ÉTHIQUE ET AGENTS AUTONOMES

Juillet 2018

Avec le soutien de l'Agence Nationale de la Recherche (ANR-13-CORD-0006)

LE CONSORTIUM DU PROJET ETHICAA

Créée en 1999, Ardans (https://www.ardans.fr) s'est imposée en leader de l'ingénierie de la connaissance en France en proposant une offre de conseil, d'expertise IT & KM et d'éditeur d'outil dédié au Knowledge Management. Ardans est une société indépendante qui est basée à Montigny-le-Bretonneux (Paris-Saclay). Ardans valorise le savoir-faire de ses clients au sein de leur système d'information. Ardans est membre de l'AFIA, de EGC, de Cap Digital, d'ASTech, du Comité Richelieu, du RAVI et d'EFEL.

L'Institut Henri Fayol est un des centres communs de recherche ARMINES et École des Mines de Saint-Etienne, école de l'Institut Mines Télécom. Ce centre mène des recherches en informatique et systèmes intelligents, génie mathématique et industriel, génie de l'environnement et des organisations, management responsable et innovation. L'équipe impliquée dans le projet ETHICAA est l'équipe Informatique et Systèmes Intelligents (www.emse.fr) dont l'expertise porte sur les systèmes multi-agents normatifs et auto-organisés, les technologies du Web Sémantique, la confiance et la privacité.

Le GREYC est une UMR CNRS et un laboratoire de l'Université de Caen-Basse-Normandie (UCBN) et de l'ENSICAEN. Ses thèmes de recherche sont l'informatique, l'automatique et la micro-électronique. L'équipe impliquée dans le projet ETHICAA est l'équipe MAD (Modèles, agents et décision – https://www.greyc.fr/?page_id=439) qui travaille sur les systèmes intelligents et autonomes en utilisant en particulier des modèles de raisonnement, de planification et de régulation multi-agents.

L'Institut Mines-Télécom (IMT) est un établissement public dédié à l'enseignement supérieur et la recherche pour l'innovation dans les domaines de l'ingénierie et du numérique. Il regroupe treize écoles d'ingénieurs et de management. L'équipe impliquée dans le projet ETHICAA est l'équipe ETHOS (Éthique, Technologies, Humains, Organisations, Société – https://labetos.wp.imt.fr/) de l'Institut Mines-Télécom Business School (en collaboration avec IMT Atlantique) qui interroge les conséquences éthiques et sociétales de la transformation numérique.

Le LIP6 est une UMR CNRS affiliée à Sorbonne Université et, avec 550 membres, est un des plus importants laboratoires de recherche en informatique de France. L'équipe impliquée dans le projet ETHICAA est l'équipe ACASA (Agents Cognitifs et Apprentissage Symbolique Automatique – https://www.lip6.fr/recherche/team.php?id=560) qui possède une expertise en Intelligence Artificielle symbolique, représentation des connaissances et raisonnement automatisé.

L'ONERA (Office National d'Études et Recherches Aérospatiales) est le centre français de recherche aérospatiale. L'ONERA effectue des recherches sur les codes de calcul, méthodes, outils, technologies, matériaux et autres produits et services qui sont utilisés dans les programmes aérospatiaux et de défense. Le département impliqué dans le projet ETHICAA est le DTIS (Département de Traitement de l'Information et Systèmes) dont une des spécialités est la prise de décision et la robotique autonome.


AGENCE NATIONALE DE LA RECHERCHE
ANR


GREYC


ARMINES


MINES
Saint-Étienne


ONERA
THE FRENCH AEROSPACE LAB


ARDANS
Knowledge Consulting & Software Solutions


Institut Mines-Télécom
Business School


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom


LIP6

SOMMAIRE

Rédacteurs

Flavien Balbo, ARMINES/Mines Saint-Etienne

Fiona Berreby, LIP6

Olivier Boissier, ARMINES/Mines Saint-Etienne

Vincent Bonnemains, ONERA

Grégory Bonnet, GREYC

Gauvain Bourgne, LIP6

Pierre-Antoine Chardel, IMT

Jean-Gabriel Ganascia, LIP6

Bruno Mermet, GREYC

Gaële Simon, GREYC

Thibault de Swarte, IMT

Catherine Tessier, ONERA

Robert Voyer, IMT

Relecteurs additionnels

Alain Berger, Ardans

Nicolas Cointe, ARMINES/Mines Saint-Etienne

1.CONTEXTE ET PROBLÉMATIQUES

Les agents autonomes artificiels[1] sont des machines logicielles ou physiques capables de calculer des décisions, de manière individuelle, coordonnée ou non avec d'autres agents ou avec des humains, en vue de la réalisation de buts de haut niveau qui leur ont été spécifiés. Leur introduction dans des domaines tels que le domaine militaire (voir page 10), la justice (voir page 25), le milieu médical (voir page 36) ou encore les transports autonomes (voir page 42), soulève de nombreuses questions éthiques. En effet, les utilisateurs (ou futurs utilisateurs) de ces systèmes ont parfois des attentes éthiques distinctes des problématiques d'optimalité ou de conformité légale du comportement des agents.

L'éthique nous invite à intégrer le fait que ce qui est technologiquement possible n'est pas toujours humainement ou socialement souhaitable.

Nous tenons à rappeler qu'il n'y a jamais de technologie en soi, mais des contextes toujours différents qui nous invitent à préciser notre regard sur les questions éthiques que les technologies nous posent, avec leurs lots de paradoxes et de tensions. Les bienfaits de l'automatisation ne sauraient donc nous faire sous-estimer la complexité qui définit toute expérience humaine, à plus forte raison lorsque celle-ci entre en interaction avec des agents autonomes. Il importe, par conséquent, d'éviter tout réductionnisme technologique dans la mesure où il n'y a pas de lien de nécessité entre « progrès technologique » et « progrès social ». Ce sont plutôt des conditions d'appropriation, toujours contingentes et hétérogènes, qui permettent d'établir un lien entre ces deux formes de progrès. Or, l'éthique nous invite à intégrer le fait que ce qui est technologiquement possible n'est pas toujours humainement ou socialement souhaitable.

C'est pourquoi l'intérêt pour le comportement éthique des agents autonomes est apparu dans les travaux de recherche traitant des agents autonomes, comme en témoignent déjà, à leur façon, les rapports internationaux ou nationaux publiés à ce jour[2][3][4]. Mais, vis-à-vis d'un certain nombre de recherches suivant une telle tendance, l'enjeu fut pour nous de privilégier un geste ample de problématisation et de réflexivité critique.

À cet égard, nous avons clairement décidé de privilégier les questions

1 Par souci de lisibilité, dans toute la suite de ce document, le qualificatif « artificiel » sera implicite dans toute expression « agents autonomes ».
2 https://ethicsinaction.ieee.org/
3 http://cerna-ethics-allistene.org
4 https://www.cnil.fr

d'éthique et non de morale. Comme l'histoire des idées nous le rappelle, étymologiquement, les deux mots « morale » et « éthique » désignent la même chose, le premier est d'origine latine (mores) et le second d'origine grecque (ethos). L'usage, en France, a cependant introduit une différence entre les deux notions. Autant la morale est **prescriptive** (elle nous aiguille sur ce qui est jugé bon ou mal de faire en fonction de valeurs qui sont censées être partagées par le plus grand nombre), autant l'éthique est **réflexive** au sens où elle nous invite à engager une réflexion sur le sens de nos actions : « *La morale se présente comme un ensemble de règles contraignantes d'un type spécial, qui consiste à juger des actions et des intentions en les rapportant à des valeurs transcendantes (c'est bien, c'est mal…) ; l'éthique est un ensemble de règles facultatives qui évaluent ce que nous faisons, ce que nous disons, d'après le mode d'existence que cela implique*[5] ». Plus précisément, la morale est universelle, elle commande de façon inconditionnelle (si nous suivons par exemple l'impératif catégorique de Kant) et s'impose à ce titre, ou devrait s'imposer, à tous. La morale répond à la question « Que dois-je faire ?[6] ». Différemment, l'éthique résulte de l'opposition, non pas de deux absolus, mais de deux valeurs relatives et subjectives : le bon et le mauvais pour nous. L'éthique est constituée de l'ensemble réfléchi de nos désirs.

Une éthique - car il y en a dès lors plusieurs - ne répond pas à la question « Que dois-je faire ? », mais à la question « Comment vivre ? ». Elle est toujours particulière à un individu ou à un groupe. Elle ne commande pas ; « elle recommande[6] ».

En partant du constat selon lequel les agents autonomes impliquent des modes d'existence et des jeux d'interactions humain – machine qui ne sont pas sans incidence sur la manière dont les subjectivités se construisent (c'est-à-dire sur les processus de subjectivation), se pose alors la question de la problématisation et de la modélisation de l'éthique dans des agents autonomes, question à laquelle le projet ANR pluridisciplinaire – engageant un dialogue entre sciences de l'informatique et sciences humaines – ETHICAA[7] a tenté de répondre. Avec le terme « problématisation » s'énonce la question suivante : quels sont les éléments permettant d'apprécier une situation lorsqu'il est question d'éthique et d'agents autonomes ? En parlant de « modélisation », il s'agit de réfléchir en quel

Quels sont les éléments permettant d'apprécier une situation lorsqu'il est question d'éthique et d'agents autonomes ?

5 Gilles Deleuze, *Pourparlers*.1990.
6 André Comte-Sponville. *C'est chose tendre que la vie*. 2015.
7 http://ethicaa.org/

sens les outils formels permettent de raisonner et de mettre en œuvre dans des agents autonomes des principes éthiques explicitant des tensions entre exigences contradictoires et respect (ou non) des comportements prévus. Le présent livre blanc a ainsi pour objectif de faire un panorama des travaux réalisés au cours du projet et des interrogations que nous avons été collectivement amenés à formuler.

Il convient de noter cependant que ce projet de recherche n'avait pas vocation à formuler des recommandations sur les applications elles-mêmes, comme pourrait le faire un comité d'éthique. Notre travail consistait à faire apparaître dans les modélisations des agents autonomes l'axiologie, les principes éthiques et les dilemmes, ce qui ne

En quel sens les outils formels permettent de raisonner et de mettre en œuvre dans des agents autonomes des principes éthiques ?

conduit pas à faire l'économie de toute analyse réflexive inhérente à notre démarche. En effet, ce n'est pas parce que des concepts éthiques sont modélisés et mis en œuvre dans une application que les usages de cette dernière seront éthiques par construction. Nous n'avons pas la prétention de maîtriser les usages et les détournements des usages et le travail réalisé par le projet ETHICAA ne prémunit pas les applications d'une évaluation par un comité d'éthique : il ne faut pas confondre les représentations de l'éthique (et les discours de légitimation qu'elle génère) et la pratique de l'éthique elle-même qui doit toujours nous inciter à assumer un certain degré de complexité.

Après cette introduction, le chapitre 2 présente les principaux concepts qui ont été utilisés dans le projet, en exposant leurs définitions, les problèmes que ces définitions posent et

Le travail réalisé par le projet ETHICAA ne prémunit pas les applications d'une évaluation par un comité d'éthique.

les choix que nous avons dû opérer. Le chapitre 3 détaille les réalisations techniques du projet ETHICAA. Les chapitres 4 et 5 s'adressent aux chercheurs en Intelligence Artificielle et aux développeurs de systèmes d'agents autonomes en leur proposant des recommandations, ainsi que des pistes de réflexion.

