interactions entre les agents.

**Exemple 29.** *Nous reprenons l'exemple précédent dans lequel l'agent* Robin Hood *observe l'agent* Prince John *taxer un agent ni pauvre ni riche puis donner à l'agent pauvre* Sheriff of Nottingham.

*Nous dotons* Robin Hood *d'une fonction de discrétisation des agrégations éthiques avec* $CV = \{improper, neutral, congruent\}$ *et*

$$
\begin{aligned}
ethical_image(\mathsf{a'}, \mathsf{a}, cv, t_0, t) \equiv & ethicAggregation(EC_{b_{\mathsf{a'},[t_0,t]}}) \in [0, 0.4[ \\
& \rightarrow cv = improper \\
& \wedge ethicAggregation(EC_{b_{\mathsf{a'},[t_0,t]}}) \in [0.4, 0.6[ \\
& \rightarrow cv = neutral \\
& \wedge ethicAggregation(EC_{b_{\mathsf{a'},[t_0,t]}}) \in [0.6, 1] \\
& \rightarrow cv = congruent
\end{aligned}
$$

*Par simplicité, la fonction weight() attribue un poids identique de 1 pour toute les actions.*

Robin Hood*, lui-même ni pauvre ni riche, observe le comportement de* Prince John *qui taxe un agent* a' *ni pauvre ni riche à* $t-1$ *puis donne à l'agent pauvre* Sheriff of Nottingham *à* $t$*.* Robin Hood*, qui n'avait à* $t-2$ *pas d'image du comportement de* Prince John*, se construit ou met à jour à chaque action une image de l'éthique du comportement de* Prince John*.*

- *l'action* $tax(\mathsf{a'})$ *est, comme nous l'avons montré à l'exemple précédent, une action qualifiée d'immorale selon* $mr_4$*. Le tableau IV.1 illustre le résultat du jugement de* Robin Hood *et montre pourquoi* $tax(\mathsf{a'})$ *est jugée non-éthique, c'est à dire que* $\neg ethical_conformity(tax(\mathsf{a'}), t-1)$ *est vrai, ce qui permet d'ajouter l'action* $tax(\mathsf{a'})$ *à l'ensemble* $EC^{-}_{b_{Prince\ John,[t_0,t-1]}}$ *des actions éthiquement non conformes de ce comportement.*

| Action | Évaluations | | | Évaluations éthiques | | | |
|---|---|---|---|---|---|---|---|
| | $dv$ | $pv$ | $mv$ | $p_1$ | $p_2$ | $p_3$ | $p_4$ |
| $a_1$ : $steal(\mathsf{a})$ | *neutral* | *possible* | *immoral* $\wedge$*moral* | $\bot$ | $\top$ | $\bot$ | $\bot$ |
| $a_2$ : $give(\mathsf{a})$ | *neutral* | *possible* | *very moral* ssi $poor(\mathsf{a})$ | $\bot$ | $\top$ | $\top$ | $\top$ |
| | | | *neutral* sinon | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| $a_3$ : $tax(\mathsf{a})$ | *neutral* | *impossible* | *immoral* | $\bot$ | $\bot$ | $\bot$ | $\bot$ |
| $a_4$ : $court(\mathsf{a})$ | *desirable* ssi a=Marian | *possible* | | $\bot$ | $\bot$ | $\top$ | $\top$ |

TABLE IV.1 – Tableau récapitulatif des évaluations lorsque Robin Hood n'est ni pauvre ni riche.

*Réévaluant son image de* Prince John*, l'agent* Robin Hood *calcule l'agrégation éthique. Ici,*

$$
\begin{aligned}
EC_{b_{Prince\ John,[t_0,t-1]}} &= EC^{-}_{b_{Prince\ John,[t_0,t-1]}} \\
&= \{tax(\mathsf{a'})\}
\end{aligned}
$$

*donc* $ethic_Aggregation(EC_{b_{Prince\ John,[t_0,t-1]}}) = 0$.
*L'image morale produite est donc :*

$$ethical_image(\text{Prince John}, \text{Robin Hood}, improper, t_0, t-1)$$

*indiquant que le comportement observé n'est en cet instant pas conforme à l'éthique de* Robin Hood.

- *l'action* $give(\text{Sheriff of Nottingham})$ *est, selon le même tableau illustrant le jugement de* Robin Hood, *jugée comme étant l'action éthique.*
  *Ainsi,* $ethcal_conformity(give(\text{Sheriff of Nottingham}), t)$ *est vérifié, ce qui permet d'ajouter l'action évaluée* $give(\text{Sheriff of Nottingham})$ *à l'ensemble* $EC^{+}_{b_{Prince\ John^t,[t_0,t]}}$ *des actions de ce comportement conformes à l'éthique utilisée.*
  *Réévaluant son image de* Prince John, *l'agent* Robin Hood *calcule l'agrégation éthique. Ainsi à l'instant* $t$ *on obtient,*

$$\begin{aligned}
EC^{+}_{b_{Prince\ John,[t_0,t]}} &= \{give(\text{Sheriff of Nottingham})\} \\
EC^{-}_{b_{Prince\ John,[t_0,t]}} &= \{tax(\text{a'})\} \\
EC_{b_{Prince\ John,[t_0,t]}} &= \{give(\text{Sheriff of Nottingham}), tax(\text{a'})\}
\end{aligned}$$

  *donc* $ethicAggregation(EC_{b_{Prince\ John,[t_0,t]}}) = 0.5$.
  *L'image éthique produite est donc :*

$$ethical_image(\text{Prince John}, \text{Robin Hood}, neutral, t_0, t)$$

  *indiquant que le comportement observé est à présent neutre par rapport à l'éthique employée pour le jugement.*

*Cet exemple a montré comment les évaluations successives de ces deux actions a conduit* Robin Hood *à considérer le comportement de* Prince John *comme non-conforme puis comme neutre du point de vue des connaissances de* Robin Hood. *Un agent peut maintenir simultanément plusieurs calculs d'images éthiques avec des fonctions de calcul de l'image éthique, des seuils ou des périodes de temps différents. Cela permet à* Robin Hood *de disposer à l'instant* $t$ *d'une image de l'éthique du comportement de* Prince John *avec cet ensemble de critères d'évaluation.*

Remarquons qu'il est possible de disposer d'images d'un agent telles qu'il serait jugé conforme par toutes les images morales en n'étant non conforme pour son image éthique. Cela peut se produire par exemple pour des raisons de divergences dans la résolution de dilemmes moraux pour lesquels toute alternative améliore au moins une image morale mais où les divergences de théories du juste de l'agent juge et de l'agent jugé les amènent à ne pas considérer la même action comme juste.

À l'inverse, dans des situations dans lesquelles aucune action morale n'est possible, le comportement d'un agent peut amener un agent juge à dégrader ses images morales de l'agent jugé en améliorant l'image éthique et aboutir à une situation dans laquelle aucune image morale du comportement n'est conforme, mais où l'image éthique le serait.

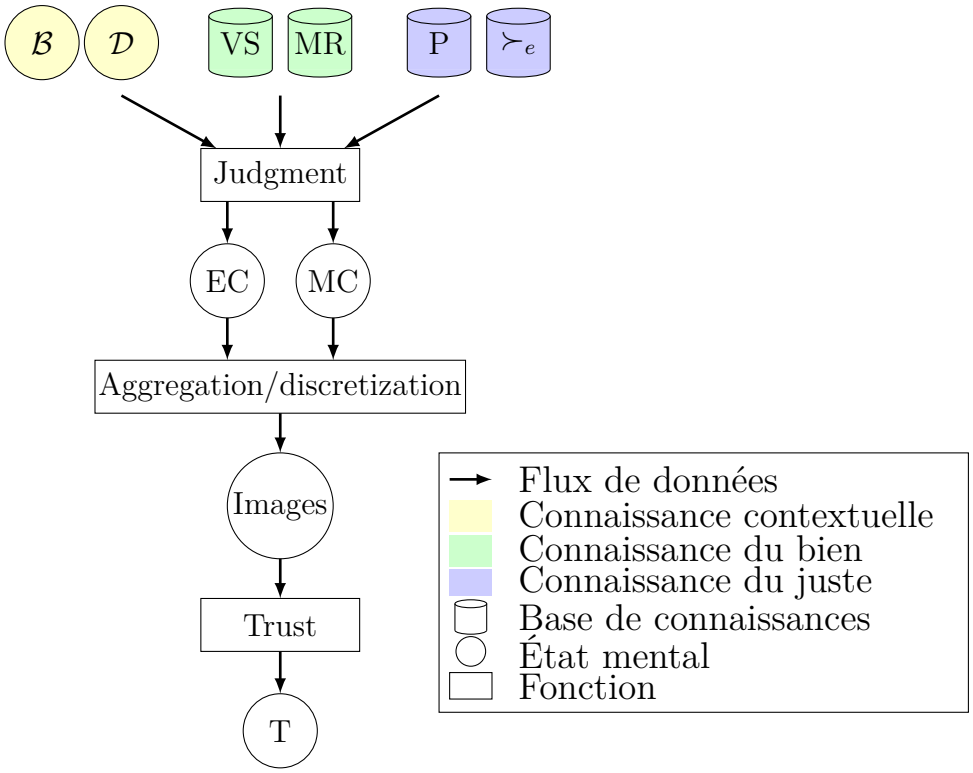


FIGURE IV.1 – Modèle de coopération entre agents autonomes fondé sur l'éthique

## IV.4 Coopération fondée sur le jugement des autres

La construction d'images du comportement des autres agents vis-à-vis d'éléments de la morale et de l'éthique permet à l'agent juge d'accorder ou non sa confiance à un autre agent. La confiance construite par ce processus peut ensuite être employée pour décrire une manière éthique d'interagir et coopérer avec les autres agents du système.

La figure IV.1 représente le mécanisme de construction de la confiance dans sa globalité : le modèle de jugement décrit au chapitre précédent est employé avec un ensemble de connaissances du contexte, de connaissances du bien et de connaissances du juste afin de générer les ensembles d'actions évaluées par l'éthique $EC$ et par la morale $MC$. Ces ensembles permettent à leur tour la construction d'un ensemble d'images du comportement de l'agent jugé. Nous décrivons à présent l'emploi d'une action permettant à l'agent de construire l'ensemble $T$ des croyances en la confiance qu'il accorde aux autres agents.

### IV.4.1 Construction de la confiance

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

d'évaluer la possibilité d'établir ces deux types de confiance.

**Définition 42** (Fonction de confiance morale)**.** *La* fonction de confiance morale $MTB_{\mathsf{a}}$ *(pour Moral Trust building) permettant d'évaluer si l'agent juge peut accorder sa confiance à l'agent* $\mathsf{a}$ *pour la conformité de son comportement vis-à-vis de l'ensemble moral* $ms$ *est définie comme :*

$$MTB_{\mathsf{a}} : Ag \times 2^{ms_{\mathsf{a}}} \times MV_{\mathsf{a}} \rightarrow \{\top, \bot\}$$

**Définition 43** (Fonction de confiance éthique)**.** *La* fonction de confiance éthique $ETB_{\mathsf{a}}$ *(pour Ethics Trust Building) permettant d'évaluer si l'agent juge peut accorder sa confiance à l'agent* $\mathsf{a}$ *pour la conformité de son comportement vis-à-vis du jugement éthique est définie comme :*

$$ETB_{\mathsf{a}} : Ag \rightarrow \{\top, \bot\}$$

Ici, ces fonctions de confiance sont abstraites et doivent être instanciées. Lorsqu'un agent $\mathsf{a}$ évalue la conformité du comportement d'un autre agent $\mathsf{a}$' au regard de $CK_{\mathsf{a}}$, $GK_{\mathsf{a}}$ et $RK_{\mathsf{a}}$ (c'est-à-dire l'image éthique), la fonction de confiance éthique produit une croyance $ethical_trust(\mathsf{a}', \mathsf{a})$. De même, lorsque l'agent $\mathsf{a}$ évalue la conformité du comportement de $\mathsf{a}$' au regard de $ms$ (c'est-à-dire vérifie que la conformité morale de l'image de son comportement par rapport à $ms$ est au moins égale à $mt$), la fonction de confiance morale produit une croyance $moral_trust(\mathsf{a}', \mathsf{a}, ms, mt)$. L'ensemble de ces croyances de l'agent représentant sa confiance dans les autres agents est noté $T$.

### IV.4.2 Éthique de la confiance

Faire confiance étant une action épistémique, il est possible de décrire la moralité de cette action en fonction du contexte et de juger s'il est juste d'accorder sa confiance à un autre agent.

La description de supports de valeurs pour l'action de « faire confiance » permet de définir de nouvelles valeurs décrivant la manière d'accorder sa confiance aux autres agents. Par exemple, l'intransigeance peut être une valeur supportée par l'action d'accorder sa confiance uniquement aux agents dont l'image est au dessus d'un seuil moral ou éthique relativement élevé. À l'inverse, l'indulgence définie comme supportée par l'action consistant à accorder sa confiance à des agents dès lors qu'il existe une image dépassant un seuil moral ou éthique relativement bas.

La description de règles morales peut ensuite décrire la moralité de la confiance. Par exemple, il est possible de définir une règle morale telle que « Il est moral d'être indulgent durant les cinq premières minutes de l'observation de leur comportement » ou bien « Il est immoral de ne pas être intransigeant lorsque la situation est critique ». La moralité de la construction de la confiance en fonction des paramètres de la construction de cette confiance peut ainsi être dépendante de la connaissance que l'agent a du contexte.

### IV.4.3 Utilisation de la confiance pour la coopération éthique

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

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

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

Enfin, puisque évaluer et juger les autres constituent des actions, il est également possible d'exprimer et évaluer leur caractère moral ou éthique. Ainsi, la valeur morale de *tolérance* peut être supportée par la construction d'une image des autres avec un seuil peu élevé tant que les ensembles $EC_{\mathsf{a}',[t_0,t]}$ ou $MC_{\mathsf{a}',[t_0,t]}$ ne sont pas assez significatifs. Le choix du seuil, des pondérations et la conversion de l'agrégation en niveau de conformité peuvent également permettre de représenter diverses formes de confiance. Une valeur telle que l'*indulgence* peut être supportée par le fait d'accorder toujours une pondération plus faible aux actions les moins récentes. Il est ainsi possible de décrire une morale de la confiance par l'emploi de règles comme "Il est immoral de construire la confiance sans tolérance ni indulgence" (Horsburgh, 1960).