2. CONCEPTS PRINCIPAUX

Il faut s'interroger sur ce qu'est une « valeur » ou un « cadre éthique » codé dans une machine : il s'agit de fait d'un élément de connaissance, mis sous une forme mathématique calculable, et dont la portée et le contenu sémantique sont très restrictifs par rapport à ce qu'on entend en philosophie par valeur ou cadre éthique. Il faut donc être prudent dans l'utilisation des vocables. Les « valeurs » ou « cadres éthiques » représentés et simulés dans une machine constituent bien des représentations, des simplifications, des interprétations de concepts complexes – tout comme le sont les « émotions » que l'on peut faire simuler à un robot : en aucun cas la machine ne sera « morale » ou « éthique ». Ce chapitre présente donc les concepts principaux avec lesquels nous avons travaillé, partant des concepts les plus généraux sur la question de l'éthique jusqu'aux concepts les plus spécifiques liés à la prise de décision.

Éthique et morale

Il convient de permettre une élucidation scrupuleuse des modes de vie qui se voient engagés car nos interactions avec des agents autonomes ne sont pas sans conséquences d'un point de vue existentiel. Toutefois, un tel horizon critique (au sens constructif du terme) doit aussi être en mesure de nous amener à considérer que dans le contexte de sociétés pluralistes, la référence à la nature sacrée de l'humanité n'est plus convaincante pour tout le monde. Il importe pour cette raison de réfléchir à des modes d'évaluation des innovations technologiques qui soient en mesure de tenir compte du pluralisme des valeurs et des différences culturelles qui interviennent dans la compréhension que nous avons des technologies.

De plus, il n'y a pas de risque technologique en soi mais des contextes (sociaux, économiques ou politiques) qui favorisent, ou non, une appropriation clairvoyante des objets technologiques. Ainsi, la même augmentation, au travers de l'exosquelette par exemple, porte en elle des conséquences fondamentalement différentes selon qu'elle se trouve employée à des fins de rééducation neuromotrice ou à des fins militaires. Il est nécessaire, par conséquent, de valoriser des démarches d'évaluation qui soient tou-

Les armes létales autonomes

Le débat est engagé à l'ONU à Genève, dans le cadre de la Convention sur certaines armes classiques (CCAC), sur la question d'une interdiction ou d'un moratoire sur le développement des armes autonomes. Loin de prendre ici une quelconque position, il s'agit avant tout de savoir précisément de quoi on parle. En premier lieu, « autonome » a des acceptions multiples : plutôt que de parler d' « arme autonome », il semble plus pertinent d'étudier quelles fonctions sont, ou pourront être, automatisées, et avec quelles limitations. En second lieu, une arme n'est pas forcément létale – et une arme létale, quel que soit son niveau d'automatisation, n'est en aucun cas animée d'une *intention* de tuer. Cela n'a aucun sens pour une machine ou un logiciel, l'intentionnalité relevant de la responsabilité humaine. Les controverses portent sur le fait qu'une arme puisse être dotée de la capacité de reconnaître des cibles complexes (par exemple, des combattants par rapport à des civils ou des blessés), dans des situations et des environnements eux-mêmes complexes, et de la capacité d'engager de telles cibles sur la seule base de cette reconnaissance. Du point de vue éthique, au-delà de la question de la délégation à une machine de la « décision » de vie ou de mort, de telles capacités supposeraient d'identifier automatiquement et de manière fine une situation, et d'évaluer si les actions envisagées respectent les principes d'*humanité* (éviter les maux superflus), de *discrimination* (distinguer les objectifs militaires des populations et biens civils), et de *proportionnalité* (adéquation entre les moyens mis en œuvre et l'effet recherché) inscrits dans le droit international humanitaire (DIH). Or d'une part, **comprendre** et **apprécier** automatiquement une situation sur la base de modèles mathématiques semble pour le moment hors de portée et d'autre part, les principes du DIH sont des principes généraux comme celui de respect de la dignité humaine, qu'il semble difficile voire impossible d'instancier automatiquement sur une situation réelle précise.

jours en contexte.

Le principal biais néfaste d'une régulation normée au travers d'un texte est la déresponsabilisation des acteurs. D'un côté, les ingénieurs et les concepteurs risquent de se contenter d'être en accord avec le texte, sans s'engager dans une démarche réflexive. De l'autre, les consommateurs ne réfléchissent plus à leurs actions et font confiance aux labels attribués par les éventuels régulateurs. Derrière cette banalisation, le risque est de perdre tout esprit critique.

Un même risque concerne l'éthique elle-même. Sa formalisation textuelle peut vite trahir la réflexivité dont elle est porteuse. Cela reviendrait à figer la réflexion éthique. En se rapportant aux travaux des développeurs en Intelligence Artificielle, il arrive toujours un moment où l'ingénieur doit en effet traduire l'éthique par une formule mathématique à intégrer dans un algorithme. Concrètement, cela peut prendre la forme d'une décision dans le domaine éthique en fonction d'une représentation structurée de connaissances (ontologie). Mais si on résume l'éthique à un problème de logique, cela devient plus que problématique : pour un drone militaire par exemple, cela voudrait dire définir un seuil de nombre de morts civils à partir

duquel la décision de tir est acceptable ? Est-ce souhaitable ? Il n'y a pas d'ontologie de l'éthique, et il est plus que problématique de se laisser emmener sur ce terrain là. Un autre exemple est celui des voitures autonomes pour lesquelles se posent de nombreuses questions quant à la manière de leur faire calculer des décisions : une voiture qui ne peut éviter un obstacle qu'en écrasant des piétons, doit-elle préserver les passagers et sacrifier les piétons, ou le contraire ? Les raisonnements sont multiples. Le penseur pragmatique privilégiera le nombre de vies ; d'autres souhaiteront que la voiture sauve son conducteur quoi qu'il arrive. Ces différences témoignent de l'impossibilité d'établir, pour cet exemple particulier, des principes éthiques universels. Pour les Anciens, les termes morale (mores en latin) et éthique (ethos en grec) signifiaient la même chose et étaient la traduction l'un de l'autre. L'usage les distingue, voire les oppose, aujourd'hui.

On peut résumer cette distinction par le tableau suivant :

Morale	**Éthique**
Champ d'application prétendu universel	Champ d'application particulier
Statut absolu	Statut relatif
Modalité impérative	Modalité hypothétique
Principe du devoir	Principe du désir
Visée de la vie bonne	Visée de la vie juste
Tout discours normatif et impératif : o opposition du bien et du mal o devoirs universels et inconditionnels	Tout discours normatif et hypothétique : o opposition du bon et du mauvais o valeurs immanentes et relatives

La morale résulte de l'opposition de deux absolus : le Bien et le Mal. Elle est constituée d'un ensemble de devoirs et d'interdits qui sont censés guider la vie des sujets humains. À cet égard, ce que nous pouvons dire avec Paul Ricœur[8], c'est que la morale, dans son déploiement de normes, structure et guide des éthiques appliquées qui lui donnent visibilité et lisibilité dans le champ pratique, dans le champ du comment vivre. Les formules impératives générales, selon Ricœur, *« ne deviennent des maximes concrètes d'action que reprises, retravaillées, ré-articulées dans des éthiques régionales, spéciales, telles éthique médicale, éthique judiciaire, éthique des affaires[8] ».*

C'est pour cela que nous parlons d'agents éthiques – expression signifiant plus précisément agents dont le comportement est considéré comme éthique – et non d'agents moraux, sans pour autant congédier les grands courants de la philosophie morale qui sont toujours des sources fondamentales d'enrichissement pour la pratique de l'éthique.

Responsabilité

La responsabilité, dans le langage courant, se réfère à des devoirs et obligations liés à un statut (parent, pilote d'avion, citoyen, etc.). C'est cet usage du terme qui est invoqué dans les implications éthiques de la technologie moderne. Il est ainsi question d'un « principe de responsabilité[9] ».

Il est alors essentiel de distinguer responsabilité et culpabilité. Les humains sont responsables de leurs erreurs et de leurs échecs, et coupables des fautes accomplies délibérément en sachant qu'elles étaient des fautes. Pour être moralement responsable, il doit y avoir une « intention coupable ». En effet, il est possible d'être responsable dans un sens autre que moral sans être coupable alors que la culpabilité implique toujours la responsabilité morale. Par exemple, l'élève qui se trompe est responsable de ses erreurs sans pour cela être coupable de s'être trompé. En revanche, le chauffard en état d'ivresse qui tue des innocents est entièrement responsable de ses actes, mais il est aussi coupable d'avoir conduit en état d'ivresse puisqu'il était conscient de boire plus que de raison avant de prendre le volant. En résumé, être responsable c'est pouvoir et devoir répondre de ses actes en assumant

8 Paul Ricœur. *Dictionnaire d'éthique et de philosophie morale.* 2014.
9 Hans Jonas. *Le principe responsabilité*. 2013.

« *le pouvoir qui est le sien, jusque dans ses échecs, et accepter d'en supporter les conséquences*[10] ».

C'est, comme l'a montré Ricœur, le concept de responsabilité entendu dans son usage juridique classique : en droit civil, la responsabilité se définit par l'obligation de réparer le dommage que l'on a causé à autrui ; en droit pénal, par l'obligation de supporter le châtiment. Nous pouvons, dans ce cas, observer la place qui est accordée à l'obligation : *« obligation de réparer ou de subir la peine*[11] ». Est responsable quiconque est soumis à ces obligations de répondre d'un dommage devant la justice et d'en assumer les conséquences. En tenant les humains responsables de leurs actes, la société fait passer l'intérêt général avant l'intérêt particulier. La société doit donc inculquer à chaque personne les règles de comportement à l'égard des autres, et les faire respecter au moyen des lois, de la force publique et de la pression sociale.

L'éthique de responsabilité exige donc qu'on tienne compte des principes (éthique de conviction) mais aussi, avant toute décision, des conséquences prévisibles de l'acte envisagé (conséquentialisme).

Liberté

Les notions de responsabilité morale, d'intention, d'autonomie, de volonté dépendent toutes à des degrés divers de la notion plus fondamentale de liberté. Aussi qu'entendons-nous par liberté (libre arbitre) et dans quelle mesure est-elle applicable aux agents autonomes ? Le libre arbitre *suppose* que ce que je fais (mon existence) n'est pas déterminé par ce que je suis (mon essence), mais le crée, au contraire, ou le **choisit** librement. Cette liberté renferme l'exigence d'une autonomie absolue, c'est l'existence qui précède l'essence. C'est, selon Sartre, le pouvoir indéterminé de se déterminer soi-même, autrement dit de se choisir ou de se créer.

Pour Spinoza, le libre arbitre n'existe pas. Si les humains se figurent être libres, c'est parce qu'ils ont conscience de leur désir mais ignorent tout des causes qui leur font désirer. C'est la nécessité qui s'impose à nous. Pour autant, cela n'implique pas qu'aucun changement ne soit possible puisque la raison, qui est en nous, est libre. Elle est libre, non pas dans le sens où elle aurait le choix, mais parce que sa nécessité propre est la marque de son indépendance. C'est l'indépendance de la vérité, c'est la libre nécessité du vrai dirait Spinoza.

10 André Comte-Sponville, *Dictionnaire philosophique*. 2013.
11 Paul Ricoeur, *Le juste*. 1995.

Finalement, l'agent autonome libre est celui qui se possède par la réflexion, celui qui compare, analyse, prévoit et juge les différentes séries de phénomènes. Sa liberté n'est pas un absolu, elle dépend de plusieurs conditions en raison desquelles elle varie. C'est pourquoi un agent autonome – **contraint par les nécessités de sa programmation** et disposant de fonctions, certes limitées, de prévision et de décision – peut être vu comme libre au sens de Spinoza mais certainement pas libre au sens de Sartre.

Autonomie et automatisme

Au sens premier, l'autonomie désigne l'aptitude à se donner ses propres lois. Cela peut avoir une signification politique : la capacité d'une nation à légiférer d'elle-même, sans se voir imposer des normes de l'extérieur, par d'autres, que ce soient des nations étrangères, des organismes supranationaux ou des autorités spirituelles, ce qui correspond, peu ou prou, à l'idée de souveraineté nationale. Cela a aussi une signification morale : un être autonome choisit lui-même les maximes sur lesquelles il décide de régir sa propre conduite. Cela s'oppose à l'hétéronomie de celui qui soumet sa volonté à une influence extérieure, par exemple à des personnages puissants ou à la satisfaction de ses besoins animaux. Enfin, il est courant d'utiliser l'adjectif autonome pour qualifier une entité technique qui ne fait pas intervenir d'agent humain entre la prise d'information et l'action. En ce sens, nous parlons de « voitures autonomes » ou d'« armes autonomes » et plus généralement d'« agents autonomes ». Notons toutefois qu'il s'agit là d'un abus de langage, car c'est bien plutôt d'automatismes, c'est-à-dire d'entités qui agissent sans qu'intervienne, dans le temps de l'action, une volonté extérieure, que d'autonomie au sens propre, puisque ces entités ne manifestent pas de volonté spontanée qui échapperait à leurs utilisateurs et *a fortiori* à leurs concepteurs.

Jugement

Un point important, et malheureusement trop souvent omis, porte sur le jugement entendu au sens classique de l'opération de connaissance qui fait passer de sensations ou de perceptions à un concept. Un individu qui aurait perdu l'usage de ses sens ou serait victime d'hallucinations ne serait pas nécessairement en mesure de porter un jugement sur une situation. Soulignons, pour éviter tout malentendu, que le jugement en question est un acte de l'entendement qui se distingue de

l'acte judiciaire consistant à décider de la culpabilité d'un prévenu ou à prononcer une sentence.

Le jugement est à l'origine de tout comportement rationnel ; sans lui, il n'y a ni liberté, ni éthique possible.

Pour les agents autonomes, le pendant du jugement correspond à l'interprétation des signaux générés par les capteurs ; on conçoit qu'elle conditionne leur comportement, car si cette interprétation est erronée, cela peut conduire à des calculs de décisions qui seraient inadaptées, voire catastrophiques. À titre d'illustration, dans l'exemple de l'accident de la voiture autonome (voir page 42), c'est l'équivalent d'un défaut de jugement, à savoir d'une mauvaise interprétation, qui a conduit à négliger l'information envoyée par les capteurs. Dans un registre différent, Ronald Arkin explique que les robots seront « plus éthiques » que des soldats humains sur un champ de bataille, car ils ne perdront pas leur sang-froid, agiront en respectant les règles dites de la guerre juste et, en conséquence, ne tueront qu'à bon escient, de façon « morale ». Or, deux des principes fondamentaux de la guerre juste tiennent à la discrimination entre civils et militaires et à la proportionnalité de la réponse à l'attaque. Et ces deux points font appel à un jugement qui apparaît très difficile à traduire sous forme algorithmique, surtout lorsque les combattants ne portent pas d'uniforme et que la supériorité technologique d'un camp surpasse considérablement celle de l'autre, comme c'est le cas dans les guerres asymétriques contemporaines.

Délibération

Après avoir identifié la situation, l'agent autonome détermine les actions à accomplir au cours d'une phase critique de délibération, soit individuelle, soit collective, que l'on met en œuvre au travers de modèles informatiques.

Dans tous les cas de délibération collective, l'explicitation des arguments avancés en faveur ou en défaveur de chacune des options autorise la contre argumentation et donc le débat collectif. Dans le cas de la délibération individuelle, il convient aussi de confronter toutes les options en déployant les arguments en faveur ou en défaveur de chacune d'entre elles.

Pour bien comprendre la nature des difficultés rencontrées tant pour la délibération individuelle que collective, il convient de distinguer trois plans dans la décision d'un agent,

- ce qui relève de la **prudence** ou en termes kantiens, de l'impératif pragmatique, à savoir de la détermination des objectifs à poursuivre,

- ce qui correspond à la **sagacité**, ou en termes kantiens à l'impératif problématique, autrement dit à la capacité à découvrir les actions à accomplir les plus appropriées pour réaliser les objectifs que l'on s'est fixés, et enfin
- ce qui relève de la **moralité**, à savoir en termes kantiens l'impératif moral, qui assure que les actions n'enfreignent pas les maximes de la propre volonté de l'agent, autrement dit les règles que l'on s'est données et qui se présentent comme des prescriptions morales.

Aujourd'hui, les finalités des agents autonomes sont fournies par le concepteur, l'usager ou d'autres agents ; leur prudence se résume donc à l'obéissance à des injonctions fixées par avance, par exemple aller d'un point A à un point B pour une voiture autonome.

La modélisation de la sagacité relève de ce que l'on appelle traditionnellement la résolution de problèmes, discipline qui a été beaucoup explorée en Intelligence Artificielle depuis de nombreuses années. Nous n'avons pas besoin d'insister là-dessus, car cela ne relève pas directement de l'éthique.

En revanche la modélisation de la moralité fait l'objet de recherches plus récentes, entre autres de celles poursuivies au sein du projet ETHICAA. Ces travaux ont visé à introduire des valeurs humaines dans la prise de décision afin de limiter les capacités d'actions en fonction de l'évaluation morale de leurs conséquences ou de leur légitimité au regard de maximes que l'on s'est données. Dans ce contexte, différents cadres éthiques, conséquentialistes, déontologiques ou utilitaristes, peuvent être mis en œuvre et comparés. La difficulté tient à l'existence de dilemmes éthiques comme ceux qui se présentent avec la voiture autonome qui doit décider soit d'écraser cinq jeunes gens imprudents, soit de tuer son passager. Pour mettre en lumière ces conflits, nous recourons à des formalismes logiques développés en Intelligence Artificielle mettant en œuvre des raisonnements « non monotones » qui permettent de faire face à des contradictions.

Ainsi, imaginons que l'on conçoive un agent destiné à accompagner des patients atteints d'un diabète qui n'ont pas le droit d'abuser du sucre. Si cet agent s'aperçoit qu'un patient s'apprête à enfreindre les règles très strictes que lui imposent ses médecins en prenant un chocolat, l'agent doit-il se contenter d'avertir son « protégé » ou doit-il immédiatement prévenir son médecin ? S'il déclenche tout de suite l'alerte, il sera insupportable à la fois au patient et au médecin. S'il choisit de limiter ses réactions dans le cas d'une infraction mineure, à partir de quand doit-il vraiment lancer l'alerte, parce qu'il y a péril ?

L'axiologie et le traitement des valeurs doivent déboucher sur des définitions formelles explicites ou des taxonomies utilisables en Intelligence Artificielle.

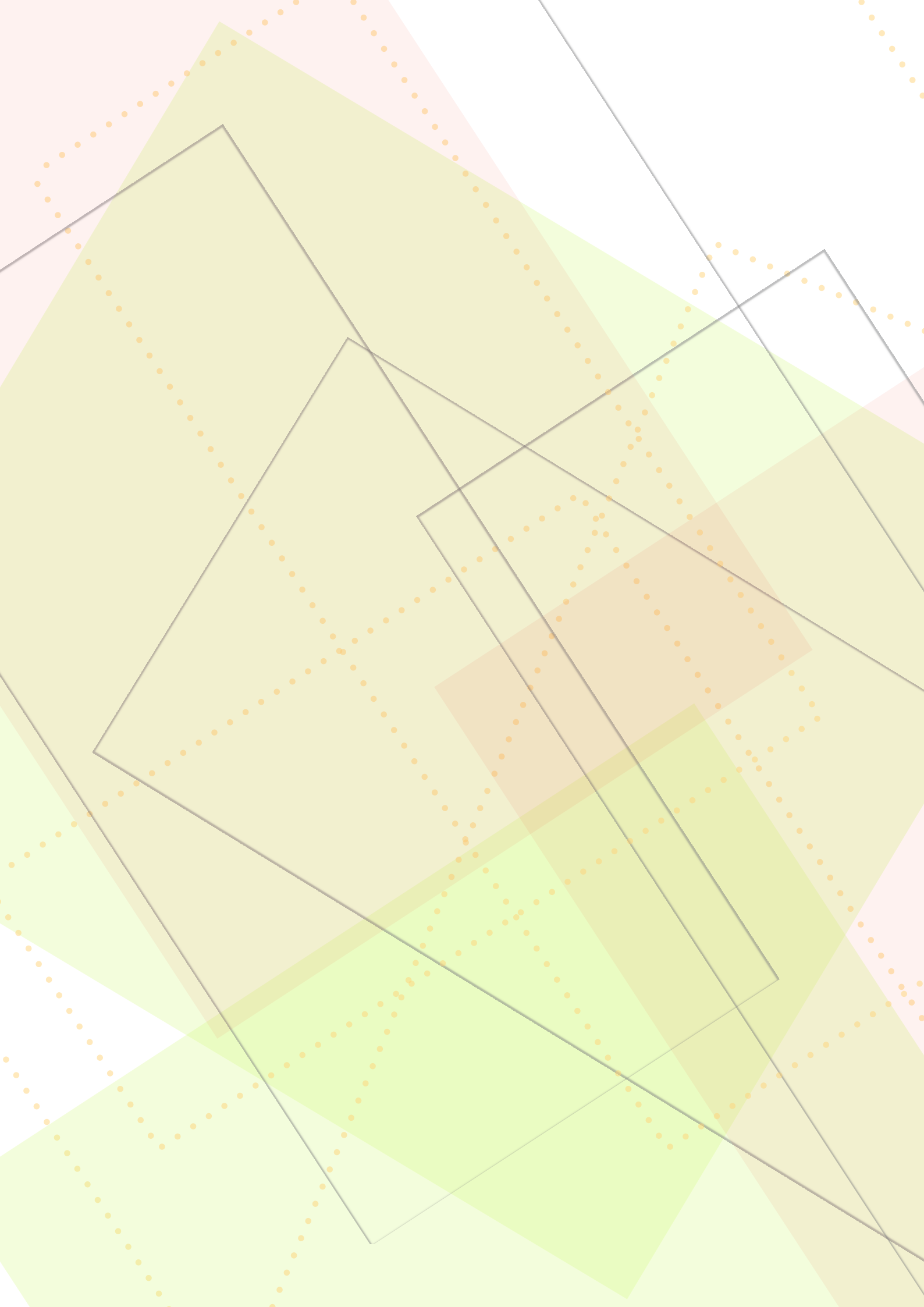
Décision

Décider signifie trancher, opter pour une conclusion, choisir entre plusieurs options possibles. Une fois la décision calculée, l'agent peut agir au moyen de ses facultés exécutives, mettant ainsi en œuvre le choix, par exemple dans le cas de la voiture autonome, en accélérant ou en tournant à gauche ou à droite. Cependant, le choix lui-même est fait par l'agent à partir de données et de règles, ce qui fait que, même s'il paraît décider de lui-même, il ne fait qu'exécuter le code informatique mis en place par ses programmeurs. Il reste bien sûr à traduire toutes les connaissances qui permettent de peser les différentes options à partir de critères éthiques.