**Exemple 30.** *Prenons cette fois l'exemple de l'agent* Prince John *doté d'un ensemble de supports de valeurs comprenant les supports $vs_1$, $vs_2$ et $vs_3$ de* Robin Hood *définissant la générosité auxquels viennent s'ajouter des supports de l'obéissance suivants :*

*$vs_5$ : « Voler un noble trahit la valeur d'obéissance », est un support identique à celui de* Sheriff of Nottingham *(voir l'exemple récapitulatif du chapitre précédent) spécifique à une action et restreint aux situations dans lesquelles la cible est noble*

$$vs_5 = \langle steal(\mathsf{a}), noble(\mathsf{a}), \top, obedience, defeat \rangle$$

*$vs_6$ : « Donner à un noble promeut la valeur d'obéissance », est un support qui, comme le précédent est restreint à une action dans la situation particulière qui est la noblesse de l'agent ciblé*

$$vs_6 = \langle give(\mathsf{a}), noble(\mathsf{a}), \top, obedience, promote \rangle$$

*L'ensemble des règles morales de* Prince John *est constitué de :*

$mr_1$ *: « Trahir la valeur générosité est immoral »*

$$mr_1 = \langle \{any\}, \top, \top, \{\langle generosity, defeat, max \rangle\}, immoral \rangle$$

$mr_2$ *: « Effectuer une action généreuse est moral »*

$$mr_2 = \langle \{any\}, \top, \top, \{\langle generosity, promote, min \rangle\}, moral \rangle$$

$mr_3$ *: « Trahir la valeur obéissance est très immoral »*

$$mr_3 = \langle \{any\}, \top, \top, \{\langle obedience, defeat, max \rangle\}, very\ immoral \rangle$$

$mr_4$ *: « Effectuer une action promouvant l'obéissance est très moral »*

$$mr_4 = \langle \{any\}, \top, \top, \{\langle obedience, promote, min \rangle\}, very\ moral \rangle$$

Prince John *est doté de deux ensembles moraux que sont* $ms_{generosity} = \{mr_1, mr_2\}$ *et* $ms_{obedience} = \{mr_3, mr_4\}$ *et d'un seul principe éthique qui est « Une action est juste si elle est possible, désirable et pas très immorale ».*

*Après avoir jugé de manière aveugle le comportement de* Sheriff of Nottingham *et* Robin Hood*,* Prince John *se construit une représentation de la morale et de l'éthique de leur comportement afin de déterminer avec lesquels il est envisageable de collaborer. En employant les fonctions présentées en section IV.3 nous supposons qu'il a produit des images suivantes :*

$moral_image(\text{Sheriff_of_Nottingham}, \text{Prince John}, ms_{obedience}, neutral, neutral, t_0, t)$
$moral_image(\text{Sheriff_of_Nottingham}, \text{Prince John}, ms_{generosity}, neutral, improper, t_0, t)$
$moral_image(\text{Robin Hood}, \text{Prince John}, ms_{obedience}, neutral, improper, t_0, t)$
$moral_image(\text{Robin Hood}, \text{Prince John}, ms_{generosity}, neutral, congruent, t_0, t)$
$ethical_image(\text{Sheriff_of_Nottingham}, \text{Prince John}, improper, t_0, t)$
$ethical_image(\text{Robin_Hood}, \text{Prince John}, improper, t_0, t)$

*Ces images signifient que, après observation de leur comportement,* Prince John *considère celui de* Sheriff of Nottingham *comme non-conforme par rapport à l'ensemble moral* $ms_{generosity}$ *et neutre par rapport à* $ms_{obedience}$ *tandis que celui de* Robin Hood *serait conforme par rapport à l'ensemble moral* $ms_{generosity}$ *et non-conforme à* $ms_{obedience}$*.*

*En ajoutant à la morale de* Prince John *une règle morale* $mr_5$ *telle que*

$mr_5$ *: « Il est très immoral d'accorder sa confiance à un agent dont un agent noble considère le comportement comme non conforme à l'obéissance »*

$$mr_5 = \langle\{build_ethical_trust(\mathsf{a})\}, moral_image(\mathsf{a}, \mathsf{a}', ms_{obedience}, neutral, improper, t_0, t) \\ \wedge noble(a')\top, \emptyset, very\ immoral\rangle$$

*Cette règle $mr_5$ est également ajoutée à $ms_{obedience}$. L'action $build_ethical_trust(\mathsf{a})$ permettant à l'agent de construire sa confiance en l'agent* a *n'est maintenant éthique selon le jugement de* Prince John *qu'à condition que l'agent concerné ait un comportement jugé comme obéissant. Cette règle morale a permis de définir la moralité et l'éthique de la confiance pour* Prince John *en fonction de sa représentation du comportement des autres.*

*Remarquons enfin que si* Prince John *transmet ses connaissances du bien et du juste $GK_{\text{Prince John}}$ et $RK_{\text{Prince John}}$ à* Sheriff of Nottingham*, celui-ci peut construire une image du comportement de* Robin hood *par une succession de jugements partiellement informés en employant ces connaissances et ainsi obtenir lui aussi l'image suivante :*

$$moral_image(\text{Robin Hood}, \text{Prince John}, ms_{obedience}, neutral, improper, t_0, t)$$

*Une fois cette image construite, il peut, grâce à un jugement partiellement informé employant $GK_{\text{Prince John}}$ et $RK_{\text{Prince John}}$, en déduire*

$$\neg moral_conformity(build_ethical_trust(\text{Robin Hood}), mr_5, neutral, t) \wedge mr_5 \in ms_{obedience})$$

$$\neg ethical_conformity(build_ethical_trust(\text{Robin Hood}), t)$$

*Ces connaissances peuvent alors être prises en compte par* Sheriff of Nottingham *qui peut déduire qu'accorder sa confiance à* Robin Hood *remettrait en question la confiance que lui accorde* Prince John*. Il peut également être doté de désirs et de règles morales lui permettant évaluer comme indésirable et immoral le fait de détériorer son image auprès d'un agent noble.*

## IV.5 Synthèse

Le modèle de jugement présenté au chapitre précédent permettait à un agent de raisonner sur des valeurs morales, règles morales et principes éthiques afin de juger de l'action juste à effectuer. L'usage de ce modèle comme un processus décisionnel ne permettait en revanche pas à cet agent d'évaluer le comportement des autres agents du système.

Nous avons montré dans ce chapitre comment l'agent peut utiliser les connaissances d'un autre dans son propre raisonnement afin de produire des jugements par rapport à différents ensembles de connaissances. Ces jugements permettent à un agent de remplacer tout ou partie de ses connaissances sur le contexte, la morale et l'éthique afin de produire un jugement lui permettant de reproduire partiellement ou totalement le jugement d'un autre. Nous appelons la proportion de cette substitution le « niveau d'information » du

jugement. Ensuite nous avons étendu le problème du jugement des actions à celui des comportements, composés de séquences d'actions.

Le jugement de comportement produit des images, c'est-à-dire des croyances associant à un comportement un niveau de conformité en fonction de la proportion d'actions conformes à des ensembles moraux ou au jugement éthique. La construction de ces images est elle aussi sujette à de nombreuses possibilités de paramétrage. Il est ainsi possible de faire preuve de plus ou moins d'exigences dans la proportion d'actions bonnes ou justes nécessaires pour obtenir une image positive. Il est également possible de pondérer les actions agrégées afin de permettre à certaines d'entre elles d'avoir plus d'impact sur l'image finale (par exemple en fonction de la distance entre le seuil de moralité et la valuation affectée par l'évaluation morale de l'action, ou bien en fonction du temps écoulé entre la réalisation de l'action et le moment où est construite l'image).

Les images permettent enfin de décrire le contexte d'une prise de décision de l'agent d'accorder ou non sa confiance à un autre agent afin de permettre des actions de coopération. La moralité de cette action de décision de coopération peut elle-même être décrite par des valeurs morales et règles morales, dépendantes ou non du contexte. C'est la définition de ces éléments qui, en permettant d'évaluer la moralité afin de juger l'action d'accorder sa confiance, permet de définir une éthique de la coopération.

Ce chapitre a ainsi montré comment le modèle de jugement peut être employé pour juger le comportement des autres et fournir un cadre permettant la coopération entre les agents en se fiant à des représentations de la morale et de l'éthique des autres. Le chapitre suivant s'attachera à illustrer l'implémentation de ce modèle de coopération et à évaluer la qualité des images construites.

# CHAPITRE V

# Mise en œuvre et expérimentations

**V.1 Implémentation du modèle de jugement . . . . . 100**
V.1.1 Le framework JaCaMo . . . . . 100
V.1.2 Reconnaissance de situation . . . . . 102
V.1.3 Évaluation de la désirabilité et de la possibilité des actions . 103
V.1.4 Moralité des actions . . . . . 104
V.1.5 Évaluation de l'éthique des actions. . . . . 106
V.1.6 Jugement des autres agents . . . . . 109
**V.2 Application à la gestion d'actifs financiers . . . . . 110**
V.2.1 Les agents autonomes sur les marchés financiers . . . . . 111
V.2.2 L'éthique dans la finance. . . . . 112
V.2.3 Description du simulateur . . . . . 113
**V.3 Évaluation de l'influence du jugement sur le comportement individuel . . . . . 116**
V.3.1 Description métier du domaine. . . . . 116
V.3.2 Paramétrage moral des agents . . . . . 117
V.3.3 Initialisation . . . . . 119
V.3.4 Comportement des agents . . . . . 120
**V.4 Évaluation du jugement des autres . . . . . 121**
V.4.1 Initialisation . . . . . 122
V.4.2 Agrégation d'images . . . . . 122
**V.5 Synthèse . . . . . 123**

Le chapitre III propose un modèle de jugement montrant comment représenter les concepts définis au chapitre I et les employer dans le raisonnement de l'agent. Le chapitre IV étend ensuite ce modèle pour proposer un cadre de coopération fondé sur l'éthique dans lequel les agents, observant le comportement des autres agents du système, jugent ces comportements au cours de leur observation et peuvent ensuite décider d'accorder leur confiance et entreprendre des actions de coopération.

Ce chapitre montre comment ce modèle peut être implémenté, paramétré, testé et évalué dans un contexte applicatif réaliste, qui est celui de la gestion éthique d'actifs financiers. La section V.1 explique comment est implémenté le modèle proposé au chapitre III. La section V.2 présente ensuite le domaine applicatif de la gestion éthique d'actifs financiers et montre comment certaines valeurs morales, règles morales, principes éthiques et préférences éthiques sont implémentés et fournis aux agents. La section V.3 montre les résultats expérimentaux démontrant les changements de comportement produits par le jugement éthique. La section V.4 montre ensuite comment ce modèle de jugement est employé pour évaluer le comportement des autres, construire une image de leur comportement par rapport à des sous-ensembles de leurs règles morales ou à leur éthique et établir une confiance pouvant servir de préalable à une coopération. Nous dressons en synthèse de ce chapitre un bilan de ces expérimentations soulignant les propriétés mises en évidence.

## V.1 Implémentation du modèle de jugement

Nous détaillons dans cette section l'implémentation du modèle présenté dans les deux chapitres précédents. L'implémentation de ce modèle de jugement fait abstraction du domaine applicatif employé pour les expérimentations et seules les connaissances du contexte et connaissances du bien emploient des croyances, désirs et actions spécifiques au domaine. Nous commençons par introduire le framework JaCaMo, conçu pour faciliter l'implémentation de systèmes multi-agents BDI en séparant la description des agents, de l'environnement et des organisations. Puis nous expliquons l'implémentation des éléments du modèle en suivant l'ordre employé dans les chapitres précédents.

### V.1.1 Le framework JaCaMo

JaCaMo est une plateforme produite par l'intégration de trois composants préexistants (Boissier *et al.*, 2016) :

— *Jason* est une implémentation étendant le langage logique AgentSpeak (A. S. Rao, 1996) inspiré du modèle BDI, proposant une architecture d'agent disposant d'une base de croyances, d'une base de plans, d'une base de buts, d'un ensemble de règles logiques et d'un mécanisme de sélection des intentions (Bordini *et al.*, 2007). Le concepteur du système peut adapter cette architecture en implémentant ses propres composants et actions internes afin de modifier le processus de sélection des intentions ou étendre l'ensemble des primitives à disposition de l'agent.

— *CArtAgO* (Ricci *et al.*, 2009) est un framework conçu pour faciliter le développement d'environnements qui peut être ensuite exploré par les agents lors de l'exécution. Il facilite notamment la conception d'artefacts en fournissant une classe et un ensemble de méthodes écrits en JAVA permettant d'implémenter rapidement des propriétés observables par les agents et d'actions leur permettant d'agir sur l'artefact.

— $\mathcal{M}oise$ (Hubner *et al.*, 2007) est un framework déclaratif pour la description d'organisations normatives, leur gestion à l'échelle du système et le raisonnement sur

ces organisations au niveau des agents.

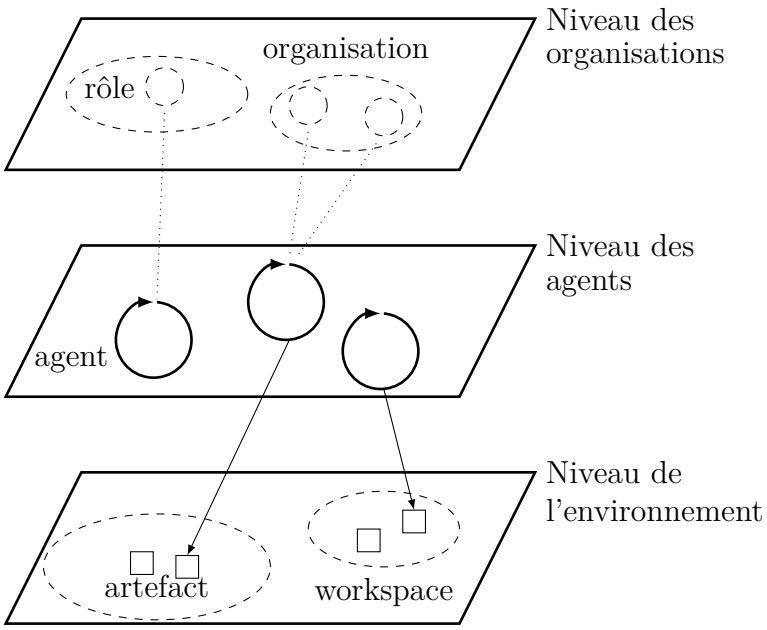


FIGURE V.1 – Représentation générale d'un système multi-agent conçu avec JaCaMo

La figure V.1 illustre l'implémentation d'un système multi-agent à l'aide de JaCaMo : les agents peuvent intégrer des organisations (traits pointillés) y adopter des rôles, et exécuter des opérations sur les artefacts de l'environnement (flèches). Pour décrire leurs connaissances et leur comportement, il est possible de décrire des croyances, règles, buts et plans en Jason.

L'exemple d'implémentation V.1 présente un exemple simple d'agent. Cet agent dispose d'une croyance initiale `beliefExample2("test")` et de deux règles unaires. La première règle est vraie s'il n'existe pas de croyance `beliefExample1(Arg)`, où `Arg` est une variable passée en argument (les variables commencent toujours par des majuscules en Jason) et le mot-clé `not` symbolise la négation faible (Jason emploie l'hypothèse d'un monde ouvert dans lequel la négation forte est notée `~`). Un plan est décrit par un évènement déclencheur (but à satisfaire, nouvelle croyance), suivi d'un contexte, c'est-à-dire une conjonction de croyances et de règles devant être vérifiée pour que le plan puisse être sélectionné comme une intention, puis d'un ensemble d'opérations. Les opérations permettent d'ajouter ou supprimer des croyances, ajouter des buts, effectuer des actions internes à l'agent ou externes (par exemple sur un artefact). Cet agent va alterner entre les deux plans connus puisque leurs contextes respectifs ne seront valides qu'après l'exécution de l'autre plan et que le but est réactivé à la fin de chaque plan.

```
1        /* Initial beliefs and rules */
2        beliefExample2("test").
3        ruleExample1(Arg):- not beliefExample1(Arg).
4        ruleExample2(Arg):- beliefExample2(Arg).
5
6        /* Initial goals */
```

```
7          !goalExample("test").
8
9          /* Plans */
10         +!goalExample(Arg) :
11             ruleExample1(Arg)          // Context of the plan
12           & ruleExample2(Arg)
13           <- +beliefExample1(Arg); // a belief addition
14              -beliefExample2(Arg); // a belief deletion
15              .wait(1000);          // an internal action
16              !goalExample(Arg).    // a goal addition
17
18         +!goalExample(Arg) : not ruleExample1(Arg)
19           & not ruleExample2(Arg)
20           <- -beliefExample1(Arg);
21              +beliefExample2(Arg);
22              .wait(1000);
23              !goalExample(Arg).
```

Implémentation V.1 – Exemple de description d'un agent en Jason

JaCaMo dispose d'une communauté d'utilisateurs relativement large et active, en témoigne la proportion d'articles disposant de preuves de concept implémentées à l'aide de ce framework dans des conférences spécialisées [1]. Libre et indépendant du système d'exploitation (son exécution nécessite cependant l'existence d'une machine virtuelle Java), la communauté s'attache à ajouter constamment des fonctionnalités, documenter son utilisation et étendre le spectre des utilisations possibles (par exemple construire un système décentralisé dans des objets connectés).

Dans le cadre de l'implémentation du modèle de jugement, nous allons principalement nous attacher à la description des agents à l'aide de Jason. Nous utilisons dans cette section l'exemple employé dans les deux chapitres précédents afin d'illustrer les concepts généraux. Le jugement prend ici la forme d'un ensemble de règles permettant d'évaluer la conformité d'une action. Cette implémentation emploie quelques simplifications : les supports de valeurs sont réduits à deux états (une action promeut ou trahit une valeur) et les connaissances de l'ontologie $\mathcal{O}$ sont ici représentées sous la forme de prédicats. Nous détaillons la façon dont sont implémentés les connaissances sur les actions, les désirs, les supports de valeurs morales, les règles morales, les principes éthiques et le préférences. Nous présentons également les mécanismes permettant de raisonner sur ces éléments.

### V.1.2 Reconnaissance de situation

Un agent s'exécutant sur la plateforme JaCaMo a accès à différentes sources d'information pour percevoir et reconnaître une situation décrite au final dans sa base de croyances. Une première source d'information consiste en la perception des propriétés observables des artefacts que l'agent observe dans l'environnement dans lequel il est

1. Voir par exemple le *Fourth International Workshop on Engineering Multi-Agent Systems* (EMAS) organisé à AAMAS en 2016 http://www.di.unito.it/~argo/papers/EMAS2016-WorkshopNotes.pdf

situé. Ces propriétés observables offrent un aperçu de l'état de l'artefact (par exemple la disponibilité d'une ressource ou la valeur d'une variable interne). Une autre source d'information est celle issue des signaux que les artefacts peuvent générer à chaque fois que leur état change ou que des opérations exécutées sur l'artefact en génèrent (par exemple si ceux-ci sont mis à jour régulièrement par un processus interne ou bien s'ils encapsulent des éléments de communication via un réseau ou une interface graphique). Les autres sources d'information sont les envois de messages entre agents et les croyances initiales ou générées en interne par raisonnement au sein de l'agent. Nous voyons ainsi que selon la manière dont l'application est conçue, la reconnaissance d'une situation par un agent peut aboutir aux mêmes problèmes que la reconnaissance de situation dans des cas réels : perception locale, incomplète et incertaine. Dans le cadre de notre application, nous avons utilisé les mécanismes décrits ci-dessus pour simplifier et éviter ce problème essentiel.

### V.1.3 Évaluation de la désirabilité et de la possibilité des actions

Nous représentons les connaissances sur les conséquences des actions à l'aide d'un prédicat indiquant en fonction du contexte les croyances qui deviennent vraies après l'exécution de l'action comme l'illustre l'exemple d'implémentation V.2.

```
1        consequence(A1,tax(A2),rich(A1)):-
2          not rich(A1) & not poor(A1).
3
4        consequence(A1,tax(A2),poor(A2)):-
5          not rich(A2) & not poor(A2).
```

Implémentation V.2 – Exemples de connaissances sur les actions

Nous distinguons dans le cadre de notre expérimentation les buts (*goals*) au sens du moteur BDI proposé par JaCaMo et le caractère désirable déterminé par le prédicat suivant `desire_eval` illustré par l'exemple d'implémentation V.3.

```
1        desire_eval("Robin_Hood",court("Marian"),desirable).
2        desire_eval("Marian",court("Friar_Tuck"),undesirable).
3        desire_eval("Prince John",A,desirable):-
4          consequence("Prince John",A,rich("Prince John")).
```

Implémentation V.3 – Exemples de désirs

Ce prédicat permet de connaître, dans le contexte connu de l'agent (c'est à dire l'état de sa base de croyances et les connaissances des conséquences de l'action), la valuation de désirabilité associée à une action pour un agent donné. Ici le premier exemple signifie que l'agent pense que courtiser Marian est désirable pour l'agent Robin_Hood, le second exprime la connaissance du caractère indésirable de l'action de courtiser Friar_Tuck pour Marian. Cela ne donne aucune information sur la désirabilité de cette action pour les autres agents. Ce prédicat peut ainsi être indifféremment employé pour représenter la désirabilité de ses propres actions et connaissance de la désirabilité d'une action pour un autre agent.

De manière analogue, nous dotons les agents de prédicats permettant de connaître le caractère possible ou impossible d'une action comme l'illustre l'exemple d'implémentation V.4.

```
1        possible_eval(Agent1,tax(Agent2),possible):-
2          noble(Agent1)& not poor(Agent2).
3
4        possible_eval(Agent1,tax(Agent2),impossible):-
5          not possible_eval(Agent1,tax(Agent2),possible).
```

Implémentation V.4 – Exemple de connaissances sur la possibilité d'une action

Dans cet exemple, l'agent sait qu'il n'est possible pour un agent de taxer un autre agent que si l'agent effectuant l'action est noble et que celui qui est taxé n'est pas pauvre. Dans le cas contraire, l'agent sait que cela est impossible. Notons que la seconde règle est indispensable pour que l'agent ait la connaissance explicite de l'impossibilité d'effectuer cette action, ce qui est différent de ne rien connaître de son caractère possible ou impossible.

### V.1.4 Moralité des actions

La moralité des actions est évaluée grâce à l'ensemble des valeurs et règles morales de l'agent. Les valeurs peuvent regrouper des sous-valeurs, c'est-à-dire des valeurs pour lesquelles, dès lors qu'une action est conforme à l'un de ses supports, cette action est également conforme à la valeur dont elle est une sous-valeur. Par exemple, si l'honnêteté est une sous-valeur de la bienveillance (comme semble l'indiquer le modèle de Schwartz, voir figure I.5), toute action supportant la valeur d'honnêteté supporte également la bienveillance.

L'exemple d'implémentation V.5 reprend l'exemple employé jusqu'ici pour montrer comment décrire les valeurs et leurs supports.

```
1      // Declaration of values
2      value("benevolence").
3      value("conformity").
4      value("security").
5
6      // Hierarchy of values
7      subvalue("honesty",     "benevolence").
8      subvalue("generosity",  "benevolence").
9      subvalue("obedience",   "conformity" ).
10     subvalue("social order","security"   ).
11
12     valueSupport(A,V2):- valueSupport(A,V1) & subvalue(V1,V2).
13     valueDefeat(A,V2):-  valueBetray(A,V1)  & subvalue(V1,V2).
14
15     // Description of some supports
16     valueSupport(give(A),"generosity") :- poor(A).
17     valueDefeat(tax(A),  "generosity") :- poor(A).
```

Implémentation V.5 – Exemple de connaissances sur les valeurs

Dans cet exemple, l'agent dispose de la description de trois valeurs morales principales (bienveillance, conformité et sécurité) et quatre sous-valeurs associées à des relations d'ordre hiérarchiques (l'honnêteté et la générosité sont des sous-valeurs de la bienveillance, l'obéissance est une sous-valeur de la conformité, l'ordre social est une sous-valeur de la sécurité). Deux supports illustrent la description de supports ou de trahison de valeurs par des actions dans certains contextes.

Le support des valeurs peut ensuite être employé pour décrire la moralité des actions par des règles morales (dans le cas d'une approche vertueuse). Ces règles morales peuvent également être exprimées sans recourir aux supports de valeurs (dans le cas d'une approche déontologique ou conséquentialiste), comme le montre l'exemple d'implémentation V.6

```
1      // It is immoral to betray the benevolence moral value
2      moral_eval(_,Action,V1,immoral) :-
3        valueBetray(Action,V1) & subvalue(V1,"benevolence").
4
5      // It is moral to give to a poor agent
6      moral_eval(_,give(A),"giving is good",moral) :- poor(A).
7
8      // It is immoral to impoverich another agent
9      moral_eval(A1,Action,"To impoverich someone is bad",immoral) :-
10       consequence(Action,poor(A2)) & A1 \== A2.
```

Implémentation V.6 – Exemple d'implémentation de règles morales

Ces règles morales permettent d'affecter à des actions, en fonction du contexte et des supports de valeurs morales, des valuations de moralité. Notons que le troisième paramètre permet de désigner la règle morale à l'aide d'une chaîne de caractère. Nous employons ensuite ces désignations pour relier les règles à des ensembles moraux (par exemple un ensemble de règles partagé par des agents ou lié à une valeur) qui seront employés pour évaluer les comportements, comme l'illustre l'exemple d'implémentation V.7.

```
1      moralSet("giving is good",              "merry men's rules").
2      moralSet("To impoverich someone is bad","merry men's rules").
```

Implémentation V.7 – Exemple d'implémentation d'un ensemble moral

L'ensemble des valuations de moralité employé par l'ensemble des règles morales pour exprimer la moralité des actions doit être défini et ordonné, par exemple à la manière de l'implémentation V.8.

```
1      // Order on moral valuations
2      order_on_moral_valuation(very_moral, moral).
3      order_on_moral_valuation(moral, amoral).
4      order_on_moral_valuation(amoral, immoral).
5      order_on_moral_valuation(immoral, very_immoral).
6
7      // Transitivity on this order
```

```
8        torder_on_moral_valuation(MV1, MV2):-
9          order_on_moral_valuation(MV1, MV2).
10       torder_on_moral_valuation(MV1, MV2):-
11         order_on_moral_valuation(MV1, MV3)
12         & torder_on_moral_valuation(MV3, MV2).
13
14       // Comparison
15       moralAtLeast(Agent,Action,MV):-
16         moral_eval(Agent,Action,_,MV1) &
17         torder_on_moral_valuation(MV1,MV).
18
19       moralAtLeast(Agent,Action,MV):-
20         moral_eval(Agent,Action,_,MV).
21
22       moralAtMost(Agent,Action,MV):-
23         moral_eval(Agent,Action,_,MV1) &
24         torder_on_moral_valuation(MV,MV1).
25
26       moralAtMost(Agent,Action,MV):-
27         moral_eval(Agent,Action,_,MV).
```

Implémentation V.8 – Exemple d'implémentation de l'ensemble ordonné des valuations morales

Dans cet exemple nous avons défini l'ensemble ordonné des valuations de moralité comme étant composé de (`very_immoral`, `immoral`, `amoral`, `moral`, `very_moral`) dans cet ordre croissant de moralité. Les prédicats `moralAtLeast` et `moralAtWorst` permettent de vérifier qu'il existe au moins une règle morale affectant une valuation de moralité au moins (respectivement au plus) aussi haute que celle utilisée en comparaison. Ces prédicats vont permettre aux principes éthiques de comparer les valuations de moralité affectées aux actions moralement évaluées.

### V.1.5 Évaluation de l'éthique des actions

Nous dotons également les agents d'un ensemble de principes éthiques ordonné par des relations de préférences. Ces principes éthiques peuvent être très simples, à l'image du principe présenté par l'exemple d'implémentation V.9 qui exprime la théorie selon laquelle « Une action possible, désirable, non indésirable, morale (ou mieux) et pas immorale (ou pire), est une action juste. ». Des principes plus complets tels que celui en partie présenté par l'exemple d'implémentation V.10 sont également employés.

```
1        ethPrinciple(Agent,"perfectAct",Action):-
2          possible_eval(Agent,Action, possible) &
3          desire_eval(Agent,Action,desired) &
4          not desire_eval(Agent,Action,undesired) &
5          moralAtLeast(Agent,Action,moral) &
6          not moralAtMost(Agent,Action,immoral).
```

Implémentation V.9 – Exemple d'implémentation de l'éthique d'Aristote

```
1          // Aristotelian ethics inspired from
```

```
2 //  - [1] J.G. Ganascia, "Modelling ethical rules of lying with Answer Set
      Programming"
3 //  - [2] J.G. Ganascia, "Ethical System Formalization using Non-Monotonic
      Logics"
4 ethPrinciple(Agent,"aristotelian",Action):-
5   possible_eval(Agent,Action, possible) &
6   aristotelian_just(Agent,Action).
7
8 // These Rules are directly translated from [1]
9 aristotelian_just(Agent,Action):-
10   not aristotelian_unjust(Agent,Action).
11
12 aristotelian_unjust(Agent,Action):-
13   aristotelian_worst_consequence(Action, Consequence)
14   & aristotelian_worst_consequence(OtherAction, ConsequenceOtherAction)
15   & aristotelian_less_moral(Agent,Consequence, ConsequenceOtherAction).
16
17 aristotelian_worst_consequence(Action,Consequence):-
18   aristotelian_consequence(Action,Consequence)
19   & not aristotelian_not_worst_consequence(Action, Consequence).
20
21 aristotelian_not_worst_consequence(Action, Consequence):-
22   aristotelian_consequence(Action, Consequence)
23   & aristotelian_consequence(Action, OtherConsequence)
24   & aristotelian_less_moral(OtherConsequence, Consequence)
25   & not aristotelian_less_moral(Consequence, OtherConsequence).
26
27 aristotelian_consequence(Action,Action).
28
29 // Here we adapt these rules in our model
30 aristotelian_less_moral(Agent,E1, E2):-
31   moral_eval(Agent,E1,V1,MV1)
32   & moral_eval(Agent,E2,V2,MV2)
33   & torder_on_moral_valuation(MV2, MV1)
34   & not aristotelian_not_less_moral(Agent,E1, E2).
35
36 aristotelian_not_less_moral(Agent,E1, E2):-
37   moral_eval(Agent,E1,V1,MV1)
38   & moral_eval(Agent,E2,V2,MV2)
39   & not torder_on_moral_valuation(MV2, MV1)
40   & not equals_actions(E1,E2).
```

Implémentation V.10 – Exemple d'implémentation d'un principe éthique simple

L'agent est également doté d'un ordre de préférence sur les principes éthiques et de règles permettant d'évaluer l'action satisfaisant le mieux cet ensemble ordonné afin de déterminer l'action juste. L'exemple d'implémentations V.11 illustre ce fonctionnement.

```
1        //* Ethical preferences *//
2       prefEthics("aristotelian","perfectAct").
3       prefEthics("perfectAct","desireNR").
4       prefEthics("desireNR","dutyNR").
```

```
5
6       // This section describe transitivity on the preferences
7       tPrefEthics(PE1,PE2):- prefEthics(PE1,PE2).
8       tPrefEthics(PE1,PE2):- prefEthics(PE1,PE3) & tPrefEthics(PE3,PE2).
```

Implémentation V.11 – Exemple d'implémentation de l'ensemble ordonné des principes éthiques

Le prédicat `ethicalJudgment` permet enfin d'évaluer l'action juste comme étant celle pour laquelle il n'existe pas d'action satisfaisant davantage l'ensemble ordonné des principes éthiques. Ainsi le jugement d'une action nécessite de comparer sa justesse à celle de toutes les autres actions au regard de l'ensemble des principes éthiques. L'exemple d'implémentation V.12 illustre la description de cette fonction.

```
1       existBetter(A,PE1,X):-
2         ethPrinciple(A,PE1,X)
3         & tPrefEthics(PE2,PE1)
4         & ethPrinciple(A,PE2,Y)
5         & not ethPrinciple(A,PE1,Y).
6
7       ethicalJudgment(A,X,PE):-
8         ethPrinciple(A,PE,X)
9         & not existBetter(A,PE,X).
```

Implémentation V.12 – Implémentation du jugement

Ce jugement éthique des actions est enfin employé comme un élément de contexte nécessaire à la sélection des plans comme des intentions. Ce mécanisme permet de faire intervenir le jugement dans la décision sans nécessiter de modifier l'architecture des agents. L'exemple d'implémentation V.13 illustre ce fonctionnement.

```
1       +!act : myName(N)
2         & ethicalJudgment(N,give(A),PE)
3         <- .print("I decide to give to ", A, " according with ", PE);
4            give(A);
5            !act.
6
7       +!act : myName(N)
8         & ethicalJudgment(N,steal(A),PE)
9         <- .print("I decide to steal ", A, " according with ", PE);
10           steal(A);
11           !act.
12
13      // If there is no rightfull action, wait a second
14      +!act : true
15        <- .wait(1000);
16           !act.
```

Implémentation V.13 – Utilisation du jugement comme processus décisionnel

Relevons que, comme l'illustre l'exemple d'implémentation V.13, dans l'éventualité où aucune des deux actions `steal` et `give` ne serait jugée juste, un plan est tout de même

présent pour permettre à l'agent d'attendre et recommencer son évaluation. L'absence de plan dont le contexte est valide pour satisfaire un but provoquerait une erreur dans Jason. Il est donc nécessaire de prévoir dans tout contexte applicatif le comportement que l'agent doit adopter en cas de dilemme éthique dans lequel aucune action ne satisfait le moindre principe éthique.