Valeur

Si en sociologie, il n'existe pas de consensus au sein de la communauté quant à une définition univoque des valeurs, la philosophie propose par l'« axiologie » l'étude des valeurs. Il est alors important de noter que les valeurs et les systèmes de valeurs forment un langage. Les énoncés évaluatifs expriment des valeurs, par opposition à des énoncés prescriptifs qui expriment des normes. De plus, les valeurs procèdent par degré en indiquant un idéal vers lequel tendre. On peut par exemple être plus ou moins courageux.

Il convient de remarquer qu'un gouffre important existe à ce jour entre normes juridiques (très générales), normes techniques (très précises) et valeurs (difficiles à hiérarchiser). Il s'ensuit des prises de positions dans le débat sociétal, scientifique et technique assez fortement divergentes dès lors qu'on cherche à appliquer une ou plusieurs valeurs sociales à des agents autonomes. C'est pourquoi l'axiologie et le traitement des valeurs doivent déboucher sur des définitions formelles explicites ou des taxonomies utilisables en Intelligence Artificielle.



3. RÉALISATIONS DU PROJET ETHICAA

Au cours du projet ETHICAA, en partant des réflexions présentées dans la section précédente, nous avons proposé un cadre de raisonnement permettant à un agent autonome d'évaluer son environnement, d'intégrer des principes éthiques et de déterminer à partir de la mise en œuvre de ces principes soit un plan d'actions mettant en place un comportement considéré comme éthique, soit une évaluation de l'éthique du comportement d'autres agents. Ce cadre s'appuie sur deux distinctions explicitement représentées ainsi que sur un ensemble de fonctionnalités.

La première distinction que nous opérons est entre morale et éthique, c'est-à-dire entre le raisonnement sur le bien et mal, et le raisonnement sur le juste et l'injuste comme vu au chapitre précédent. La seconde distinction explicite est entre éthique individuelle et éthique collective car, dans le deuxième cas, il s'agit de permettre à un agent de tenir compte de la pluralité des valeurs et principes des autres agents.

Il s'agit de permettre à un agent de tenir compte de la pluralité des valeurs et principes des autres agents.

Enfin, notre cadre est construit à partir d'un ensemble de fonctionnalités nécessaires au traitement de l'éthique dans les systèmes d'agents autonomes : percevoir les situations de dilemme, attribuer la causalité et les responsabilités, juger, décider et agir en fonction de principes éthiques et moraux, collaborer et faire confiance aux autres agents, expliquer et justifier les décisions, et enfin être capable de vérifier formellement l'éthique du comportement d'un agent.

Perception des dilemmes

L'identification automatique d'un dilemme en tant que tel est une tâche complexe et néanmoins pertinente. En effet, un agent autonome n'étant pas confronté systématiquement à des prises de décision nécessitant un raisonnement intégrant des considérations éthiques, il peut être utile de ne procéder à un tel raisonnement qu'après avoir identifié la situation comme nécessitant ces considérations éthiques. De plus, si l'on se place dans le cadre d'un système opérateur-robot, il est intéressant de pouvoir signaler à l'opérateur par la levée d'une alarme que l'agent autonome rencontre une situation nécessitant son intervention.

Afin d'identifier formellement un dilemme éthique, il convient de le

définir. En nous appuyant sur les travaux d'Aroskar[12], nous pouvons constater qu'un dilemme se définit par l'absence de « bonne solution ». Autrement dit, toute décision possible est insatisfaisante. S'il existe une décision satisfaisante, alors la situation n'est pas un dilemme. Remarquons que dans la définition originelle, il existe aussi une situation de dilemme lorsqu'il n'existe aucune solution. Cependant, dans le cadre d'un agent ayant à mener un raisonnement sur les décisions, nous estimons qu'il n'est pas pertinent de considérer cette définition. Il convient alors de définir ce que l'on entend par « insatisfaisant ». Si nous modélisons une situation comme un ensemble de décisions ayant des conséquences, nous pouvons poser l'insatisfaction pour une décision comme une disjonction entre deux conditions :

1. Les conséquences de la décision sont insatisfaisantes lorsque au moins un fait faisant partie de ses conséquences est apprécié comme négatif.
2. La décision est insatisfaisante en elle-même, c'est-à-dire qu'elle ne doit pas l'être pour ses conséquences, puisque étudiées séparément, mais par sa nature. Nous pouvons imaginer une situation où « mentir » mène à un ensemble de conséquences satisfaisantes. Cependant, cette décision n'est pas satisfaisante car la décision en elle-même est appréciée négativement.

Notons que dans les deux cas il s'agit d'une appréciation propre à l'agent qui juge la décision. Qu'il s'agisse d'identifier un dilemme ou de raisonner sur ce dilemme, il est important de toujours garder à l'esprit les limites de perception d'un agent.

Quelles conséquences ?

Lorsqu'une décision a un impact sur le monde, il est difficile voire impossible d'en appréhender toutes les conséquences. En effet, les conséquences étant nombreuses, parfois imperceptibles pour certains agents, et entraînant elles-mêmes d'autres conséquences (et ainsi de suite), être exhaustif dans leur analyse est impossible. Un choix est donc obligatoirement réalisé dans la sélection des faits pertinents qui seront étudiés. Ce choix a un impact direct sur l'identification d'un dilemme. En effet, si une conséquence négative n'est pas détectée, un dilemme peut ne pas être identifié comme tel.

12 Mila Aroskar. *Anatomy of an ethical dilemma: The theory*. American Journal of Nursing, 80(4):

Quel agent ?

Un agent a des capacités de perception limitées, et de plus ces perceptions varient d'un agent à l'autre. Non seulement un agent pourra ne pas prendre en compte une conséquence qu'un autre agent considère, mais en outre il pourra évaluer cette conséquence différemment. Ainsi, une situation pourra être un dilemme pour un agent mais pas pour un autre.

Causalité et responsabilité

Constituant un bloc important d'un cadre de raisonnement éthique, la causalité est une notion subtile qui a été et reste amplement discutée en philosophie. Elle revêt aussi une grande importance pour d'autres domaines tels que le droit, où elle est centrale pour l'attribution de la responsabilité légale. Or une investigation de cette notion fait émerger de nombreuses nuances. Nous proposons quelques pistes pour les aborder.

Causalité générale et causalité effective

Une distinction se fait entre une *causalité générale* (« Rouler vite provoque des accidents ») et *une causalité effective* (« Le fait que Caitlyn ait roulé vite a provoqué son accident d'aujourd'hui »). Une distinction peut aussi être faite au sein de la causalité effective entre ce qui est considéré comme la cause réelle et ce qui ne doit être vu que comme circonstanciel. Dans le cas de l'accident de Caitlyn, nous pouvons nous demander si c'est la conduite de Caitlyn, la puissance du moteur ou l'état de la route qui est la cause de l'accident. Déterminer cela demande d'évaluer l'importance de chaque critère pour choisir une cause première. À l'inverse, nous pouvons considérer que dans la mesure où tous ces éléments participent d'une façon ou d'une autre au résultat final, ils peuvent tous être considérés comme des causes.

Insuffisance de la définition contre-factuelle de la causalité

Suivant le philosophe anglais David Hume, une approche courante est de définir la causalité en termes de *dépendance contre-factuelle* : A est une cause de B quand, si A n'était pas arrivé, B ne serait pas arrivé non plus. Cette définition échoue cependant à capturer certaines subtilités de la causalité comme les cas de préemption (une cause peut être remplacée par une autre) ou de surdétermination (il y a plus de causes que nécessaire pour provoquer l'effet). L'exemple suivant illustre un cas de préemption : deux enfants, Suzy et Billy, décident de lancer tour à tour une pierre sur une bouteille, sachant que chaque pierre la brisera. Suzy lance sa pierre (s-lance) et brise la bouteille (brisée). Comme Billy aurait tout de même brisé la bouteille

en lançant sa pierre (b-lance) si Suzy n'avait pas lancé la sienne alors (s-lance) ne dépend pas contre-factuellement de (brisée). Ainsi, selon Hume, Suzy n'a pas causé le bris de la bouteille et ne saurait en être tenue responsable, ce qui est problématique dans un cadre de responsabilité morale si nous remplaçons par exemple le bris de la bouteille par la mort d'un être humain.

Événements causés et événements évités

Comme le pouvoir causal des actions d'un agent semble constituer le seul lien entre lui et le monde, il est naturel de suggérer que les agents sont responsables des états du monde qu'ils ont causés. Cependant, il arrive aussi que l'on veuille tenir responsable un agent non pour quelque chose qu'il a causé, mais pour quelque chose qu'il a évité. C'est par exemple le cas lorsqu'un médecin sauve la vie d'un patient, prévenant sa mort. Cette notion de prévention demande un traitement informatique particulier puisqu'il s'agit de raisonner sur des événements qui n'ont pas eu lieu.

Actions et omissions

Le choix d'un agent dans une situation donnée peut se traduire par une action, mais aussi par une inaction. Or, il est des cas où l'inaction endosse sans conteste un poids moral. C'est le cas par exemple si un agent choisit de ne pas agir pour sauver un enfant qui se noie. Le fait qu'il ait eu le pouvoir de changer le résultat implique sa responsabilité. Cette notion d'omission demande donc une attention particulière en parallèle à celle portée aux actions, et comme dans la distinction précédente, il s'agit de traiter d'un événement qui n'a pas eu lieu – ici une action.

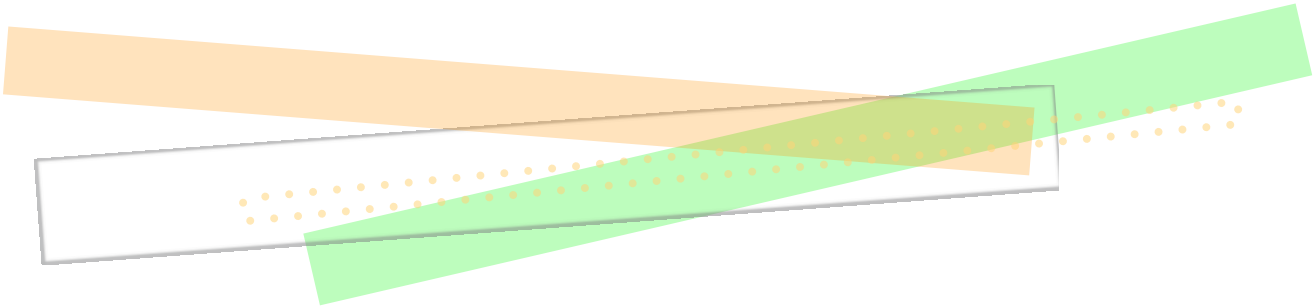
Les valeurs et principes éthiques modélisés

Les modélisations et mises en œuvre réalisées dans le cadre d'ETHICAA ont d'abord pour objectif de juger une décision d'un point de vue éthique ou axiologique, soit dans l'absolu, soit par rapport à d'autres décisions possibles.