### V.1.6 Jugement des autres agents

Lors de l'observation des actions des autres agents (par exemple par un signal envoyé par un artefact de l'environnement), les agents jugent le comportement des agents impliqués dans l'action et mettent à jour les images de la morale et de l'éthique du comportement concernées. Pour cela, la reconnaissance de situation est enrichie de plans déclenchés lors d'observations qui ajoutent à l'ensemble des buts actifs de l'agent deux buts représentant la nécessité d'évaluer respectivement la moralité et l'éthique du comportement de l'agent. En outre, ce plan ajoute également l'évaluation de possibilité indiquant que cette action a été possible (ce qui est manifestement le cas puisqu'elle a été réalisée). L'exemple d'implémentation V.14 montre l'un de ces plans.

```
1       +executedAction(DateOfExecution,Agent,Action):-
2         +possible_eval(Agent,Action,possible);
3         !evaluateMorality(DateOfExecution, Agent,Action);
4         !evaluateEthics(DateOfExecution, Agent,Action).
```

Implémentation V.14 – Éléments de la reconnaissance de situation permettant d'effectuer le jugement d'un comportement

Pour évaluer l'éthique du comportement, l'agent dispose d'un ensemble de plans afin d'atteindre le but généré. Ces plans permettent de faire face aux différents cas de figure (si l'action est jugée éthique ou non, si une image existe déjà ou non). L'exemple d'implémentation V.15 montre le plan permettant d'incrémenter la croyance représentant le nombre d'actions dans $EC^+$ si l'action est jugée conforme à l'éthique et qu'une image de l'éthique de ce comportement existe déjà. Notons l'usage du mot-clé `atomic` permettant d'empêcher que ce plan soit interrompu en cours d'exécution par la réalisation d'un autre plan. Il est indispensable ici pour ne pas interrompre l'évaluation d'une action et l'apparition de problèmes d'accès concurrents à la base de croyances.

```
1     @evalethics1[atomic]
2     +!evaluateEthics(DateOfExecution, Agent,Action) :
3       ethicalJudgment(Agent,Action,PE) & ethicalAggr(Agent,X)
4       <-  .abolish(evaluateEthics(DateOfExecution, Agent, Action));
5         .abolish(ethicalAggr(Agent,_)); // remove the previous |EC+| value
6         .abolish(possible_eval(Agent,Action, possible));
7         +ethicalAggr(Agent,X+1); // increment |EC+|
8         +rebuildImageOf(Agent);
9         !buildEthicalImage.
```

Implémentation V.15 – Exemple de plan permettant d'évaluer l'éthique d'un comportement

Ce plan se termine par l'ajout d'un nouveau but indiquant que maintenant que l'évaluation du comportement est mise à jour, l'agent juge doit réactualiser son image du comportement de l'agent jugé. Cette actualisation est effectuée par un ensemble de plans tels que celui présenté par l'exemple d'implémentation V.16 permettant d'attribuer une valeur qualitative à l'image concernée en fonction de l'intervalle sur lequel se situe la valeur retournée par la fonction d'agrégation éthique (voir section IV.3.2).

```
1     +!buildEthicalImage : rebuildImageOf(Agent)
2       & unethicalAggr(Agent,U) // get |EC-|
3       & ethicalAggr(Agent,E)   // get |EC+|
4       & E/(U+E)<0.4    // comparison with a threshold
5       <- .abolish(ethicalImage(Agent,_));
6          changeImageof(Agent,"ethics",E,U);// save |EC+| and |EC-| in file
7          -rebuildImageOf(Agent);
8          +ethicalImage(Agent,improper); // set the discrete value
9          !buildEthicalImage.
```

Implémentation V.16 – Exemple de plan permettant de mettre à jour l'image de l'éthique d'un comportement

Nous enregistrons l'évolution de ces images dans un fichier afin de visualiser leur évolution en section V.4. De manière analogue, un ensemble de plans permet de construire et faire évoluer les images de la moralité du comportement. Ces images sont construites par rapport aux ensembles de règles regroupées en ensembles moraux comme l'a illustré l'exemple d'implémentation V.7. L'agent dispose ainsi à la fin de ces fonctions d'un ensemble de croyances, mises à jour à chaque perception d'une action, décrivant la conformité du comportement des autres agents au regard de ses ensembles moraux et de son éthique. Ces croyances peuvent alors être employées pour décrire de nouvelles valeurs morales et règles morales décrivant le comportement qu'il est moral d'adopter envers un agent en tenant compte de l'image que l'on a de lui, par exemple pour lui accorder sa confiance comme l'illustre l'implémentation V.17.

```
1       +!trust : myName(N)
2         & ethicalJudgment(N,trust(Agent,MoralSet),PE)
3         & moralImageOf(Agent,MoralSet,ConformityValuation)
4         & trustThreshold(Threshold)
5         & not tOrderOnConformityValuation(Threshold,ConformityValuation)
6         & not trust(Agent,MoralSet)
7         <-.print("I trust ",Agent, "for",MoralSet," according with ",PE);
8           +trust(Agent,MoralSet);
9           !trust.
```

Implémentation V.17 – Exemple de plan permettant d'accorder une confiance en un ensemble moral au dessus d'un seuil

## V.2 Application à la gestion d'actifs financiers

Afin d'illustrer le fonctionnement de l'implémentation du modèle que nous avons présenté dans la section précédente, nous avons choisi le domaine applicatif de la gestion

d'actifs financiers dans lequel la problématique de l'éthique du comportement des investisseurs est connue et documentée (voir section V.2.2). Cette section a pour but d'illustrer le fonctionnement de l'environnement dans lequel les agents seront en interaction lors des expérimentations que nous présentons dans les sections suivantes.

### V.2.1 Les agents autonomes sur les marchés financiers

Un marché financier est un espace dans lequel des agents peuvent échanger des actifs financiers (*assets*) de diverses natures : des monnaies (*currency*) sont échangées sur les marchés des changes, des parts de capitaux (*stock*) sont échangées sur les marchés d'actions, des dettes publiques ou privées sur les marchés monétaires et obligataires, etc. Les marchés ont de nombreuses fonctions dans les économies dites « de marché » (par opposition aux économies dites « planifiées » ) parmi lesquelles nous pouvons compter :

— permettre *la rencontre de l'offre et de la demande*, c'est-à-dire mettre en relation des vendeurs et des acheteurs de manière à procéder à des transactions. Par exemple dans le cadre du financement d'un état, il s'agit de faciliter la rencontre entre les services financiers de l'état souhaitant emprunter de la monnaie et les investisseurs à la recherche d'un placement.

— déterminer *la valeur des actifs*, qui fluctue en fonction des variations de l'offre et de la demande. Certains économistes affirment que ce modèle permet aux prix de trouver leur propre équilibre (Smith, 1937) nous parlons alors d'hypothèse de l'efficience des marchés.

La dématérialisation des marchés financiers initiée à la fin du vingtième siècle a mené à une situation actuelle dans laquelle ces fonctions sont assurées par des systèmes informatiques recevant des ordres d'achat et de vente, dont la forme est définie par des protocoles (par exemple : FIX), et pour lesquels la décision peut être prise par des opérateurs humains ou des agents autonomes. Quelques études montrent que ces derniers sont responsables de la majorité des transactions effectuées et que leur comportement impacte le fonctionnement des places de marché (Tuominen *et al.*, 2012). La conception d'agents autonomes pour la finance soulève un large éventail de problématiques (Aldridge, 2009). Bon nombre d'entre elles, hors du cadre de ce chapitre, concernent la maximisation de la vitesse de la prise de décision et de l'émission d'ordre. Ce domaine est celui du trading haute fréquence, axé sur la recherche de bénéfices optimaux par spéculation. D'autres problématiques concernent l'intégration dans la prise de décision de notions plus complexes : estimation de risque à moyen ou long terme, impact sur le cours de l'actif ou encore discrétion de la transaction. Nous montrons dans la section suivante que la dimension éthique de la décision est une question complexe et, bien que de plus en plus de gestionnaires proposent des placements intégrant des critères éthiques dans la sélection des actifs (Bono *et al.*, 2013), la littérature dans ce domaine ne semble pas encore proposer de modèles permettant de déléguer ce type de décision à des agents autonomes.

### V.2.2 L'éthique dans la finance

L'échange d'actifs sur des marchés financiers fait l'objet de nombreuses réflexions éthiques [2]. L'une des principales raisons à ces interrogations provient du fait que les décisions des agents autonomes, auxquels sont confiés les décisions d'achats et de ventes d'actifs appartenant à des humains, impactent l'économie réelle (Economic and Financial Affairs, 2009). Certains analystes considèrent l'usage de techniques d'automatisation des activités financières comme introduisant en soi de nombreux effets pervers tels que des formes de manipulation de marchés, de concurrence déloyale envers les petits investisseurs et des brusques variations des cours par effet cascade. D'autres arguent que cela réduit la volatilité, accroît la transparence, la liquidité et la stabilité des marchés avec un coût d'exécution des ordres plus faibles (Aldridge, 2009). Les fonds d'investissement éthiques se multiplient et semblent peu à peu prendre une place significative sur les marchés [3]. Toutefois, si des indicateurs objectifs efficaces permettent de mesurer la performance de ces fonds en termes de rendement, l'éthique de leur comportement reste plus difficile à évaluer et reste au moins partiellement sensible aux valeurs de l'évaluateur.

La prise en compte de critères éthiques ne semble pas remettre en question la rentabilité de ces placements (Kempf, Osthoff, 2007) et apporte aux investisseurs une satisfaction supplémentaire qui est celle de voir leurs convictions respectées. Ces convictions peuvent être d'origines diverses : les doctrines religieuses (Guéranger, 2009) et politiques (Hamilton *et al.*, 1993) proposent des ensembles de critères permettant d'évaluer la qualité éthique d'un investissement. Des outils mis à la disposition du grand public [4] permettent de trouver les placements les plus adaptés à leurs convictions personnelles. Nous cherchons à montrer comment notre modèle de jugement peut permettre de concevoir des agents autonomes gestionnaires d'actifs financiers dotés de théories du bien et du juste variées pouvant être paramétrés pour adopter des comportements conformes aux attentes des usagers qui leurs délèguent la gestion de leur portefeuille sur les marchés.

Notons enfin qu'en plus de ces considérations liées aux actifs échangés, la bonne conduite d'un agent sur un marché passe aussi par la manière dont ses échanges sont effectués. Le bon fonctionnement d'un marché financier nécessite une certaine transparence des investisseurs (c'est-à-dire la non-dissimulation de ses intentions pour ne pas fausser la perception de l'offre et de la demande des autres investisseurs) et de bienveillance à l'égard des cours (non-manipulation des cours, minimisation de l'impact des transactions sur les prix). Certaines pratiques telles que les *ordres icebergs* consistant à découper un ordre d'achat en plus petits ordres, pour minimiser l'impact d'une transaction au détriment de la visibilité des autres sur l'intention du donneur d'ordre, ont été grandement facilitées par l'usage croissant de systèmes autonomes capables d'automati-

2. http://sevenpillarsinstitute.org/

3. Par exemple, les placements socialement responsables, solidaires ou écoresponsables de https://www.amundi.fr/fr_part/Nos-Fonds ou http://www.aberdeen-asset.fr/

4. Citons par exemple l'outil http://www.ethicalconsumer.org/buyersguides/money/ethicalinvestmentfunds.aspx proposant au consommateur des placements éthiques ordonnés grâce à une interface lui permettant de pondérer les différents critères (respect de l'environnement, des droits des animaux, qualité des produits, neutralité politique des entreprises, etc.).

ser ces techniques. De telles pratiques sont moralement discutables : elles diminuent la volatilité et accroissent la liquidité (deux bonnes propriétés du point de vue de nombreux acteurs des marchés), mais nuisent fortement à la transparence (ce qui est une mauvaise chose car cela fausse la perception que les acteurs ont de l'offre et de la demande à un instant donné). Ce type d'ordre cause donc un conflit entre des valeurs et règles morales, ce qui met en évidence la nécessité d'un raisonnement éthique.

### V.2.3 Description du simulateur

Dans le système conçu pour nos expérimentations et illustrée par la figure V.2, nous considérons une place de marché sur laquelle des agents autonomes peuvent échanger des actifs présents dans leurs portefeuilles. Ces actifs sont à la fois des devises (monnaies) et des participations de capitaux privés.

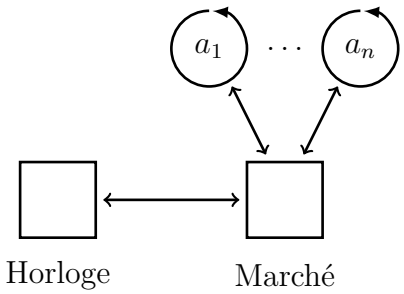


FIGURE V.2 – Architecture du système de place de marché pour la preuve de concept de notre proposition : des agents $a_i$ interagissent sur une place de marché partagée dont le fonctionnement est rythmé par une horloge commune

La place de marché est mise en œuvre par un artefact situé dans l'environnement partagé et donc accessible aux agents par un ensemble d'opérations et un ensemble de propriétés observables et de signaux qui leur permettent de reconnaître la situation.

L'ensemble des actions à la disposition d'un agent sont des ordres d'achat (`buy`), de vente (`sell`) ou d'annulation (`cancel`). Nous ajoutons à ces actions mises à disposition par l'artefact de place de marché une quatrième action consistant à attendre une seconde, comme présentée en section V.1.5, qui reste possible en toute situation et permet d'offrir une alternative en cas de dilemme éthique dans lequel toute autre action serait injuste au regard de la totalité des principes éthiques connus. Les actions `buy`, `sell` et `cancel` consistent respectivement en l'échange de devises contre des participations, l'échange de participations contre des devises et l'annulation d'un ordre en attente d'exécution sur le marché. Pour chaque ordre de vente ou d'achat, l'agent peut spécifier un prix limite ou accepter le prix actuel du marché. De même, le volume de titres échangés est précisé lors de l'envoi de l'ordre. L'artefact de place de marché met également à disposition des agents une information leur permettant de s'abonner aux informations concernant l'évolution de la cotation d'un ou plusieurs actifs.

Les titres de participation sont cotés sur le marché par des structures classiques de *Central Limit Order Book* (CLOB) (Aldridge, 2009 ; Gould *et al.*, 2013). Un CLOB

conserve et trie par ordre de prix l'ensemble des ordres d'achat et de vente, placés respectivement du côté de la demande (*bid*) et de l'offre (*ask*) du marché. Chaque agent peut ainsi insérer ses ordres d'un côté ou de l'autre et le CLOB se comporte alors selon les règles simples qui suivent :

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

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

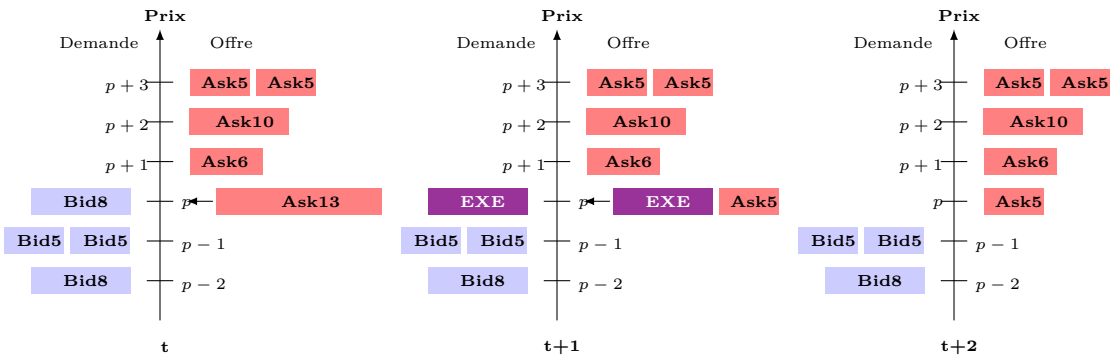


FIGURE V.3 – Exécution d'un ordre lors de son ajout sur le marché

L'exemple présenté sur la figure V.3 illustre l'insertion d'un ordre de vente de treize participations au prix $p$. Avant l'insertion, la meilleure demande est au prix $p$ et la meilleure offre est à $p+1$. L'ordre inséré correspond en termes de prix à un ordre situé de l'autre côté. Le plus grand des deux est alors partiellement exécuté, le plus petit est exécuté dans sa totalité (c'est-à-dire Bid8 est supprimé), et le reste du plus grand est placé dans le CLOB (c'est-à-dire Ask5, voir représentation à t+1). À l'issue de l'insertion, la nouvelle meilleure demande est de prix $p-1$ et la nouvelle meilleure offre est de prix $p$. Tous ces changements sont perçus par les agents par la mise à jour de croyances sur l'évolution du marché et l'état de leur portefeuille. Les indicateurs sont recalculés et diffusés toutes les minutes tandis que les informations sur la présence d'ordres sur le marché ou leur exécution sont diffusées en temps réel. Les croyances sur l'état courant du marché sont présentées par l'exemple d'implémentation V.18.

```
1        indicators(Date,Marketplace,Asset,Close,Volume,
2            Intensity,Mm,Dblmm,BollingerUp,BollingerDown).
3
4        onMarket(Date,Agent,Portfolio,Marketplace,
5            Side,Asset,Volume,Price).
6
7        executed(Date,Agent,Portfolio,Marketplace,
8            Side,Asset,Volume,Price).
```

Implémentation V.18 – Croyances des agents sur la situation courante produites par la reconnaissance de situation

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

Les titres (`Asset`) présents dans le portefeuille de l'agent sont représentés par des croyances tenues à jour à chaque changement sous la forme présentée par l'exemple d'implémentation V.19.

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

Implémentation V.19 – Croyances des agents sur l'état de leur portefeuille produites par la reconnaissance de situation

En raisonnant sur ses croyances, un agent est alors capable d'évaluer la possibilité de passer un ordre d'achat ou de vente (en vérifiant s'il possède ou non assez de devises ou de titres) afin de générer l'ensemble des actions possibles $\mathcal{A}_{c_{a_i}}$. L'objet de ces expérimentations n'étant pas de montrer une quelconque performance dans la sélection des investissements, une fonction d'évaluation de la désirabilité génère l'ensemble $\mathcal{A}_{d_{a_i}}$ à l'aide d'une méthode simple et classique basée sur la comparaison des moyennes mobiles et des bandes de Bollinger (Leung, Chong, 2003).

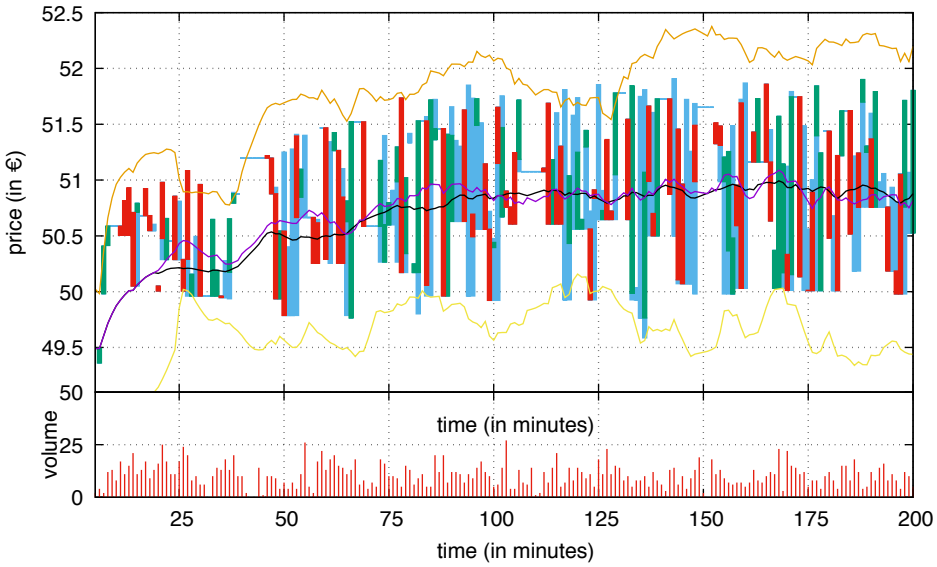


FIGURE V.4 – Représentation en chandeliers de l'évolution du cours d'un titre lors d'une expérimentation.

La figure V.4 présente une représentation graphique sous la forme de chandeliers des informations sur l'évolution du cours d'une action. Les chandeliers montrent l'évolution du prix, et les deux courbes situées au milieu représentent les moyennes mobiles

évoquées précédemment. Nous remarquons qu'il est rare que les chandeliers passent en dehors des bandes de Bollinger représentées par les courbes encadrant les évolutions de prix. Le volume représenté en histogramme dans la partie inférieure montre une activité souvent intensifiée lorsque les moyennes mobiles s'entrecroisent. Ces volumes d'échanges sont explicables en raison de la fonction d'évaluation de la désirabilité qui considère l'intersection des moyennes comme le signe d'un renversement de tendance (haussière ou baissière), entraînant un changement de position des agents (acheteur ou vendeur).

Notons que l'implémentation d'un tel artefact représentant une place de marché nécessite la prise de nombreuses précautions dues aux problèmes occasionnés par les accès concurrents. En effet, ce composant logiciel est implémenté en employant des techniques de programmation événementielle afin de réagir aux signaux envoyés par les agents et les autres artefacts (horloge, interfaces graphiques de visualisation de la situation, etc.).

## V.3 Évaluation de l'influence du jugement sur le comportement individuel

Après avoir décrit en section V.1 l'implémentation du modèle de jugement et présenté en section V.2 le domaine applicatif et l'implémentation de sa représentation pour effectuer des expérimentations, nous présentons dans cette section un ensemble d'expérimentations permettant d'éprouver l'influence sur le comportement des agents du modèle de jugement employé dans la décision. Cette démarche a pour objectif de montrer comment il est possible de décrire dans ce modèle des valeurs et règles morales réellement employées par les acteurs du domaine applicatif et apporter des résultats empiriques montrant que le modèle de jugement proposé permet d'amener les agents à se conformer à la morale et l'éthique qui lui sont confiées. Nous présentons pour cela les connaissances du contexte, du bien et du juste données en paramètres du modèle de jugement des agents en précisant la méthodologie employée pour leur écriture. Ensuite nous présentons la méthode utilisée pour l'initialisation des expérimentations afin de générer des situations réalistes. Enfin, nous observons et interprétons les résultats issus des expérimentations.

### V.3.1 Description métier du domaine

Les agents reçoivent dès le début de l'expérimentation un ensemble de croyances décrivant les actifs présents sur le marché. L'exemple d'implémentation V.20 illustre quelques unes de ces croyances.

```
1  activity("EDF","nuclear_energy_production").
2  activity("LEGRAND","Diversified_Machinery").
3  activity("ACCOR","Lodging").
4  activity("AB SCIENCE","Drug Manufacturers - Major").
5  activity("ENGIE","nuclear_energy_production").
6  activity("ENGIE","natural_gas").
7  activity("ENGIE","renewable_energy_production").
8  activity("TOTAL","oil_production").
```

```
9
10  label("LEGRAND","FSC").
11
12  lobby("Bilderberg Group").
13  lobby("Business Roundtable").
14  lobby("European Round Table of Industrialists").
15  lobby("World Business Council for Sustainable Development").
16  lobby("World Economic Forum").
17
18  // Source : http://www.ert.eu/members
19  memberOf("ENGIE","European Round Table of Industrialists").
20  memberOf("TOTAL","European Round Table of Industrialists").
21
22  // Source : http://www.wbcsd.org/Overview/Our-members
23  memberOf("EDF","World Business Council for Sustainable Development").
24  memberOf("ENGIE","World Business Council for Sustainable Development").
25
26  // Source : https://www.weforum.org/about/industry-affiliations
27  memberOf("ACCOR","World Economic Forum").
28  memberOf("ENGIE","World Economic Forum").
29  memberOf("TOTAL","World Economic Forum").
```

Implémentation V.20 – Exemples de croyances initiales sur les entreprises cotées

Ces croyances ont pour principal intérêt de permettre de définir le contexte des supports de valeurs et des règles morales.

## V.3.2 Paramétrage moral des agents

Dans le cadre de ces expérimentation, nous avons doté les agents autonomes de valeurs et règles morales diverses représentant des convictions réelles d'usagers. Le paramétrage éthique des agents reste invariable dans le cadre de ces expérimentations. N'étant pas expert en éthique financière ni en contact direct avec un grand nombre de ces usagers, nous avons utilisé des documents disponibles [5] détaillant des ensembles de critères (liés au respect de l'environnement, à la neutralité politique, au respect des droits des travailleurs, etc.) permettant d'évaluer la moralité d'un investissement dans une approche vertueuse de l'éthique et de la morale.

### V.3.2.1 Valeurs

Chaque agent éthique est alors doté d'un ou plusieurs ensembles de valeurs hiérarchisées (voir section III.5.1) correspondant aux ensembles de critères évoqués précédemment. Nous détaillons dans la suite de cette section l'implémentation de l'ensemble moral `ecology` regroupant des valeurs morales et règles morales décrivant des convictions liées à la protection de l'environnement. Nous commençons par décrire l'ensemble hierarchisé des valeurs de l'agent. Par exemple « *environmental reporting* »est décrit comme une sous-valeur de la valeur « *environment* ». Les valeurs sont décrites sous la forme de

5. Par exemple, le document mis à disposition par le site http://www.ethicalconsumer.org/ détaille un ensemble de conditions permettant de qualifier d'éthique une entreprise cotée.

prédicats logiques illustrés par l'exemple d'implémentation V.21.

```
1        value("environment").
2        subvalue("promote_renewable_energy","environment").
3        subvalue("environmental_reporting","environment").
4        subvalue("fight_climate_change","environment").
```

Implémentation V.21 – Déclaration de la valeur environment et de ses sous-valeurs

### V.3.2.2 Supports de valeurs

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

```
1        valueDefeat(buy(Asset,_,_,_), "promote_renewable_energy"):-
2          activity(Asset,"nuclear_energy_production").
3        valueDefeat(sell(Asset,_,_,_), "promote_renewable_energy"):-
4          activity(Asset,"nuclear_energy_production").
```

Implémentation V.22 – Déclarations de support de valeurs liés aux énergies renouvelables

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

```
1        valueSupport(buy(Asset,_,_,_), "environmental_reporting") :-
2          label(Asset,"FSC").
3        valueSupport(sell(Asset,_,_,_), "environmental_reporting") :-
4          label(Asset,"FSC").
```

Implémentation V.23 – Déclarations de support de valeurs liés au reporting environnemental

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

```
1        valueSupport(buy(Asset,_,_,_), "fight_climate_change") :-
2          activity(Asset,"nuclear_energy_production").
3        valueSupport(sell(Asset,_,_,_), "fight_climate_change") :-
4          activity(Asset,"nuclear_energy_production").
```

Implémentation V.24 – Déclarations de support de valeurs liés à la lutte contre le changement climatique

### V.3.2.3 Règles morales

Ces valeurs sont ensuite employées pour écrire des règles morales leur permettant de lier la promotion des valeurs morales à une valuation morale. Par exemple la transcription des règles morales « Il est moral d'agir conformément à la valeur *environment* » et « Il est immoral de trahir la valeur *environment* » est illustrée par l'exemple d'implémentation V.25.

```
1        moral_eval(X,"Promote environment is good",moral):-
2          valueSupport(X,"environment").
3
```

```
4        moral_eval(X,"Betray environment is bad",immoral):-
5          valueDefeat(X,"environment").
```

Implémentation V.25 – Déclarations de règles morales évaluant les actions trahissant ou promouvant la valeur de respect de l'environnement.

Ces règles morales sont regroupées dans l'ensemble moral définissant l'écologie, comme le montre l'exemple d'implémentation V.26.

```
1        moralSet("Promote environment is good","ecology").
2        moralSet("Defeat environment is bad",  "ecology").
```

Implémentation V.26 – Rassemblement des règles morales liées à l'écologie en un ensemble moral

À ce stade, un agent éthique est capable de déduire que, au regard de ses croyances et de sa théorie du bien, échanger des titres d'une société labelisée FSC est moral, tandis qu'échanger des actifs d'un producteur d'énergie nucléaire est à la fois moral et immoral au regard d'une même valeur. L'agent a donc besoin d'une théorie du juste pour déterminer s'il est éthique ou non d'échanger le second titre. Nous employons dans cette série d'expériences le même ensemble de principes éthiques et de préférences éthiques que ceux décrits en section V.1.5, profitant de l'abstraction faite des actions dans la description de ces principes. L'agent peut ensuite employer la fonction de jugement décrite en section V.1 afin de connaître dans la situation courante les transactions qu'il est juste d'effectuer.

### V.3.3 Initialisation

Afin d'effectuer un ensemble d'expériences, nous avons écrit un programme permettant de générer une série de situations initiales en fonction de plusieurs paramètres. Cette situation initiale est décrite dans plusieurs fichiers produits par ce programme dont : un fichier intitulé « ethicalAssetManagement.jcm » décrivant pour la plateforme JaCaMo les agents et l'environnement à instancier, et un fichier « scenario.xml » décrivant la composition des portefeuilles d'actifs générés aléatoirement et confiés aux agents, la valeur initiale des actifs et les ordres initialement présents sur le marché. Le générateur prend en paramètre une graine permettant d'initialiser le générateur de nombres aléatoires employé, le nombre d'expériences à effectuer, le nombre d'agents à instancier, la valeur des portefeuilles confiés aux agents et la valeur initiale des actifs.

Deux modes de fonctionnement sont prévus pour animer les marchés :

— soit l'animation du marché est assurée par un groupe d'agents sans éthique agissant de manière aléatoire afin de simuler l'activité d'un marché réel ;
— soit un agent spécifique est instancié pour faire suivre au cours de l'action une série de valeurs décrite dans un fichier donné en paramètre.

Ce dernier mode de fonctionnement permet de confronter le comportement des agents à une évolution réelle des cours (nous fournissons des relevés de valeurs produits à partir de sites d'investissement en ligne). Le premier mode est en revanche celui à utiliser si l'expérimentateur veut évaluer l'impact d'un ensemble d'agents sur l'évolution du cours

des actifs.

### V.3.4 Comportement des agents

Cette section présente et commente les résultats d'une série d'expérimentations simulant un marché sur lequel six actifs, dont les cours sont animés par des agents aléatoires, sont échangés par deux agents dotés de la morale ne prenant en compte que l'ensemble moral « ecology », deux agents dotés de règles morales portant sur un ensemble moral décrivant la « bienveillance politique » et deux agents disposant de ces deux ensembles moraux. Les agents emploient les principes éthiques présentés en section V.1 avec le même ordre de préférences. Nous cherchons à illustrer l'utilisation du modèle de jugement éthique comme processus de décision.

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

La figure V.5 représente l'évolution du portefeuille d'un agent éthique écologiste au cours d'une expérimentation de plus de deux heures. Son tracé, caractéristique des agents écologistes, indique par des couleurs l'évolution des proportions d'actions composant le portefeuille au cours de l'expérience. Le tableau V.1 donne les résultats cumulés de dix simulations de trente minutes.

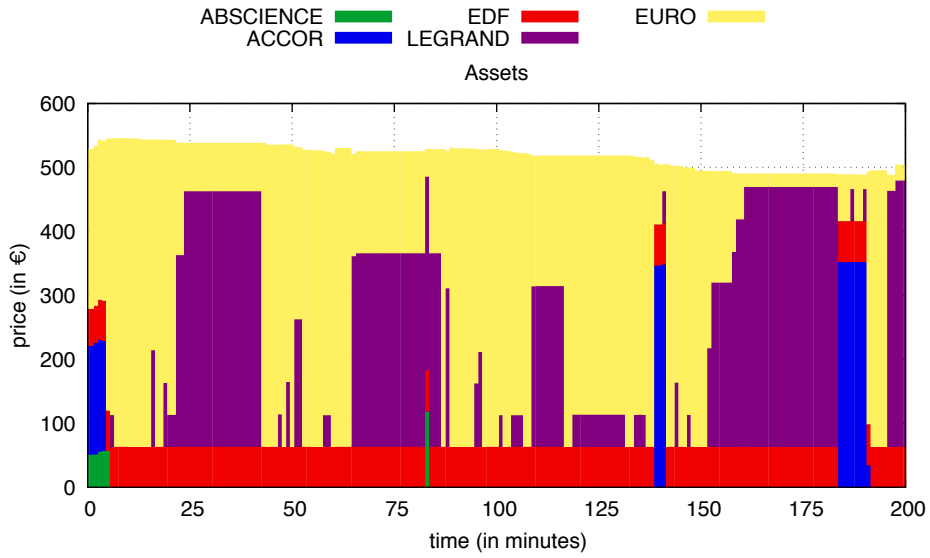


FIGURE V.5 – Évolution de la composition du portefeuille d'actifs d'un agent écologiste

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

Deuxièmement, nous pouvons aussi observer que, durant certaines périodes, le portefeuille intègre un grand nombre de titres de LEGRAND. De fait, l'agent ayant connais-

sance d'une labellisation de cette entreprise, une motivation morale s'ajoute à la recherche de bénéfice et en fait un investissement privilégié (près de quatre-vingt-dix pourcents des transactions selon le tableau V.1 qui récapitule le cumul de dix simulations plus courtes mais paramétrées de façon similaire). Ces actes d'achat et de vente seront en outre les seuls à satisfaire le principe `perfectAct` évoqué en section précédente.

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

L'agent n'ayant reçu aucune action TOTAL ou ENGIE et ne pouvant à aucun moment considérer qu'il est juste d'y investir, ils n'apparaissent pas dans cet histogramme.

Le tableau V.1 récapitule le nombre de transactions par actifs effectuées pour chaque type d'agent sur un ensemble de dix simulations de trente minutes. Le pourcentage indiqué représente la proportion de transactions concernant cet actif par rapport à l'ensemble des transactions opérées par l'agent.

| | AB SCIENCE | | ACCOR | | EDF | | ENGIE | | LEGRAND | | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | achats | ventes | achats | ventes | achats | ventes | achats | ventes | achats | ventes | achats | ventes |
| Écologiste | 281 | 39 | 341 | 31 | 0 | 0 | 0 | 0 | 3254 | 3251 | 0 | 0 |
| | 4.5 % | | 5.2 % | | 0 % | | 0 % | | 90.3% | | 0 % | |
| Bienveillant politique | 3202 | 2941 | 0 | 0 | 0 | 0 | 0 | 0 | 1731 | 1435 | 0 | 0 |
| | 66 % | | 0 % | | 0 % | | 0 % | | 34 % | | 0 % | |
| Les deux | 5152 | 4831 | 179 | 19 | 0 | 0 | 0 | 0 | 3836 | 3315 | 0 | 0 |
| | 58 % | | 1.1 % | | 0 % | | 0 % | | 40.9 % | | 0 % | |

TABLE V.1 – Nombre de transactions par types d'agents en dix simulations de 30 minutes

Les résultats concernant les agents écologistes vient corroborer ceux présentés précédemment. De même les proportions indiquées montrent que les agents dotés uniquement de l'ensemble moral de « Bienveillance politique » n'investissent que dans les actions AB SCIENCE et LEGRAND, dont l'achat ou la vente est motivée par une règle indiquant qu'il est bon d'échanger des actifs d'une entreprise qui semble n'appartenir à aucun lobby. Enfin les agents dotés des deux ensembles moraux semblent avoir adopté un comportement intermédiaire.