L'éthique par raisonnement a été privilégiée par rapport aux approches d'éthique par conception. Ce choix s'appuie sur le fait qu'une représentation explicite de connaissances et la mise en place d'outils de raisonnement associés permettent d'envisager la proposition de décisions argumentées et justifiées à un opérateur ou un utilisateur dans le cadre d'une aide à la décision. Ces modélisations et mises en œuvre ont été étudiées sur des cas-jouets de dilemmes éthiques (par exemple : dilemme du trolley et va-

riantes) puis testées sur des cas plus réalistes comme les enchères à haute fréquence.

Ces modélisations et réalisations concernent (i) les théories du bien et du juste, concrétisations de la distinction entre morale et éthique qui fonde les travaux menés au sein du projet, (ii) le jugement éthique pour décider au sein d'un agent, (iii) le jugement de l'éthique des autres agents et (iv) le jugement de l'éthique pour la coopération entre agents.



Théorie du bien et théorie du juste
La modélisation s'appuie sur :

A. Un **modèle du bien** qui donne une appréciation de la valeur morale intrinsèque de finalités ou d'événements. Parmi les modalités du bien, on trouve par exemple les vertus de l'agent décideur (altruisme, courage…), le droit des agents qui subissent les conséquences des décisions (droit à la vie, droit à la propriété…), le bien-être de tous les agents.

B. Un **modèle du juste** qui détermine l'action éthiquement la plus adaptée selon des circonstances données, le champ des possibles et les désirs des agents. Ce modèle prend en compte le contexte dans lequel l'action est choisie. Parmi les cadres éthiques qui composent en partie un modèle du juste, nous trouvons :

   - des cadres conséquentialistes, dont le jugement porte sur les conséquences d'une décision ou d'une action (par exemple le conséquentialisme positif – maximisation du bien, ou négatif – minimisation du mal),
   - des cadres déontologiques, dont le jugement porte sur la décision ou l'action elle-même (par exemple le code de déontologie médicale),
   - des cadres mixtes, dont le jugement considère à la fois la décision elle-même et ses conséquences positives et négatives (par exemple la Doctrine du Double Effet).

Éthique dans un cadre mono-agent

L'éthique pour la décision a été abordée en s'intéressant d'une part à un modèle d'action et de causalité, et d'autre part à un modèle de jugement individuel.

- Au sein même d'un agent, nous représentons la prise de décision éthique en adjoignant au modèle du bien et au modèle du juste deux autres modèles interdépendants qui permettent à l'agent d'évaluer son environnement et de raisonner sur sa responsabilité. Ces deux modèles sont : (1) un **modèle d'action** permettant à l'agent de représenter son environnement et les changements qui s'y déroulent ; (2) un **modèle de causalité** qui piste les conséquences des actions, rendant possible un raisonnement sur l'imputabilité des agents. Ces deux modèles produisent une compréhension du monde qui ne fait pas intervenir de considérations éthiques, à la différence des deux modèles de la théorie du bien et de la théorie du juste qui superposent une compréhension éthique du monde. Les différents modèles sont interdépendants à des degrés variables. Les modèles du bien et du juste reposent toujours sur un modèle d'action et un modèle de causalité. Cependant, alors qu'un modèle de causalité est toujours nécessaire, la formulation particulière du moteur causal peut varier, par exemple pour représenter différentes définitions des causes et des conséquences. En ce qui concerne les modèles éthiques, avoir un modèle du bien est nécessaire pour modéliser les théories conséquentialistes (puisque leur but est de promouvoir ce bien), mais n'est pas nécessaire pour formuler des théories purement déontologiques. En effet, les théories conséquentialistes expriment des principes de la forme « Il faut maximiser le bien de telle ou telle manière » et demandent donc à ce que ce bien soit défini (par des valeurs ou des indicateurs de bien-être). D'un autre côté, les théories déontologiques peuvent parfois simplement exprimer des principes comme « Il ne faut jamais mentir » ou « Il ne faut pas utiliser les autres pour ses propres fins » sans qu'il y ait pour autant une théorie de ce qui est bien ou mal intrinsèquement.
- Un **modèle de jugement éthique** a été intégré au sein d'une architecture réflexive d'agent de type « Belief/Desire/Intention ». S'appuyant sur une représentation locale et à jour de la situation courante perçue par l'agent, ce modèle permet de réaliser un jugement de l'éthique en contexte. Une approche rationaliste excluant la prise en compte d'émotions dans le processus de jugement et de décision des agents a été privilégiée : les buts et croyances de l'agent sont les seuls éléments pris

La justice prédictive

La notion de justice prédictive peut s'entendre d'au moins trois façons, selon que l'on s'intéresse au jugement, auquel cas cela peut avoir deux significations, ou à la législation. Nous nous intéresserons ici deux premiers volets qui portent l'un et l'autre sur la façon dont la prédiction pourrait être prise en considération dans l'acte de juger, car le troisième, à savoir l'intervention de l'anticipation dans l'élaboration des lois, va bien au-delà de nos compétences et de notre propos. Le jugement judiciaire dont il est question ici correspond soit à la détermination de la culpabilité d'un prévenu, soit à l'établissement et à l'énoncé de la sanction. Dans la première acception, à savoir pour établir la culpabilité, l'idée d'une justice prédictive signifierait que la faute serait anticipée avant même que l'acte ait été commis. Cela aurait l'avantage, si cela marchait, de diminuer considérablement la criminalité. Cependant, cela serait éthiquement fort discutable, puisqu'en assimilant le crime virtuel au crime actuel, on rendrait coupables ceux qui aurait un simple profil de délinquants et on les empêcherait de s'amender. C'est malheureusement ce que quelques hommes politiques préconisent en demandant d'enfermer tous les « fichés S ». Dans le second sens, la justice prédictive signifie que l'on calcule la peine en fonction de l'anticipation de récidive. Cela peut à son tour s'entendre de deux façons. Dans la première, on détermine la sévérité de la peine en fonction de la probabilité de récidive évaluée avec des systèmes prédictifs, autrement dit plus l'anticipation du risque de récidive est élevée, plus la peine serait importante. Cela pose de nombreux problèmes liés aux indicateurs qui anticipent la récidive et à leur caractère potentiellement discriminatoire. On peut aussi entendre l'introduction de la prédiction dans l'établissement de la sanction en se demandant quel est l'effet de la sanction sur la probabilité de récidive. Cela conduit, entre autres, à des peines de substitution à la prison, ou au contraire, si l'on constate que la prison a une vertu rédemptrice, à accroître la peine de prison. En conclusion, dans le contexte de la justice prédictive, les questionnements éthiques portent sur l'évaluation de la récidive avec des systèmes prédictifs fondés sur l'apprentissage machine et entraînés avec de grandes masses de données. Plus généralement, l'usage quasiment systématique du profilage des identités exclut le changement de cap et le décentrement, éliminant, de ce fait, la possibilité d'explorer différents modes d'action. Dans le cadre du projet ETHICAA, nous nous sommes centrés sur la problématisation et la modélisation de conflits éthiques, sans travailler directement sur les questions d'apprentissage et de profilage. Néanmoins, pour donner aux individus les moyens de se penser comme des sujets autonomes et créatifs, il convient de leur restituer la maîtrise des choix en leur permettant de comprendre et de critiquer les suggestions de la machine.

en compte dans le processus de jugement en dehors des théories du bien et du juste. Une telle architecture permet à un agent autonome de sélectionner une ou plusieurs actions à exécuter parmi l'ensemble des actions qu'il juge éthiques selon les théories du bien et du juste dans la situation courante.

Éthique dans un cadre multiagent

De nombreux travaux se sont déjà intéressés à la modélisation de l'axiologie et de l'éthique, en s'appuyant pour la plupart d'entre eux sur la caractérisation d'éléments fondamentaux que peuvent être la causalité, la responsabilité, les droits et devoirs spécifiques à des applications ou des principes éthiques plus généraux. Cependant, ces travaux portent essentiellement sur l'éthique du comportement individuel de l'agent dans l'accomplissement de ses objectifs. Toutefois, de nombreux domaines applicatifs mettent en présence plusieurs agents qui doivent interagir, décider conjointement et coopérer. Dans ce contexte, un agent doit non seulement tenir compte de principes éthiques au regard de ses propres objectifs mais aussi quant à la manière dont il coopère, et par extension quant à la manière dont il tient compte des principes éthiques des autres agents. Dans le cadre du projet ETHICAA, deux directions de recherche d'agents autonomes avec une pluralité d'éthiques ont été initiées dans un cadre multiagent : jugement de l'éthique des autres pour coopérer et formation éthique de collectifs d'agents.

Jugement de l'éthique des autres pour coopérer

Dans le cadre de la coopération fondée sur l'éthique, les travaux ont conduit à la prise en charge du jugement des autres agents comme valeur d'entrée dans la décision d'un agent de coopérer avec eux. Pour ce faire, le processus de jugement éthique intégré au sein de l'architecture d'agent a été utilisé pour juger de l'éthique du comportement d'autres agents du système. Ce point de vue externe utilise le processus de jugement pour qualifier l'éthique du comportement d'un ou de plusieurs autres agents à partir des actions qu'ils ont exécutées. Selon le degré de connaissance de l'« agent juge » sur les théories du bien et du juste de l'« agent jugé », plusieurs types de jugements peuvent être mis en œuvre :

1. jugement aveugle (l'agent juge utilise ses propres théories uniquement),
2. partiellement informé (l'agent juge se fonde sur une connaissance partielle des théories de l'agent jugé),
3. totalement informé (l'agent juge se fonde sur une connaissance totale).

Ce processus se traduit par un calcul de la confiance et de la proximité éthique entre les agents. Cela permet ensuite à chaque agent de décider de coopérer ou non avec d'autres agents et de potentiellement leur déléguer la réalisation de certaines actions. Ce travail a été testé et validé sur la vente et l'achat de biens financiers sur une place de marché simulée.

Formation éthique de collectifs

Dans le cadre de la formation éthique de collectifs d'agents, c'est-à-dire de groupes d'agents qui vont devoir par la suite collaborer pour effectuer une action, nous avons proposé une modélisation d'une éthique du bien fondée sur les **vertus**. Pour ce faire, nous avons proposé un nouveau modèle de jeux de coalitions – les *jeux de déviation hédoniques* – où chaque agent exprime des conditions qui lui sont propres pour caractériser la manière dont il tient compte des autres agents. Nous avons formellement défini ce qu'était une solution d'un tel jeu qui faisait consensus entre tous les agents et nous avons montré que certaines compositions de conditions permettent de retrouver les solutions classiques de la littérature. Nous avons alors pu nous servir de notre modèle pour caractériser trois éthiques fondées sur des vertus : la liberté, l'altruisme et l'hédonisme.