Ces résultats permettent d'évaluer l'usage du jugement comme processus de décision. Nous avons montré comment, à l'aide de valeurs morales, règles morales et principes éthiques ordonnées par un ensemble de préférences, il est possible d'obliger les agents à adopter un comportement conforme à une certaine éthique habituellement employée par les humains dans ce domaine.

## V.4 Évaluation du jugement des autres

La section précédente a illustré l'influence du jugement sur le comportement des agents. Nous cherchons maintenant à montrer comment les agents peuvent employer ce jugement pour évaluer le comportement des autres.

## V.4.1 Initialisation

À chaque action exécutée sur le marché, les agents reçoivent un signal de l'artefact et réévaluent leurs images des agents impliqués dans la transaction. Dans la suite de cette section nous détaillons la construction de la confiance éthique. La confiance morale se construit de manière analogue.

Dans l'implémentation actuelle, nous utilisons une agrégation linéaire (c'est-à-dire associant la même pondération à chaque action, voir section IV.3.2). Ensuite une valuation de conformité est attribuée en comparant la proportion d'actions conformes à des seuils afin de construire l'image. Dans cette expérimentation nous n'utilisons que trois niveaux de conformité (arbitrairement `neutral` pour un résultat compris dans $[0.4, 0.6[$, `improper` pour les résultats inférieurs à 0.4 et `congruent` pour les résultats supérieurs ou égaux à 0.6).

## V.4.2 Agrégation d'images

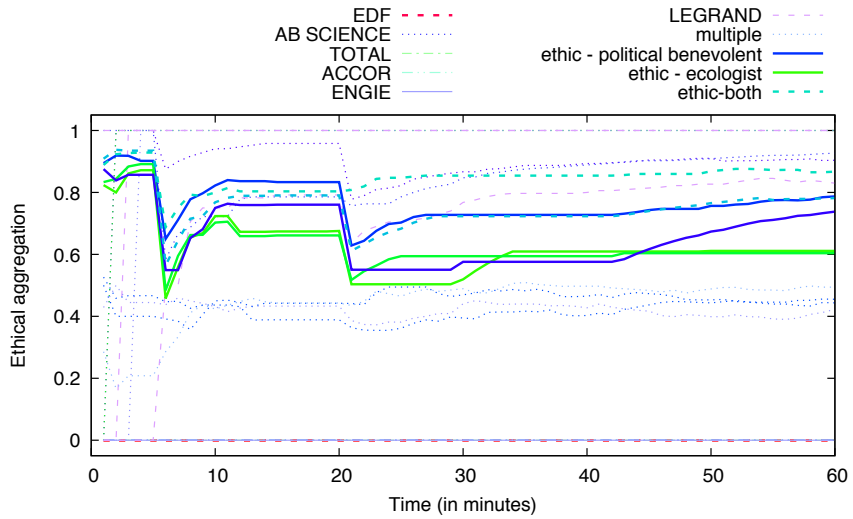


FIGURE V.6 – Évolution des images des agents en sortie de la fonction d'agrégation éthique générées au sein d'un agent de type ethic-both

Afin d'évaluer la construction de ces agrégations, nous observons l'évolution de l'image des agents présents dans une simulation d'une heure. Ces agents sont les vingt agents au comportement aléatoire affectés à l'animation du cours d'un ou plusieurs actifs, deux agents dotés exclusivement de l'ensemble moral « écologie », deux agents dotés exclusivement de l'ensemble moral « bienveillance politique » et deux agents, désignés par « ethic-both », employant les deux ensembles moraux « écologie » et « bienveillance politique ». La figure V.6 montre l'évolution de l'agrégation des images éthiques construites par un jugement aveugle en sortie de la fonction d'agrégation d'un agent de type « ethic-both ».

Remarquons premièrement qu'un agent peut construire une image d'un agent aléatoire ou d'un agent doté d'une éthique différente. C'est l'une des propriétés de ce modèle : nous

n'évaluons que la conformité d'un comportement observé au regard d'une éthique, sans nécessairement chercher à connaître l'intention et les connaissances des autres agents. Remarquons tout d'abord que l'image que l'agent observateur a de lui-même (celle qui est également employée dans la légende pour désigner les agents de type « ethic-both ») n'est pas au maximum bien qu'elle soit plus élevée que toutes celles des autres agents dotés d'une éthique à partir d'une vingtaine de minutes. Cela est explicable par la possible différence de contexte entre le moment ou la décision de transaction est prise et celle où elle est exécutée (les agents ne sont jugés qu'au moment de l'exécution des actions).

Ensuite, relevons que les agents aléatoires en charge de l'animation des cours d'actifs qu'il est moral d'échanger du point de vue de l'agent juge (ACCOR et LEGRAND) restent durant toute l'observation à la valeur maximale de 1.0 (traits épais). À l'inverse, les agents aléatoires affectés à l'animation du cours d'un actif immoral aux regard des connaissances de l'agent juge restent à 0.0 (traits pointillés). Quatre agents aléatoires qui ne sont affectés à aucun actif en particulier oscillent aux alentours de 0.5, ce qui signifie que l'agent juge évalue la moitié de leurs actions comme justes et les autres comme injustes. Enfin les agents en charge de l'action AB SCIENCE voient leur image osciller à une valeur élevée sans pour autant atteindre la valeur maximale. Effectivement, comme l'échange de cet actif n'est éthique que dans l'éventualité où il est impossible d'échanger un autre actif, certaines transactions sont considérées comme injustes.

Enfin, nous nous intéressons à la confiance produite par l'emploi de ces images à la fin de la simulation avec les seuils présentés à la section précédente. Les agents éthiques écologistes ou politiquement bienveillants (traits épais et continus respectivement verts et bleus) se trouvent au dessus du seuil requis fixé précédemment pour recevoir la valuation de conformité la plus élevée. Toutefois leur image atteint un niveau moins élevé que celles des agents dotés des deux ensembles moraux.

Ces résultats montrent des propriétés intéressantes de l'utilisation du jugement pour construire une image des autres agents. Tout d'abord, relevons que les agents paramétrés de la même manière produisent des images relativement semblables, même s'ils n'emploient pas eux-même le jugement éthique comme processus de décision. De plus, l'agent s'évalue lui-même comme étant l'agent qui, parmi ceux qui n'ont pas un comportement nécessairement éthique en raison de contraintes limitant leurs actions, a le comportement le plus conforme à sa propre éthique. Enfin ce modèle permet de différencier de manière notable les agents partageant une partie des valeurs de l'agent juge de ceux qui ont un comportement purement aléatoire.

## V.5 Synthèse

Nous avons montré dans ce chapitre comment il était possible d'implémenter le modèle de jugement présenté en chapitre III et IV à l'aide de la plateforme JaCaMo dédiée à la création de système multi-agents BDI. Cette série d'expérimentations a illustré plusieurs caractéristiques de ce modèle.

Le premier but est d'abord de montrer que notre modèle est opérationnel, puisque

nous en avions fait au chapitre II l'un des objectifs de notre démarche. Bien qu'une première implémentation du modèle de jugement ait été implémentée en Answer Set Programming pour fournir une preuve de concept (Cointe *et al.*, 2016a), il était nécessaire d'implémenter ce modèle dans un véritable cadre multi-agent pour observer des comportements composés d'un grand nombre de décisions successives et confronter ce modèle de jugement au comportement d'autres agents.

Ces expérimentations montrent également comment une même implémentation de ce modèle peut être paramétrée de manières différentes, pour illustrer la généricité des fonctions de raisonnement par rapport au domaine applicatif, à la morale et à l'éthique confiées aux agents. Dans le cadre du développement de cette série d'expérimentations, la totalité des agents dotés d'un modèle de jugement partage une implémentation unique des fonctions à employer. Générer des agents dotés d'éthiques différentes nécessite uniquement de décrire les valeurs morales, règles morales, principes éthiques et préférences éthiques confiés aux agents. La conception modulaire de ces agents permet de faciliter l'implémentation d'une population d'agents dotés d'éthiques hétérogènes. Comme annoncé au chapitre II, cela permet de minimiser le temps nécessaire au déploiement de nouveaux types d'agents.

Nous voulions illustrer l'influence d'un tel modèle sur le comportement des agents. En ce sens, la section V.3 montre que les agents utilisant le modèle de jugement comme processus de décision adoptent effectivement un comportement qui est la conséquence de la conciliation à l'aide de principes éthiques de leurs désirs et des règles morales qui leur sont donnés.

Enfin nous souhaitions illustrer l'emploi du jugement du comportement des autres agents comme un moyen de distinguer les agents dotés d'éthiques similaires ou dissemblables à celle de l'agent juge. Les résultats présentés montrent que les images construites par le jugement progressif du comportement des autres est cohérent avec les éthiques qu'ils emploient dans leurs décisions.

Ces expérimentations pourraient être complétées sur différents aspects. Premièrement, les éthiques données en paramètres dans ces expérimentations ne diffèrent que sur les valeurs morales et règles morales. Il serait intéressant d'évaluer les différences de comportement dues à des ensembles différents de principes éthiques ordonnés de manière aléatoire. Deuxièmement, les agents partagent ici la même connaissance sur l'état du marché, mais pas la composition du portefeuille des agents qu'ils jugent. Le jugement évalué ici est donc toujours partiellement informé. Permettre aux agents de justifier leurs actes en transmettant aux agents juges les croyances dont ils disposaient lors de la prise de décision pourrait produire des résultats différents. De plus, les agents éthiques sont tous équipés d'une même fonction d'évaluation de la désirabilité. Enfin, implémenter des agents dotés d'appréciations différentes de la rentabilité d'un investissement permettrait de rendre plus hétérogène la population d'agents et poserait la question de la distinction de différences de comportements dues à des divergences éthiques ou à des divergences de désirs.

# Conclusion et perspectives

La problématique de la proposition d'un modèle de jugement éthique dans les systèmes multi-agents a été abordée au long de ce mémoire en partant de la définition des concepts employés et des travaux existants dans le domaine de la philosophie, source essentielle des théories liées à l'éthique. Nous avons ensuite présenté une proposition de modèle de jugement pour la décision puis pour la coopération et enfin une démarche expérimentale évaluant une implémentation de ce modèle. La première section apporte des éléments de réflexion personnelle sur cette démarche de recherche en intelligence artificielle. Puis la section suivante propose une vision synthétique des apports de ce travail avant d'en aborder en dernière section les limites et les perspectives.

## Prise de recul sur la démarche de recherche

Les avancées récentes et moins récentes dans le domaine de l'intelligence artificielle ont provoqué diverses réactions dans la communauté scientifique et, de manière plus large, auprès du grand public. Ces réactions semblent dues à la fois à une prise de conscience de la part du public des possibilités de reproduction par des systèmes autonomes de facultés mentales que les individus pensaient être une spécificité humaine, et en même temps au déploiement rapide et visible de systèmes autonomes dans leur environnement (citons par exemple le cas des véhicules autonomes, des drones civils et militaires ou encore des chatbots présents sur internet).

La prise de conscience de la possibilité de doter des agents autonomes de facultés mentales reproduisant des comportements et raisonnements humains mènent certains penseurs à désigner la thèse de Church-Turing (Turing, 1937) comme la quatrième blessure narcissique de l'humanité (Bouzou, 2016) après les trois premières énoncées par (Freud, 1961) : la révolution copernicienne a brisé l'illusion d'une Humanité placée au centre de l'univers ; la théorie de l'évolution a replacé l'Homme dans le règne animal ; psychanalyse a montré que les êtres humains n'ont pas le plein contrôle de leur esprit. L'idée que toute série d'opérations réalisable par un humain puisse être effectuée par une machine de Turing vient encore dégrader cette image d'une condition humaine privilégiée. L'actualité est marquée par des réactions variées face à cet état de fait : pour les uns, la panique ou la peur d'une domination ou d'un remplacement des êtres humains les incite à appeler à un arrêt des recherches dans le domaine. Pour d'autres, une eupho-

rie et une surinterprétation des résultats existants les incitent à imaginer dans un futur plus ou moins proche des entités toutes puissantes capable de les « guérir de la maladie de la mort » [6]. Nous avons également mentionné en section I.2.3 des travaux montrant que la théorie de l'esprit pousse l'humain à prêter à des objets ou des représentations, et à plus forte raison ceux qui font preuve de capacité à effectuer des raisonnements et avoir des comportements proactifs, des processus mentaux pourtant inexistants. L'un des véritables risques introduits par les agents autonomes serait une confiance excessive et indue des utilisateurs envers des systèmes qu'ils croient conscients et infaillibles, pouvant amener à leur déléguer des responsabilités ou à faire aveuglement confiance à leurs décisions sans comprendre les limites de leur fonctionnement (O'Neil, 2017).

La responsabilité du chercheur, producteur de connaissance et de résultats expérimentaux semble être engagée dès lors que ses résultats peuvent servir l'argumentaire d'un camp ou de l'autre (*Éthique de la recherche en robotique*, 2014). Il est de son devoir de mentionner dans son discours les limites de ses travaux, et de fournir un maximum d'informations aux décideurs et législateurs qui devront évaluer l'impact potentiel de ses apports sur la société. En omettant de telles limites, voire en provoquant volontairement des espoirs irréalistes chez le grand public, les auteurs exposent l'ensemble de la communauté à un risque de perte de crédibilité et de confiance pouvant être fatale pour nos financements et nos institutions (Hendler, 2008).

Enfin la conception d'agents autonomes, embarqués ou non, et destinés à être introduits dans un environnement partagé avec des humains, pose la nécessaire question de leur responsabilité et de l'éthique de leur comportement. Notre travail, réalisé dans le but d'apporter des éléments de réponse nouveaux à cette problématique, a été pensé avec la vision d'une humanité hétérogène dans laquelle les convictions de l'utilisateur et de tout être humain en interaction avec les agents autonomes doit être prise en compte pour ne pas imposer le seul point de vue du concepteur ou propriétaire du système.

# Synthèse globale

Ce document est structuré en trois parties principales. La première a pour but de définir les composants du jugement et de définir les caractéristiques d'un modèle de jugement pour la coopération des agents autonomes artificiels. La seconde partie propose un modèle de jugement pour la décision et la coopération entre agents fondée sur le jugement de comportements. La troisième partie illustre la mise en œuvre de ce modèle et une démarche expérimentale permettant d'évaluer le fonctionnement de ce modèle.

## Définition des composants et caractéristiques du jugement

Le début de ce document a permis de poser des définitions des constituants du jugement, montrer comment ils interagissent et en déduire un ensemble de propriétés attendues d'une modélisation de ce jugement pour la prise de décision et le jugement des autres par des agents autonomes artificiels. Divers travaux de la littérature ont été

6. Type de propos tenus sur https://iatranshumanisme.com par exemple.

comparés à l'aide de ces critères. Notre conclusion est qu'à ce jour, à notre connaissance, aucun modèle existant ne permet de les satisfaire entièrement.

Dans le premier chapitre, nous avons rassemblé et articulé entre elles des définitions de concepts de morale et d'éthique afin de présenter les termes employés dans la suite de ce mémoire et montrer comment les diverses doctrines philosophiques s'expriment à l'aide de ces éléments. Nous y avons introduit la distinction entre la morale, ou théorie du bien, et l'éthique, ou théorie du juste. La première permet de décrire à l'aide de valeurs morales et de règles morales le caractère bon ou mauvais d'une action en fonction de son contexte. Cette distinction nous a permis d'introduire l'éthique comme étant un niveau de réflexion supplémentaire permettant de déterminer l'action juste par le jugement de l'ensemble des actions connues dans une situation à partir de principes éthiques. L'éthique est ainsi abordée comme une solution au risque d'indécision qu'entraînent les dilemmes moraux : ces situations exceptionnelles se produisent lorsque les actions à disposition sont toutes qualifiées de mauvaises par au moins une règle morale. Ce chapitre se poursuit par une brève présentation des apports de la littérature en neurologie à la compréhension du fonctionnement du jugement chez l'Homme. Nous y avons montré l'intervention de diverses aires cérébrales impliquées, entre autres, dans des processus émotionnels et dans la capacité à se représenter et reproduire les états mentaux d'autrui (cette faculté est appelée *théorie de l'esprit*). Le rôle des émotions semble essentiel dans les premiers instants du jugement grâce à la rapidité de la formulation d'une réponse. Le raisonnement vient, dans de nombreux modèles, confirmer ou réviser ce premier jugement dans un second temps à l'aune des connaissances de l'individu. La fin de ce chapitre aborde des travaux de psychologie et de sciences sociales présentant des modèles permettant de représenter l'hétérogénéité des théories du bien et du juste au sein d'une population en montrant que les raisonnements éthiques des individus reposent sur des principes divers (recherche de l'approbation des proches, conformité à des normes sociales, utilisation de principes considérés comme universels, etc.).

Muni de cette première analyse de l'éthique nous avons poursuivi notre analyse de l'état de l'art en nous intéressant à notre domaine de recherche Intelligence Artificielle et Systèmes multi-agents pour identifier les approches existantes. Nous avons choisi de comparer ces travaux vis-à-vis d'un ensemble de critères que sont la représentation explicite ou implicite de l'éthique employée, l'approche rationaliste ou intuitionniste, la généricité du modèle vis-à-vis du domaine applicatif, de la morale et de l'éthique et le caractère opérationnel de la proposition. La présentation de ces modèles permet de proposer en synthèse de ce chapitre une grille de lecture mettant en évidence l'absence, en l'état, de proposition correspondant à l'ensemble des caractéristiques pertinentes. Nous justifions la nécessité d'une représentation explicite par le besoin de pouvoir formuler l'éthique de l'agent afin de la communiquer à d'autres agents, artificiels ou humains. Notre décision d'exclure toute simulation émotionnelle du jugement s'appuie sur des études montrant que les humains préfèrent, dans le cadre de la délégation de décisions à des agents, que ceux-ci adoptent un comportement rationnel. La recherche d'un modèle générique est motivée par le coût élevé du développement de nouvelles architectures

d'agents *ad hoc* à chaque changement dans le domaine applicatif et plus encore lors du déploiement d'agents dotés de morales et d'éthiques différentes. Enfin le caractère opérationnel du modèle est une nécessité pour pouvoir expérimenter le déploiement de tels agents et évaluer le bon fonctionnement de ce modèle dans des situations réalistes.

## Proposition d'un modèle de jugement

À partir de cette analyse de l'état de l'art, nous avons proposé un modèle permettant de juger de l'action juste dans une situation, tant pour permettre à l'agent lui-même de décider des actions à réaliser que pour évaluer le comportement d'un autre.

Le troisième chapitre détaille une première partie de notre proposition qu'est le modèle de jugement. Ce modèle emploie la représentation des actions, croyances et désirs de l'agent, puis une description des modèles de reconnaissance de situation et d'évaluation de la désirabilité et de la possibilité des actions. Nous introduisons les modèles d'évaluation de la moralité et de l'éthique des actions, spécifiques de notre modèle. L'évaluation de la moralité s'appuie sur une représentation de supports de valeurs et de règles morales permettant de produire un ensemble d'actions moralement évaluées, c'est-à-dire auxquelles est associée une valuation de moralité en fonction du contexte. Nous montrons à travers un ensemble d'exemples qu'il est possible de représenter à l'aide de ce formalisme plusieurs approches philosophiques présentées au premier chapitre, ce qui nous permet d'affirmer que, dans la limite de l'ensemble des approches morales exprimables de cette manière, le modèle de l'évaluation de la moralité est générique par rapport aux supports de valeurs et règles morales employés pour évaluer les actions. De même, le modèle d'évaluation de l'éthique des actions utilise des ensembles de représentations de principes et de préférences éthiques afin de représenter et employer des théories variées pour évaluer les actions justes au regard de l'ensemble ordonné des principes éthiques de l'agent et des ensembles d'actions évaluées pour leur caractère possible, désirable et moral. L'ensemble de ce chapitre est illustré par des exemples volontairement simples afin de montrer le déroulement, étape par étape, de ce jugement.

La deuxième partie de notre modèle étend cette proposition afin de permettre aux agents d'employer leur modèle de jugement non seulement comme un processus décisionnel, mais également comme un moyen d'évaluer la conformité du comportement des autres agents à un ensemble de connaissances de la situation, de connaissances du bien et de connaissances du juste. Ces ensembles de connaissances, lorsqu'ils sont enrichis ou remplacés par des informations acquises sur l'agent jugé ou toute autre connaissance que celles propres à l'agent juge, permettent de produire ce que nous appelons un jugement partiellement ou totalement informé. De plus le jugement de comportements, définis comme des successions d'actions, nécessite la définition de mécanismes d'agrégation des jugements ponctuels de ces actions afin de construire une image du comportement, c'est-à-dire une croyance en un certain niveau de conformité de ce comportement à un ensemble de critères. Ces images permettent ensuite à l'agent de décider d'accorder ou non sa confiance à un autre agent en ayant connaissance de ce comportement. Enfin,

l'action de construction de cette confiance peut faire l'objet de la définition de nouvelles valeurs et règles morales permettant de définir la morale et l'éthique de la coopération afin de permettre aux agents d'intégrer une dimension sociale de leurs décisions à leur jugement.

## Évaluations du modèle

Ce travail de recherche s'achève sur des travaux d'expérimentation permettant d'illustrer la mise en œuvre du modèle avec quelques concessions et simplifications, dans un cadre BDI. Les résultats expérimentaux montrent que le comportement des agents et leur évaluation par les autres agents sont cohérents avec les connaissances du bien et du juste données en paramètres.

Le cinquième chapitre propose un ensemble d'expérimentations de déploiement d'agents autonomes éthiques dans un domaine applicatif de gestion d'actifs financiers afin d'évaluer l'utilisation du jugement comme processus de décision et la qualité des images produites par le jugement des autres agents autonomes. Nous détaillons tout d'abord les problématiques éthiques liées au domaine de la gestion d'actifs en présentant à la fois le fonctionnement des marchés et la dimension éthique des décisions. N'étant pas expert de l'éthique financière, nous présentons des documents proposant des ensembles de valeurs morales et de règles morales diverses qui, une fois exprimées dans un langage manipulable par les agents, nous permettent de paramétrer des ensembles d'agents avec des morales diverses. Nous présentons ensuite une implémentation de cette expérimentation à l'aide de la plateforme JaCaMo dédiée à la conception de systèmes multi-agents BDI dans le langage déclaratif Jason. Enfin nous examinons les résultats produits en comparant le comportement d'agents dotés de connaissances du bien différentes afin de mettre en évidence l'influence du modèle de jugement sur le processus de décision des agents. Ensuite nous évaluons la qualité de la construction des images en montrant comment elles permettent de discriminer les agents conformes à des ensembles de règles morales ou à des éthiques.

## Limites et perspectives

Ce travail s'appuie sur les définitions présentées au premier chapitre et qui sont, par nécessité, réductrices. Bien que soucieux de définir les composants du jugement en faisant abstraction des convictions qu'il emploie, il est probable que certaines doctrines moins classiques que celles présentées au premier chapitre ne puissent pas être représentées par ce modèle. De plus, nous avons considéré l'éthique et la morale de l'agent comme des connaissances sur lesquelles repose le raisonnement sans plus de précisions sur les mécanismes ou méthodologies permettant de produire ces connaissances. Il serait intéressant d'explorer des pistes telles que l'apprentissage de la morale par observation du comportement des autres, ou bien par construction lors d'un dialogue avec l'utilisateur. Il serait également intéressant d'étudier les propriétés d'une morale et d'une éthique plus dynamique, qui serait ajustée par l'observation des conséquences des actions de

l'agent, tant sur l'environnement que par les réactions des autres agents (amélioration ou dégradation de l'image de l'agent chez les autres).

De manière générale, de nombreuses questions sont maintenant posées par la dimension collective de l'éthique. De nombreuses problématiques ont été soulevées et structurées pour guider la suite de nos réflexions (Cointe *et al.*, 2015 ; 2017). Notre proposition de modèle pour la coopération est ici focalisée sur les effets de ce jugement et de sa capacité à s'en servir pour coopérer d'un point de vue individuel. Il serait intéressant d'étudier la stabilité et les caractéristiques des ensembles d'agents formés par des relations de confiance mutuelle. Serait-il possible d'identifier une forme d'éthique collective ? L'appartenance à un collectif inciterait-elle l'agent à se conformer à des valeurs morales, règles morales et principes éthiques qui ne sont pas les siens pour ne pas risquer son exclusion ? En quoi l'appartenance à de tels collectifs modifierait les rapports entre les agents membres de celui-ci et les autres agents ? Quelles relations peuvent exister entre différents collectifs ? Ces collectifs sont-ils résistants à d'importants changements de situations ?

La démarche expérimentale pourrait également être étendue à des aspects inexplorés jusqu'ici : les agents gestionnaires implémentés n'échangent pas encore d'information permettant d'informer le jugement des autres. De plus, même si nous avons observé le bon fonctionnement de la construction des images et de la confiance, aucun mécanisme rendant utile la coopération n'a été implémenté dans ces expérimentations. Il serait intéressant de pouvoir visualiser l'effet de cette coopération sur le comportement des agents et l'évolution du système. Enfin, il serait intéressant d'employer les résultats de ces jugements pour construire des collectifs d'agents et exploiter des structures pour faire de la coopération. De nombreuses questions seraient ensuite soulevées sur la stabilité de ces collectifs face à de nouvelles situations, la réaction des agents face à un changement de comportement d'un membre du collectif ou encore la possibilité pour les agents d'interagir pour identifier les concepts partagés avec le reste du collectif.

# Bibliographie


Abdul-Rahman A., Hailes S. (2000). Supporting trust in virtual communities. In *33th IEEE international conference on systems sciences*, p. 1-9.

Ahmed O. (2001). Islamic equity funds : The mode of resource mobilization and placement. *Islamic Development Bank.*

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.

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

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

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

Anderson M., Anderson S. L. (2014a). Geneth : A general ethical dilemma analyzer. In *AAAI*, p. 253–261.

Anderson M., Anderson S. L., Berenz V. (2017). A value driven agent : Instantiation of a case-supported principle-based behavior paradigm. In *Workshops at the thirty-first AAAI conference on artificial intelligence.*

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.

Atkinson K., Bench-Capon T. (2006). Addressing moral problems through practical reasoning. In *International workshop on deontic logic and artificial normative systems*, p. 8–23.

Baertschi B. (2011). Neurosciences et éthique : que nous apprend le dilemme du wagon fou ? *Igitur : arguments philosophiques*, vol. 3, n° 3, p. 1–17.

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

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

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.

Berreby F., Bourgne G., Ganascia J.-G. (2017a). Cadre déclaratif modulaire pour représenter et appliquer des principes éthiques. In *Journées d'intelligence artificielle fondamentale.*

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

Blair R. J. R. (1995). A cognitive developmental approach to morality : Investigating the psychopath. *Cognition*, vol. 57, n° 1, p. 1–29.

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

Boella G., Torre L. van der. (2003). BDI and BOID argumentation : Some examples and ideas for formalization. *Procs. of IJCAI-Computational Models of Natural Argument.*

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

Boissier O., Hübner J. F., Ricci A. (2016). The JaCaMo framework. In *Social coordination frameworks for social technical systems*, p. 125–151. Springer.

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

Bordini R. H., Hübner J. F., Wooldridge M. (2007). *Programming multi-agent systems in AgentSpeak using Jason* (vol. 8). John Wiley & Sons.

Bouzou N. (2016). *L'innovation sauvera le monde.* Plon.

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

Bratman M. (1987). *Intention, plans, and practical reason.* Harvard University Press, Massachusetts.

Bringsjord S., Arkoudas K., Bello P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, vol. 21, n° 4, p. 38–44.

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

Bringsjord S., Taylor J. (2012). The divine-command approach to robot ethics. In, p. 85–108. MIT Press Cambridge, MA.

Bzdok D., Schilbach L., Vogeley K., Schneider K., Laird A. R., Langner R. *et al.* (2012). Parsing the neural correlates of moral cognition : ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, vol. 217, n° 4, p. 783–796.

Canto-Sperber M., Ogien R. (2004). *Dictionnaire d'éthique et de philosophie morale.* PUF.

Carbo J., Molina J., Davila J. (2002). Comparing predictions of SPORAS vs. a fuzzy reputation agent system. In *3rd international joint conference on fuzzy sets and fuzzy systems*, p. 147-153.

Carter J., Bitting E., Ghorbani A. (2002). Reputation formalization for an information-sharing multi-agent system. *Computational Intelligence*, vol. 18, n° 2, p. 515-534.

Castelfranchi C. (2000). Artificial liars : Why computers will (necessarily) deceive us and each other. *Ethics and Information Technology*, vol. 2, n° 2, p. 113–119.

Castelfranchi C., Falcone R. (1998). Principles of trust for MAS : Cognitive anatomy, social importance, and quantification. In *Multi agent systems, 1998. proceedings. international conference on*, p. 72–79.

Castelfranchi C., Falcone R. (2010). *Trust theory : A socio-cognitive and computational model* (vol. 18). John Wiley & Sons.

Chatila R., Firth-Butterflied K., Havens J. C., Karachalios K. (2017). The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. *IEEE Robotics & Automation Magazine*, vol. 24, n° 1, p. 110–110.

Chellas B. F. (1980). *Modal logic : an introduction.* Cambridge university press.

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

Coelho H., Rocha Costa A. C. da, Trigo P. (2010). Decision making for agent moral conducts. In *Proceedings of the inforum*, p. 9–10.

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

Cointe N., Bonnet G., Boissier O. (2015). De l'intérêt de l'éthique collective pour les systèmes multi-agents. In *Plate-forme intelligence artificielle 2015.*

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.

Cointe N., Bonnet G., Boissier O. (2017). Éthique collective dans les systèmes multi-agents. *Revue d'Intelligence Artificielle*, n° vol. 31, no 1-2, p. 71–96.

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

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

Constant B. (2013). *De la force du gouvernement actuel de la France et de la nécessité de s'y rallier : Des réactions politiques. des effets de la terreur.* Flammarion.

Constant B., Kant E. (2003). Le droit de mentir. *Paris, Mille et une nuits.*

Conte R., Paolucci M. (2002). *Reputation in artificial societies : Social beliefs for social order* (vol. 6). Springer Science & Business Media.

Cova F. (2014). *L'architecture de la cognition morale.* Thèse de doctorat non publiée, Atelier national de reproduction des thèses.

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

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

Descartes R. (1960). 1637. Discours de la méthode. *Œuvres de Descartes*, vol. 6, p. 1964–1974.

Dien D. S.-F. (1982). A Chinese perspective on Kohlberg's theory of moral development. *Developmental Review*, vol. 2, n° 4, p. 331–341.

Dogan E., Chatila R., Chauvier S., Evans K., Hadjixenophontos P., Perrin J. (2016). Ethics in the design of automated vehicles : The AVEthics project. In *1st workshop on ethics in the design of intelligent agents*, p. 10–13.

Downie R. (1980). Ethics, morals and moral philosophy. *Journal of medical ethics*, vol. 6, n° 1, p. 33.

Driessen T. (1991). A survey of consistency properties in cooperative game theory. *SIAM Review*, vol. 33, n° 1, p. 43-59.

Economic and Financial Affairs D.-G. for. (2009, June). *Impact of the current economic and financial crisis on potential output.* Occasional Papers n° 49. European Commission.

Esfandiari B., Chandrasekharan S. (2001). On how agents make friends : Mechanisms for trust acquisition. In *4th workshop on deception, fraud, and trust in agent societies*, p. 27-34.

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

Farrow T. F., Zheng Y., Wilkinson I. D., Spence S. A., Deakin J. W., Tarrier N. *et al.* (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*, vol. 12, n° 11, p. 2433–2438.

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

Fieser J. (2015). Ethics. *The Internet Encyclopedia of Philosophy.*

Fikes R. E., Nilsson N. J. (1971). Strips : A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, vol. 2, n° 3-4, p. 189–208.

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

Forte G., Miglietta F. (2011). A comparison of socially responsible and islamic equity investments. *Journal of Money, Investment and Banking*, vol. 21, p. 116–132.

Foster M. E., Deshmukh A., Janarthanam S., Lim M. Y., Hastie H., Aylett R. (2015). Influencing the learning experience through affective agent feedback in a real-world treasure hunt. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems*, p. 1711–1712.