Justifications et explications

Il convient dans un premier temps de différencier explications et traces d'exécution concernant une décision. Les *traces d'exécution* peuvent être analysées par un expert humain afin de comprendre pourquoi un agent autonome a calculé une certaine décision. Par exemple, un système formel peut produire une preuve complète montrant qu'une décision permet d'atteindre un but donné. Toutefois, ces traces sont difficiles à interpréter pour un non-expert et si cela nécessite un effort cognitif trop important, ces traces risquent de ne pas être utilisées. C'est pourquoi il est préférable de considérer des agents autonomes capables de fournir des *justifications* ou des *explications*.

Du besoin de justification

Une explication décrit comment une décision donnée a été calculée. Il s'agit alors d'une représentation compacte d'une trace d'exécution mettant en lumière les principaux éléments qui ont conduit à la prise de décision. Plus concrètement, une explication vise à répondre aux deux questions suivantes : quels sont les principaux facteurs conduisant à la décision, et quels changements dans les entrées auraient changé la décision ? Cependant, vouloir produire des explications au regard de critères éthiques pose de nouvelles questions. En effet, comme l'éthique est un raisonnement en contexte, les

explications doivent tenir compte de cette notion. Il ne s'agit plus seulement d'expliquer les décisions mais aussi *d'expliquer la manière dont un agent évalue la situation dans laquelle il prend une décision*. C'est pour cela qu'il est préférable de considérer des justifications.

Si une explication décrit *comment* une décision donnée a été calculée, une justification décrit *pourquoi* la décision calculée est une bonne décision. Les études en psychologie ne donnent pas de réponses définitives quant à ce que doivent être ces justifications, mais plusieurs éléments pertinents sont classiquement mis en avant : pourquoi la décision a-t-elle été prise ? Pourquoi une décision donnée n'a-t-elle pas été prise ? Pourquoi la décision qui a été prise est la meilleure au regard d'un critère ? D'un point de vue concret, la mise en œuvre de ces justifications peut passer par la mise en évidence de liens de causalité, d'un événement indésirable ou d'une absence de solution si une autre décision avait été prise ou même une métrique de performance explicite. De plus, afin de calculer ces éléments, il est possible de considérer soit des *systèmes externes aux agents* qui vont les interroger pour construire ces réponses, soit des *systèmes internes aux agents* où la construction des justifications fait partie intégrante de leurs mécanismes de décision. Par exemple, les mécanismes de raisonnement pratique fondés sur de l'argumentation formelle constituent une approche pertinente. En effet, cette technique représente les arguments pour et contre les différentes décisions possibles et calcule un ensemble d'arguments acceptables au regard d'une certaine sémantique. Les arguments sont représentés dans une structure hiérarchique qu'un utilisateur peut parcourir pour avoir de plus en plus d'arguments spécifiques.

Arguments pour représenter l'éthique

Au cours du projet ETHICAA, nous avons ainsi proposé un **modèle formel d'argumentation éthique**, inspiré du modèle des ordres d'André Comte-Sponville[13]. Dans ce modèle, un agent à éthique artificielle est doté de plusieurs théories logiques : une théorie des croyances, une théorie des désirs, une théorie d'action, une théorie normative, une théorie morale et une théorie des valeurs. Chaque théorie permet, en fonction du contexte, de générer des arguments. Par exemple, la théorie normative va générer des arguments exprimant qu'il est permis de faire A dans le contexte C et la théorie morale qu'il est bon de

13 André Comte-Sponville. *Le capitalisme est-il moral ?* Albin Michel. 2004.

faire A' dans le contexte C au nom d'une valeur V. Ces arguments interagissent en mélangeant une approche déontologique (certains arguments sont plus ou moins préférés *a priori* par l'agent) et conséquentialiste (plus il y a d'arguments en faveur d'une décision, plus cette dernière est acceptable).

Enfin, produire des justifications au sein d'un système multiagent pose de nouvelles questions. En effet, en raison de la nature distribuée, décentralisée, voire hétérogène et ouverte, de tels systèmes, il convient de se demander : comment justifier une décision lorsqu'elle est influencée par les décisions d'autres agents ? Comment réutiliser des justifications fournies par un autre agent ? Comment faire lorsqu'un agent est incapable de fournir sa part de justification ?

Vérification de l'éthique d'un comportement

Dans le cadre du projet ETHICAA, il est apparu que le problème de savoir si un agent autonome vérifiait une certaine éthique pour l'utiliser dans un contexte donné était crucial. Cela est d'autant plus nécessaire que l'agent est *a priori* autonome, c'est-à-dire que son comportement, défini par ses développeurs, peut être complexe et non maîtrisé. Aussi avons-nous cherché à montrer comment il était possible de garantir formellement que le comportement d'un agent respecte une éthique particulière. Nous avons donc proposé un environnement de spécification et de vérification formelle du comportement éthique d'un agent autonome.

En partant des travaux d'Abramson et Pike[14], une règle morale est représentée par une propriété formelle qu'un agent doit respecter dans certains contextes et un agent a un comportement éthique si son comportement est conforme aux règles morales relevant des différents contextes dans lesquels il se trouve.

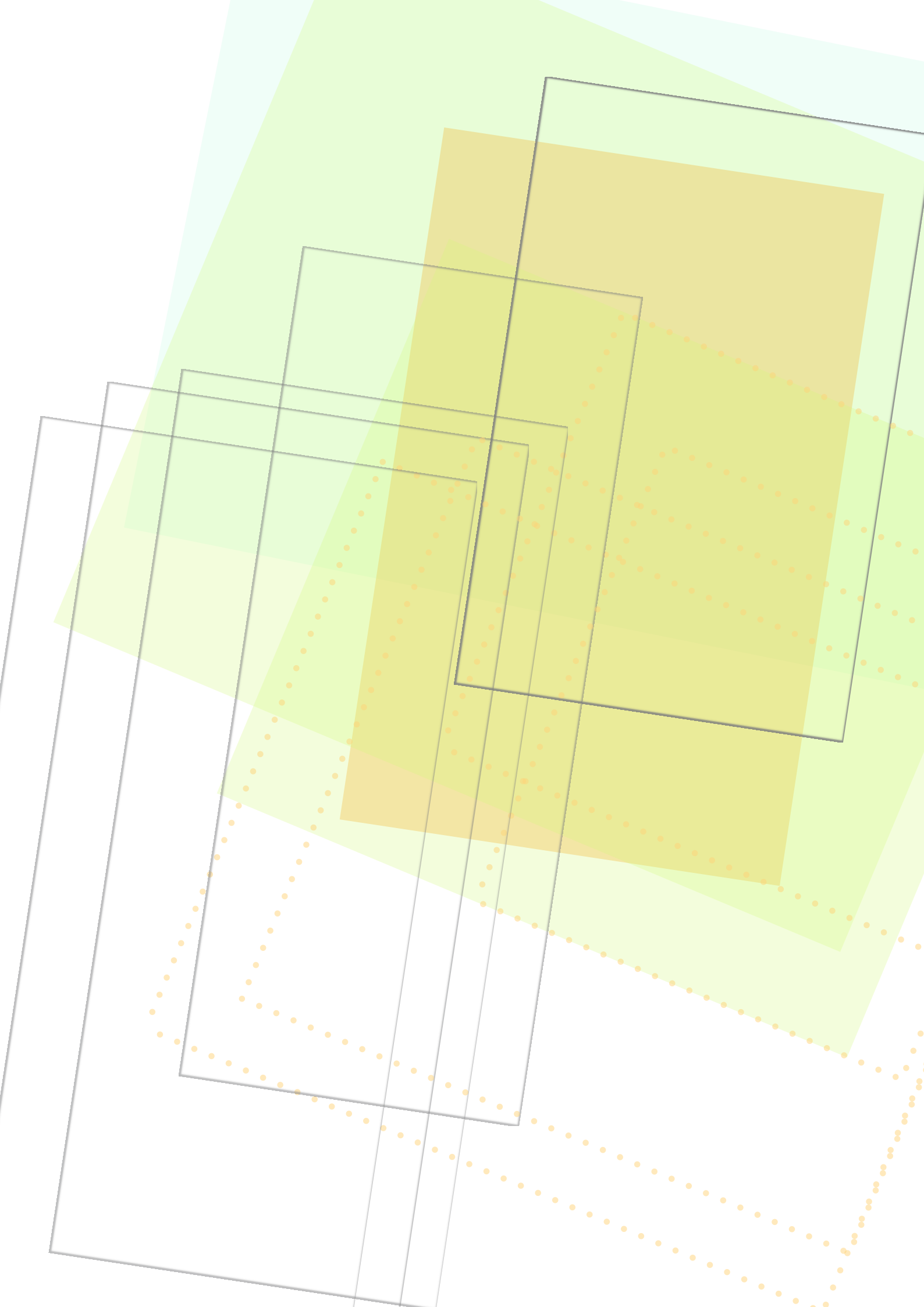
Nous avons tout d'abord proposé un cadre de formalisation assez général pour les règles morales et pour les principes éthiques. Ainsi, nous avons choisi de représenter chaque règle morale sous la forme d'un ensemble d'états corrects dans différentes situations. En effet, une règle morale peut engendrer différents comportements selon les conditions dans lesquelles on se trouve. Nous avons également formalisé les principes éthiques comme des règles établissant des priorités variables entre les différentes règles morales. En effet, là encore, suivant les situations, le respect d'une règle morale peut devenir plus ou moins prioritaire. Ainsi, un principe éthique permet de préciser selon les conditions, quelle est

14 Darren Abramson et Lee Pike. *When formal systems kill: computer ethics and formal methods*. APA Newsletter. 2011.

la relation de préséance qui existe entre les différentes règles morales.

Nous avons ensuite défini un système de transformateurs de prédicats qui permet de convertir un ensemble de règles morales associées à un principe éthique E en une propriété d'invariance qui doit être vérifiée par un agent devant respecter le principe éthique E. Il a également fallu spécifier formellement les agents et leur comportement, afin de pouvoir faire une vérification formelle du respect du principe éthique choisi. Pour cela, nous avons choisi le modèle GDT4MAS qui permet, outre une spécification formelle des agents sous la forme d'un arbre de décomposition de buts en sous-buts, de générer un ensemble de théorèmes (appelés *Obligations de Preuve*) vérifiables par un démonstrateur automatique.

Le modèle en question présente notamment l'avantage d'être associé à une plate-forme qui permet de faire le travail de génération des obligations de preuve et de vérification automatiquement dans la plupart des cas. Cela nous permet ainsi de garantir qu'un agent respectera bien un principe éthique donné dans un contexte d'utilisation particulier, même si cette propriété implique que l'agent ne respecte pas à tout moment toutes les règles morales associées à ce principe éthique.



4. PRÉCONISATIONS AUX CHERCHEURS ET DÉVELOPPEURS

Le cadre développé au cours du projet ETHICAA n'a bien entendu pas la prétention de constituer une solution définitive au problème de l'intégration de l'éthique dans des systèmes d'agents autonomes. Plus que les aspects techniques développés, ce sont les écueils et les réflexions entourant ces aspects techniques qui nous semblent importants. Ainsi, l'expérience du projet ETHICAA nous permet de formuler les préconisations suivantes.

Être intelligible et lisible par l'humain

Si les systèmes techniques automatisés – dans le cas d'agents d'assistance à la personne par exemple – peuvent contribuer au bien-être des personnes, une réflexion sur leurs conditions d'appropriation doit être ouvertement menée. Car comme l'a écrit le philosophe des techniques Gilbert Simondon, l'ensemble des machines doit toujours supposer l'humain comme interprète vivant des machines les unes par rapport aux autres : *« Loin d'être le surveillant d'une troupe d'esclaves, l'est l'organisateur permanent d'une société des objets techniques qui ont besoin de lui comme les musiciens ont besoin du chef d'orchestre*[15] ». Dans une telle perspective, et compte tenu du développement présent et futur des agents autonomes dans nos espaces privés, un enjeu éthique de premier ordre sera de concevoir des interactions où les usagers seront, autant que possible, en situation de participation cognitive et intellectuelle, et où l'exercice du libre-arbitre demeurera aussi entier que possible, en reprenant la main sur les systèmes, en étant capables, par exemple, de neutraliser un système de géolocalisation.

Les réflexions de Norbert Wiener peuvent à cet égard constituer une source importante d'inspiration. Le père de la cybernétique ne manquait pas de souligner le risque de voir se développer une société qui ne serait plus à même de questionner le développement de ses propres inventions, le danger social de la machine ne tenant pas à la machine elle-même mais à l'usage que l'humain peut en faire ainsi qu'au risque d'absence de questionnement que la machine peut induire. C'est en ce sens et au regard de la complexité des interactions qui peuvent se nouer entre les humains et les agents autonomes que ces derniers doivent impérativement répondre à des exigences d'intelligibilité et de lisibilité.

15 Gilbert Simondon, *Du mode d'existence des objets techniques*. Aubier, 1989.

Du danger de la quantification

Dès lors que l'on cherche à modéliser, et *a fortiori* à programmer, des connaissances et des cadres qui permettent de juger, donc à terme de **choisir**, des décisions ou des actions, il peut être tentant – pour des questions de simplicité de mise en œuvre – d'adopter des approches qui visent à ordonner totalement les choix possibles, à pondérer les références (tel cadre éthique est pris en compte à 0,6 et tel autre à 0,2), à définir une conformité numérique d'une décision à une valeur morale, des seuils numériques, etc.

Dans certains cas particuliers, la quantification est pertinente et justifiée : dans le cadre conséquentialiste par exemple, on peut être amené à comparer des conséquences de décisions qui sont par nature chiffrées (un nombre de victimes, un gain financier, etc.)

Cependant de manière générale, un raisonnement prenant en compte des considérations éthiques ne saurait être projeté dans l'ensemble des nombres réels. Une telle tentative dénaturerait le principe même d'un raisonnement prenant en compte des considérations éthiques et altérerait totalement la signification d'un jugement « éthique » d'une décision ou action. En effet, une quantification non fondée sur une réalité objective n'a aucun sens, et le résultat – le choix de la décision donc – ne peut plus être justifié autrement que par le fait qu'il est le meilleur en référence à l'ordre total sur les nombres. Ainsi, même dans le cadre conséquentialiste, les quantifications peuvent mener à de dangereuses comparaisons entre faits incomparables, de par leur réduction à de simples chiffres.

Du besoin de modularité

Un processus de raisonnement s'appuyant sur une représentation explicite des considérations éthiques doit permettre une expression claire et générale de ces considérations. Pour cela, il paraît essentiel d'adopter une démarche modulaire séparant clairement les différents aspects du raisonnement. En particulier, il convient non seulement d'identifier précisément les aspects relevant purement de l'éthique, mais aussi de bien mettre en évidence les éléments sur lesquels ils se fondent, pour mieux en cerner les dépendances. Afin de proposer des mécanismes généraux et adaptables, introduire une modularité selon le niveau de généralité (les règles proposées sont-elles générales, spécifiques à un domaine ou à une situation ?) facilite la réutilisation des modèles dans d'autres applications ainsi que leur remplacement et leur adaptation.

Dans les propositions du projet ETHICAA, le raisonnement sur la fai-

sabilité des actions et leurs conséquences directes, voire leur désirabilité vis-à-vis des objectifs de l'agent, est ainsi effectué en amont du jugement éthique, que ce soit dans un modèle d'action ou un processus d'évaluation, et informe le module spécifique qui s'intègre ainsi comme une couche supplémentaire dans le processus de décision d'un agent. Cela permet de s'appuyer sur des modèles ou architectures classiques (comme le calcul des événements ou l'architecture BDI) qui représentent la dynamique du système et de la prise de décision. Si l'agent considère les ramifications de ses actions, un modèle de causalité permet de plus d'exprimer ce qui découle indirectement de ses actes.

Les architectures proposées séparent de plus le jugement éthique lui-même en plusieurs aspects, différenciant les théories du bien – appréciation directe du caractère bon ou mauvais d'un événement au regard de certaines valeurs ou règles morales – et les théories du juste – qui s'appuient sur ces appréciations pour juger de façon plus globale du caractère juste ou non d'un acte selon un ou plusieurs principes éthiques, en tenant compte du contexte, des préférences entre valeurs, des alternatives envisagées et des conséquences directes ou indirectes. Le processus de jugement peut de plus être complété en articulant ces différents principes éthiques selon des préférences préétablies. Cette modularité assure l'indépendance entre les composants de l'architecture vis-à-vis de chacune des théories d'une part, et entre chacune des théories elles-mêmes d'autre part.

La modularité au sein d'un agent est renforcée par la modularité introduite par chacun des agents d'un système multiagent : chaque agent intègre ses propres théories du bien et théories du juste. Il est donc envisageable de mettre en place au sein d'un système multiagent une pluralité de théories du bien et du juste.

Cette pluralité peut être accrue en envisageant la dimension organisationnelle des systèmes multiagents qui vient structurer les agents et leurs coopérations selon les organisations auxquelles ils appartiennent. Les agents d'une même organisation pourront ainsi intégrer des théories du bien et du juste soutenues par l'organisation à laquelle ils appartiennent. Ces théories seront à intégrer et à concilier éventuellement avec leurs propres théories. Un agent appartenant à plusieurs organisations pourra ainsi également développer différents comportements éthiques selon l'organisation à laquelle il se réfère en cours d'exécution.

Subjectivité des modélisations et importance de la réflexion méthodologique

Si, en sciences sociales, le chercheur et le phénomène qu'il modélise forment un « couple » voire un système dynamique de couples dans lequel de complexes actions, inactions ou rétroactions se mettent en place, peu de travaux de « sciences de l'ingénieur » ont étudié l'impact de la subjectivité du chercheur sur les modèles et algorithmes qu'il produit. Or, dans le cas des algorithmes, la question de la subjectivité des modélisations est ambiguë. D'un côté, les algorithmes peuvent paraître permettre une décision humaine plus objective mais ils sont eux-même soumis à la subjectivité de ceux qui les ont conçus. De plus, le risque de déléguer à une machine des choix qui, au plan éthique, relèvent de la subjectivité humaine est *a priori* problématique.

Dans quelle mesure alors des considérations axiologiques ou éthiques peuvent-elles être mathématisées et programmées ? Il s'agit là de se poser d'une part la question de la modélisation et d'autre part la question du calcul. Modéliser, c'est-à-dire mettre sous une forme mathématique, suppose des hypothèses, des simplifications, des choix, et comprend des biais. Par exemple, quels faits considère-t-on pour décrire une situation ? Qu'est-ce qui permet d'affirmer qu'une conséquence d'une décision est « positive » ? qu'une action est intrinsèquement « bonne » ? quelles conséquences d'une décision considère-t-on ? Comment le calcul est-il optimisé, au détriment de quelles notions ?

De manière plus générale, il convient de se poser les questions suivantes :

- un algorithme qui implique des considérations axiologiques ou éthiques doit-il être calqué sur les considérations axiologiques ou éthiques de l'humain, et si oui de quel humain ? Il faut en particulier remarquer que les attentes que l'on a vis-à-vis d'un algorithme peuvent être très différentes de celles qu'on a vis-à-vis d'un humain qui fournirait le « même » résultat ;
- un être humain peut choisir de ne pas agir de façon morale : jusqu'où traduire cela dans un algorithme, programmer la dérogation aux règles, la transgression ?

Il y a une indétermination de l'éthique qui semble incompatible avec une tentative de preuve ou de certification des algorithmes qui la mettraient en œuvre. De plus, il y a un paradoxe consistant à calquer le raisonnement humain, qui est faillible, dans un algorithme, et vouloir que cet algorithme soit infaillible.

Enfin, il y a certainement un danger à vouloir déléguer l'entendement à un algorithme, et de fonder des décisions uniquement sur le calcul.

Le suivi de patient à domicile

La multiplication et la diversification d'objets connectés déployables dans des environnements domestiques ouvre des perspectives sur le suivi d'activités quotidiennes pour l'assistance ou la supervision. Dans le domaine médical, cela s'applique tout particulièrement au suivi en temps réel des patients, leur prise en charge, la collecte et l'échange d'informations entre patients, soignants et accompagnants. Dans ce cadre, l'utilisation d'agents autonomes, embarqués dans les objets connectés, est pertinente quand il s'agit d'identifier l'état du patient, décider des informations à collecter et à transmettre. Il convient alors d'une part de traiter du patient et des informations associées en respectant le code de déontologie médicale, ce qui soulève une problématique de modélisation de l'éthique médicale. De plus, le maintien à domicile mobilise de nombreux acteurs qui ne sont pas nécessairement affiliés à l'hôpital : médecin de famille, aide à domicile, services de premier secours et, bien entendu, proches au sein du cercle familial. Ces acteurs peuvent être aussi concernés par certaines informations, jugées non pertinentes par le personnel soignant, sans pour autant oublier que le patient a un droit de regard sur l'usage qu'il est fait de ses données personnelles. Il y a donc aussi une problématique du respect de la vie de la privée du patient en fonction du contexte : quelle information pour quel acteur du maintien à domicile ? Enfin, peut se poser la question de la prise d'autorité du patient sur ces informations et décisions. En effet, même si cela n'est pas recommandé, un patient a toujours loisir de mentir à son médecin ou ses proches pour diverses raisons, ou à tout le moins de ne pas suivre les prescriptions qui lui sont faites. Nous pouvons alors nous demander si un système d'agents autonomes conçu pour le maintien à domicile peut assister le patient tout en respectant sa vie privée et son droit à la désobéissance.

Il convient donc d'accompagner la modélisation et la programmation de considérations axiologiques ou éthiques d'une réflexion elle-même éthique afin en particulier de bien comprendre la démarche et d'en identifier les limites.