Freud S. (1961). Introduction à la psychanalyse (1916) trad. *S. Jankélévitch, Payot*, p. 151.

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

Friedman B., Hendry D. G., Huldtgren A., Jonker C., Hoven J., Wynsberghe A. (2015). Charting the next decade for value sensitive design. *Aarhus series on human centered computing*, vol. 1, n° 1.

Friedman B., Kahn P., Borning A. (2002). Value sensitive design : Theory and methods. *University of Washington technical report*, p. 02–12.

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

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.

Gendron C. (2005). *Les codes d'éthique : de la déontologie à la responsabilité sociale*. Chaire de responsabilité sociale et de développement durable, ESG, UQAM.

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

Ghallab M., Howe A., Knoblock C., McDermott D., Ram A., Veloso M. *et al.* (1998). PDDL the planning domain definition language.

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

Gould M. D., Porter M. A., Williams S., McDonald M., Fenn D. J., Howison S. D. (2013). Limit order books. *Quantitative Finance*, vol. 13, n° 11, p. 1709–1742.

Gowans C. (2016). Moral relativism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Winter 2016 éd.. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/moral-relativism/.

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

Guéranger F. (2009). *Finance islamique : une illustration de la finance éthique*. Dunod.

Haidt J. (2001). The emotional dog and its rational tail : a social intuitionist approach to moral judgment. *Psychological review*, vol. 108, n° 4, p. 814.

Haidt J., Björklund F., Murphy S. (2000). Moral dumbfounding : When intuition finds no reason. *Unpublished manuscript, University of Virginia.*.

Hamilton S., Jo H., Statman M. (1993). Doing well while doing good ? The investment performance of socially responsible mutual funds. *Financial Analysts Journal*, vol. 49, n° 6, p. 62–66.

Harenski C. L., Hamann S. (2006). Neural correlates of regulating negative emotions related to moral violations. *Neuroimage*, vol. 30, n° 1, p. 313–324.

Hendler J. (2008). *Avoiding another AI winter [open letter]*. IEEE Computer Society.

Herzig A., Lorini E., Hübner J. F., Vercouter L. (2009). A logic of trust and reputation. *Logic Journal of IGPL*, vol. 18, n° 1, p. 214–244.

Hill T. E. (2010). How clinicians make (or avoid) moral judgments of patients : implications of the evidence for relationships and research. *Philosophy, Ethics, and Humanities in Medicine*, vol. 5, n° 1, p. 11.

Hofstede G., Bond M. H. (1984). Hofstede's culture dimensions : An independent validation using rokeach's value survey. *Journal of cross-cultural psychology*, vol. 15, n° 4, p. 417–433.

Horsburgh H. (1960). The ethics of trust. *The Philosophical Quartely*, vol. 10, n° 41, p. 343-354.

Hubner J. F., Sichman J. S., Boissier O. (2007). Developing organised multiagent systems using the MOISE+ model : programming issues at the system and agent levels. *International Journal of Agent-Oriented Software Engineering*, vol. 1, n° 3-4, p. 370–395.

Hume D. (2006). *An enquiry concerning the principles of morals.* Oxford University Press.

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

Imm Ng S., Anne Lee J., Soutar G. N. (2007). Are Hofstede's and Schwartz's value frameworks congruent ? *International marketing review*, vol. 24, n° 2, p. 164–180.

Inglehart R., Welzel C. (2005). *Modernization, cultural change, and democracy : The human development sequence.* Cambridge University Press.

Jennings N. R., Sycara K., Wooldridge M. (1998). A roadmap of agent research and development. *Autonomous agents and multi-agent systems*, vol. 1, n° 1, p. 7–38.

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

Josang A., Ismail R. (2002). The beta reputation system. In *15th bled conference on electronic commerce.*

Josang A., Ismail R., Boyd C. (2007). A survey of trust and reputation systems for online service proposition. *Decision Support Systems*, vol. 43, n° 2, p. 618-644.

Kang K. I., Freedman S., Mataric M. J., Cunningham M. J., Lopez B. (2005). A hands-off physical therapy assistance robot for cardiac patients. In *Rehabilitation robotics, 2005. icorr 2005. 9th international conference on*, p. 337–340.

Kant I. (1785). Grundlegung zur Metaphysik der Sitten, Werkausgabe bd. *VII, Frankfurt/M. : Suhrkamp.*

Kauppinen A. (2017). Moral sentimentalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Spring 2017 éd.. Metaphysics Research Lab, Stanford University. https://plato.stanford .edu/archives/spr2017/entries/moral-sentimentalism/.

Kempf A., Osthoff P. (2007). The effect of socially responsible investing on portfolio performance. *European Financial Management*, vol. 13, n° 5, p. 908–922.

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.

Korsgaard C. M. (1996). *The sources of normativity.* Cambridge University Press.

La Rochefoucauld F. duc de, Vauvenargues L. d. C. de. (1867). *Réflexions ou sentences et maximes morales.* Garnier.

Larroque S. (2014). Simulation des raisonnements éthiques par logiques non-monotones. In *Rencontre des Jeunes Chercheurs en Intelligence Artificielle.*

Lesser V., Gasser L. (Eds.). (1995). *Proceedings of the first international conference on multiagent systems, San Francisco, California, USA.* The MIT Press.

Leung J. M.-J., Chong T. T.-L. (2003). An empirical comparison of moving average envelopes and bollinger bands. *Applied Economics Letters*, vol. 10, n° 6, p. 339–341.

Lin P., Abney K., Bekey G. (2011). *Robot ethics : the ethical and social implications of robotics.* MIT press.

Lorini E. (2012). On the logical foundations of moral agency. In *11th international conference on deontic logic in computer science*, p. 108-122.

Malle B. F., Scheutz M., Arnold T., Voiklis J., Cusimano C. (2015). Sacrifice one for the good of many ? people apply different moral norms to human and robot agents. In *Proceedings of the tenth annual international conference on human-robot interaction*, p. 117–124.

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.

Marsh S. (1994). *Formalising trust as a computational concept.* Thèse de doctorat non publiée, University of Stirling, Stirling.

Marti P., Pollini A., Rullo A., Shibata T. (2005). Engaging with artificial pets. In *Proceedings of the 2005 annual conference on European association of cognitive ergonomics*, p. 99–106.

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.

McNamara P. (2014). Deontic logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Winter 2014 éd.. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2014/entries/logic-deontic/.

Moll J., Oliveira-Souza R. de, Eslinger P. J., Bramati I. E., Mourao-Miranda J., Andreiuolo P. A. *et al.* (2002). The neural correlates of moral sensitivity : a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of neuroscience*, vol. 22, n° 7, p. 2730–2736.

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

Muller G., Vercouter L., Boissier O. (2003). Towards a general definition of trust and its application to openness in MAS. In *6th workshop on deception, fraud and trust in agent societies*, p. 49-56.

Murakami Y. (2004). Utilitarian deontic logic. *AiML-2004 : Advances in Modal Logic*, vol. 287, p. 287–302.

Nietzsche F. (2015). *Ainsi parlait Zarathoustra : Nouvelle édition augmentée.* Arvensa Editions.

Norling E. (2003). Capturing the quake player : using a BDI agent to model human behaviour. In *Proceedings of the second international joint conference on autonomous agents and multiagent systems*, p. 1080–1081.

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

O'Neil C. (2017). *Weapons of math destruction : How big data increases inequality and threatens democracy.* Broadway Books.

Parkinson C., Sinnott-Armstrong W., Koralus P. E., Mendelovici A., McGeer V., Wheatley T. (2011). Is morality unified ? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, vol. 23, n° 10, p. 3162–3180.

Partiot A., Grafman J., Sadato N., Wachs J., Hallett M. (1995). Brain activation during the generation of nonemotional and emotional plans. *Neuroreport*, vol. 6, n° 10, p. 1397–1400.

Patil I., Cogoni C., Zangrando N., Chittaro L., Silani G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social neuroscience*, vol. 9, n° 1, p. 94–107.

Pereira L. M., Saptawijaya A. (2017). Counterfactuals, logic programming and agent morality. In *Applications of formal philosophy*, p. 25–53. Springer.

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.

Premack D., Woodruff G. (1978). Does the chimpanzee have a theory of mind ? *Behavioral and brain sciences*, vol. 1, n° 4, p. 515–526.

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

Rai T. S., Holyoak K. J. (2013). Exposure to moral relativism compromises moral behavior. *Journal of Experimental Social Psychology*, vol. 49, n° 6, p. 995–1001.

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

Rao A. S. (1996). Agentspeak (l) : BDI agents speak out in a logical computable language. In *European workshop on modelling autonomous agents in a multi-agent world*, p. 42–55.

Rao A. S., Georgeff M. P. (1991). Modeling rational agents within a BDI-architecture. , p. 473-484.

Ricci A., Piunti M., Viroli M., Omicini A. (2009). Environment programming in CArtAgO. *Multi-agent programming : Languages, platforms and applications*, vol. 2, p. 259–288.

Richardson H. S., Williams M. S. (2009). *Moral universalism and pluralism* (vol. 49). NYU Press.

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

Rocha Costa A. da. (2016). Moral systems of agent societies : Some elements for their analysis and design. , p. 34–39.

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

Ross W. (1930). *The right and the good*. Oxford University Press.

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/)

Rutherford D. (2017). Descartes' ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Fall 2017 éd.. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/descartes-ethics/.

Ruvinsky A. I. (2007). Computational ethics. In *Encyclopedia of information ethics and security*, p. 76–82. IGI Global.

Sabater J., Sierra C. (2001). REGRET : A reputation model for gregarious societies. In *4th workshop on deception, fraud, and trust in agent societies*, p. 61-69.

Sabater J., Sierra C. (2005). Review on computational trust and reputation models. *Artificial Intelligence*, vol. 24, n° 1, p. 33-60.

Sabater-Mir J., Vercouter L. (2013). Trust and reputation in multiagent systems. *Multiagent Systems*, p. 381.

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.

Saptawijaya A., Pereira L. M. (2016). Logic programming for modeling morality. *Logic Journal of the IGPL*, vol. 24, n° 4, p. 510–525.

Saxe R., Kanwisher N. (2003). People thinking about thinking people : the role of the temporo-parietal junction in “theory of mind”. *Neuroimage*, vol. 19, n° 4, p. 1835–1842.

Schmeidler D. (1969). The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, vol. 17, n° 6, p. 1163-1170.

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

Schwartz S. 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. (1994). *Beyond individualism/collectivism : New cultural dimensions of values.* Sage Publications, Inc.

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.

Sen S., Sajja N. (2002). Robustness of reputation-based trust : Boolean case. In *1st international joint conference on autonomous agents and multi-agent systems*, p. 288-293.

Shanahan M. (1999). The event calculus explained. In *Artificial intelligence today*, p. 409–430. Springer.

Shapley L. (1953). A value for n-person games. In H. Kuhn, A. Tucker (Eds.), *Contributions to the theory of games ii*, p. 307-317. Princeton University Press.

Smith A. (1937). The wealth of nations [1776].

Snarey J., Kohlberg L., Noam G. (1983). Ego development in perspective : Structural stage, functional phase, and cultural age-period models. *Developmental Review*, vol. 3, n° 3, p. 303–338.

Snarey J. R. (1985). Cross-cultural universality of social-moral development : a critical review of Kohlbergian research. *Psychological bulletin*, vol. 97, n° 2, p. 202.

Spinoza B. (1677). *L’éthique.*

Sponville A. C. (2012). *La philosophie.* PUF.

Thoreau H. D. (2016). *Civil disobedience.* Broadview Press.

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

Tisseron S. (2015). *Le jour où mon robot m’aimera : vers l’empathie artificielle.* Albin Michel.

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.

Tufis M., Ganascia J.-G. (2012). Normative rational agents-A BDI approach. *Autonomous Agents (RDA2) 2012*, p. 38.

Tuominen A. *et al.* (2012). The role of high frequency trading in limit order book activity : Evidence from Helsinki stock exchange.

Turing A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*, vol. 2, n° 1, p. 230–265.

Van Dimarsch H., Herzig A., De Lima T. (2007). Raisonnement sur les actions : de Toronto à Amsterdam.

Vercouter L., Muller G. (2010). L.I.A.R. : Achieving social control in open and decentralized multiagent systems. *Applied Artificial Intelligence*, vol. 24, n° 8, p. 723-768.

Vincent N. (2011). A structured taxonomy of responsibility concepts. In *Moral responsibility*, p. 15–35. Springer.

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

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

Yamamoto M., Hagiwara M. (2014). Moral judgment system using evaluation expressions. In *Joint 7th international conference on soft computing and intelligent systems (SCIS) and 15th international symposium on advanced intelligent systems (ISIS), kitakyushu 2014*, p. 1040–1047.

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.

Young L., Dungan J. (2012). Where in the brain is morality ? Everywhere and maybe nowhere. *Social neuroscience*, vol. 7, n° 1, p. 1–10.

Yu B., Singh M. (2002). Distributed reputation management for electronic commerce. *Computational Intelligence*, vol. 18, n° 4, p. 535-549.

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

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2017LYSEM043

Nicolas COINTE

ETHICAL JUDGMENT FOR DECISION AND COOPERATION IN MULTIAGENT SYSTEMS

Speciality : Computer Science

Keywords : Computational ethics, multiagent systems, judgment, cooperation

Abstract :

The increasing use of multiagent systems in various fields raises the need of autonomous agents able to take into account such ethical principles in their decisions. More and more propositions are published, but they are often agent-centered and they don't consider the issues raised by the interactions between artificial agents and possibly humans, potentially using another ethics. Our goal is to give the agents the ability to reason on ethics to enable an ethics-based cooperation in multiagent systems. This work presents a model of ethical judgment for artificial autonomous agents in multiagent systems both useful to influence their decisions and behaviors, and describes an ethics-based cooperation framework. This model distinguishes the morality (or theory of the good), describing the goodness of actions in a context regarding a set of moral values and moral rules, and ethics (or theory of the right), describing the rightness of an action regarding a set of ethical principles. The use of this model in the decision process generates a conform behavior regarding the chosen theories of good and right. An agent may also use this model to judge the observed behavior of the other agents and employ this judgment to adapt its own behavior towards the judged agents. The detailed presentation of this model is followed by some experimentations to show the use of this model in a realistic application based on an ethical asset management scenario. The results show how the behaviors of the agents might be impacted and the efficiency of this model to discriminate the behaviors of the others.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2017LYSEM043

Nicolas COINTE

# JUGEMENT ÉTHIQUE POUR LA DÉCISION ET LA COOPÉRATION DANS LES SYSTÈMES MULTI-AGENTS

Spécialité : Informatique

Mots clefs : Éthique computationnelle, systèmes multi-agents, jugement, coopération

Résumé :

L'usage croissant des systèmes multi-agents dans divers domaines d'application soulève la nécessité de concevoir des agents capables de prendre des décisions s'appuyant sur des principes éthiques. De plus en plus de travaux proposent de telles approches. Toutefois, ces systèmes considèrent principalement une perspective centrée sur l'agent et mettent de côté le fait que ces agents sont en interaction avec d'autres agents, artificiels ou humains, qui utilisent d'autres concepts éthiques. Notre objectif est d'équiper les agents de capacités de raisonnement éthique pour permettre la mise en place de coopérations fondées sur l'éthique dans un SMA. Ce travail propose un modèle de jugement éthique pour les agents autonomes artificiels dans les systèmes multi-agents permettant de guider leurs décisions afin d'influencer leur comportement individuel d'une part, et de décrire un cadre de coopération fondée sur l'éthique d'autre part. Les éléments de ce modèle reposent sur une distinction entre la morale (ou théorie du bien), décrivant le caractère bon ou mauvais des actions d'un agent en faisant appel à la définition de valeurs morales et de règles morales, et l'éthique (ou théorie du juste), permettant de juger de l'action qu'il est juste d'effectuer dans une situation au regard d'un ensemble ordonné de principes éthiques et des actions moralement évaluées. L'agent, en employant ce modèle de jugement comme un processus décisionnel, adopte alors un comportement éthique du point de vue des théories du bien et du juste qui lui sont confiées. Il lui est également possible d'employer ce modèle pour juger le comportement des autres agents et tenir compte de ce jugement dans son propre comportement vis-à-vis des agents jugés. La présentation de ce modèle est accompagnée d'expérimentations illustrant l'utilisation de ce modèle dans un domaine applicatif réaliste de gestion éthique d'actifs financiers permettant d'éprouver l'influence du jugement sur le comportement des agents et la qualité de l'image obtenue par le jugement du comportement des autres agents dans diverses situations.