Prendre en compte la multiplicité des agents et des humains au sein du système

Dans la littérature en éthique computationnelle, la plupart des travaux s'intéressent uniquement à l'éthique à l'échelle du comportement individuel de l'agent, c'est-à-dire dans l'optique de créer des agents autonomes dont le comportement est conforme à des principes éthiques. Or, dans un système multiagent, une simple contrainte de son comportement peut permettre à un agent d'agir individuellement de manière éthique dans un collectif, mais le laisse démuni lorsqu'il doit tenir compte de l'éthique des autres agents. Par exemple, un agent gestionnaire d'investissements financiers pourrait être tout à fait capable de se comporter selon des principes de gestion responsable sans pour autant être capable d'évaluer le caractère éthique du comportement des autres.

L'éthique des individus diffère d'une personne à l'autre pour des raisons culturelles (les valeurs de l'individu proviennent de son entourage familial, social, éducatif,

religieux) et de développement cognitif (capacité à manipuler mentalement des concepts plus ou moins complexes). Outre ces variations dans l'éthique individuelle, différentes éthiques coexistent au sein d'une même société, parfois même au sein d'un même individu. De plus, selon l'École de Francfort, aucune éthique ne peut s'élaborer hors d'une discussion ouverte et contradictoire : seul l'examen par les autres de notre éthique nous permet d'en évaluer la prétention à l'universalité et il ne peut y avoir d'éthique sans s'obliger à se placer du point de vue de tous les autres[16].

Dans ces circonstances, toute approche de mise en œuvre du raisonnement éthique dans des agents autonomes doit prendre en considération cette dimension plurielle des éthiques et ne peut se cantonner à un raisonnement monolithique. Ceci a d'autant plus d'importance dans le contexte actuel de déploiement dans notre environnement d'un nombre croissant d'agents, collaborant entre eux ou avec des humains.

16 Karl-Otto Apel, *Éthique de la discussion*, Traduit de l'allemand par Mark Hunyadi, Éditions du Cerf, 1994.

5. PERSPECTIVES POUR LES CHERCHEURS ET DÉVELOPPEURS

De l'identification du contexte

L'appréhension du contexte est fondamentale pour **apprécier** une situation. L'exemple donné par Paul Scharre du Lieutenant Colonel Stanislav Petrov évitant une troisième guerre mondiale en est l'illustration parfaite[17]. En effet, l'éthique est non seulement dépendante du contexte, mais aussi issue de ce dernier. C'est ainsi que Malle *et al*. proposent « d'activer » les normes nécessaires au jugement éthique en fonction du contexte dans lequel il va s'exercer[18]. Il est donc fondamental lorsque nous parlons d'éthique de définir ce que nous entendons par contexte et d'être capables de l'identifier.

Définition du contexte

Il est difficile de définir de manière exhaustive ce que contient le contexte. Cependant, nous pouvons constater que si *« la situation ne porte pas en elle-même le jugement [éthique]*[19] », le contexte, lui, en porte une part non négligeable. En effet, le contexte ne se limite pas à la situation à un instant donné, mais embarque aussi l'historique de cette situation, et avec lui les normes sociales, valeurs morales, liens et interactions entre parties prenantes, etc. C'est pourquoi il est tout à fait possible de rencontrer des situations similaires ayant des contextes fondamentalement différents.

Il est aussi important de noter que le contexte embarque une notion d'évaluation du temps. En effet, le fait de devoir agir rapidement, ou d'avoir un impact à court, moyen ou long terme est à prendre en compte dans le contexte.

Le contexte est donc un ensemble complexe et évolutif, dépendant du passé, du groupe social, de la place de l'agent au sein de ce groupe, du temps, etc.

Identifier le contexte

D'après la définition précédente, nous constatons qu'identifier un contexte s'approche de la problématique de perception d'un monde ouvert, en y ajoutant une complexité supplémentaire liée entre autres au contre-factuel (ce qui doit être), à l'histoire et à la société dans laquelle nous nous plaçons. Il est donc impossible d'être exhaustif. Nous pouvons alors essayer d'identifier un contexte en nous limitant à l'en-

17 Paul Scharre. *Autonomous weapons and operational risk*. 2016.
18 Bertram Malle, Matthias Scheutz et Joseph Austerweil. *Networks of Social and Moral Norms in Human and Robot Agents*. A World with Robots. 2017.
19 Mark Hunyadi. *Artificial Moral Agents, really ?* 2017.

semble des faits pertinents pour juger la situation.

Notons que l'identification du contexte peut fortement varier d'un agent à l'autre, selon ses perceptions (fiabilité de détection, incertitude sur les faits, etc.) mais aussi selon ses propres normes sociales, ses valeurs, ses buts, etc. La sélection des faits pertinents ajoute une subjectivité supplémentaire à cette identification du contexte.

Le contexte est pratiquement toujours réduit à la situation.

Dans l'ensemble des travaux liés à la modélisation de l'éthique, le contexte est pratiquement toujours réduit à la situation. Cette simplification est un argument supplémentaire à la thèse soutenant qu'un système autonome ne peut avoir un raisonnement éthique à proprement parler, mais peut seulement embarquer des concepts éthiques dans son raisonnement.

Du raisonnement éthique sous contrainte de temps de calcul limité

Selon les domaines d'applications et plus particulièrement lorsque des agents autonomes agissent dans le monde physique, la dimension temporelle doit être prise en compte. Par exemple, les véhicules autonomes doivent être capables de décider de leurs actions en temps réel. Dans un contexte de *trading* haute-fréquence, qui peut aussi être soumis à des principes éthiques, les agents artificiels doivent prendre leurs décisions dans un temps de l'ordre de la milliseconde. Toutefois, les modèles de décision fondés sur du raisonnement automatisé s'exécutent dans des temps considérablement plus longs, et plus encore lorsqu'il s'agit de faire appel à du raisonnement contre-factuel, du raisonnement multiagent ou du raisonnement sur une multitude de principes éthiques. Il semble donc important de pouvoir mettre en œuvre des agents à éthique artificielle pouvant prendre des décisions dans des laps de temps réduits. De manière intéressante, cette problématique introduit une nouvelle forme de dilemme entre le temps passé à calculer une décision « éthique » risquant d'empêcher toute prise de décision effective et le risque de ne pas prendre le temps de calculer une décision « éthique » . Au-delà des aspects d'optimisation de mise en œuvre, deux pistes nous semblent pertinentes :

Modélisation d'un raisonnement éthique anytime

Ici, le mécanisme de raisonnement doit produire une décision initiale qui pourra évoluer au fur et à mesure que l'agent dispose de temps pour raisonner. Parmi les approches possibles, nous pouvons penser à une approche hiérarchique par élagage. Partant d'un ensemble de

principes éthiques hiérarchisés, l'agent élimine toutes les décisions qui ne satisfont pas le premier principe et choisit aléatoirement une décision parmi les restantes ; puis si le temps le permet, l'agent élimine toutes les décisions ne satisfaisant pas le second principe et choisit aléatoirement une décision ; et ainsi de suite. Toutefois, hiérarchiser ainsi des principes éthiques n'est pas une tâche triviale ni forcément bien fondée.

Compilation du raisonnement éthique hors ligne

Ici, les décisions pouvant être prises en fonction du contexte sont calculées au préalable et embarquées au sein de l'agent par exemple sous forme d'arbres de décision. Toutefois, l'agent n'est pas à l'abri de se retrouver dans un contexte imprévu. Dans ce cas, une approche possible serait d'apprendre à l'aide d'un réseau de neurones le résultat du raisonnement éthique afin de le généraliser. Il convient cependant de faire attention ici aux problèmes des cas particuliers qui pourraient ne pas être pris en considération par une généralisation. En effet, l'éthique s'accommode mal des généralisations hâtives.

Certification et éthique d'agents autonomes artificiels

La certification de produit ou de système consiste en une attestation délivrée par une tierce partie indépendante et impartiale que le produit ou le système répond, à un instant donné, à des **exigences prédéfinies**, répertoriées dans un **référentiel**[20]. Cette question de la certification et des problématiques connexes est mise en avant dans les nombreux rapports et plans de recherche nationaux et internationaux parus ces dernières années, en lien avec le développement de systèmes s'appuyant sur des technologies de l'Intelligence Artificielle.

Ainsi, dans le rapport issu de la mission Villani[21], il est question de la conduite d'étude d'impact, d'auditabilité de tels systèmes par *« un corps d'experts publics assermentés, en mesure de procéder à des audits d'algorithmes, des bases de données et de procéder à des tests par tout moyen requis »*.

Le rapport #FranceIA[22] préconise la mise en place de *« l'ensemble de la chaîne d'outils (protocoles, méthodes, simulateurs, environnements de conception et d'intégration, etc.) permettant d'expliquer, de garantir et de certifier les technologies utilisant de l'IA »*. Ces questions corroborées

20 https://www.lne.fr/fr/comprendre/la-certification
21 https://www.aiforhumanity.fr/
22 https://www.economie.gouv.fr/France-IA-intelligence-artificielle

Les véhicules autonomes

L'autonomie des véhicules est classée selon une échelle de 0 à 5 (ou 4 selon l'échelle américaine) sur laquelle nous sommes actuellement entre le niveau 2 (le véhicule est équipé d'une combinaison de services permettant le contrôle de la vitesse et de la voie) et 3 (le véhicule perçoit son environnement et alerte l'humain en cas de problèmes pour qu'il intervienne en quelques secondes). Le niveau le plus élevé concentre les interrogations car c'est le véhicule qui calcule toutes les décisions, le conducteur pouvant dormir, voire ne pas être dans le véhicule. A cause de l'inertie du renouvellement du parc automobile, il y aura pendant de nombreuses années une cohabitation entre véhicules autonomes et véhicules classiques. Ce trafic mixte impose que les véhicules autonomes soient capables de prendre en compte des comportements non rationnels. Considérer que ce problème disparaîtra avec la fin de la mixité est illusoire. L'accident du dimanche 18 mars 2018 provoquant la mort d'une cycliste par un véhicule Uber illustre tragiquement le besoin de répondre à ces questionnements. L'enquête ultérieure a montré que les caméras, les radars et les lidars fonctionnaient correctement au moment de l'accident. La voiture a bien détecté la piétonne qui traversait en tenant sa bicyclette à la main, et cela vraisemblablement mieux qu'une personne ne l'aurait fait. Cependant, pour des raisons liées au confort des passagers, la société Uber avait décidé d'ignorer certains obstacles considérés comme anodins, tels les sacs plastiques ou les feuilles mortes, car ils généraient une conduite chaotique provoquant soubresauts et inconfort. En conséquence, même si la piétonne a bien été détectée par les capteurs de la voiture, les réglages imposés par Uber pour le confort des passagers ont conduit à l'ignorer, puisqu'elle a été assimilée à un obstacle anodin. Nous avons là une tension entre des exigences contradictoires, en l'occurrence ici, conduire rapidement et de façon fluide, et éviter tous les obstacles. Cette tension est intéressante dans le sens où elle est très éloignée du scénario classique fort prisé du grand public et largement imaginaire (le célèbre dilemme du trolley) où la voiture doit choisir entre le sacrifice du passager et celui d'un groupe de jeunes gens insouciants traversant par inadvertance la rue au feu vert.

par les nombreux rapports publiés par des agences ou missions gouvernementales dans le monde, font également l'objet de travaux de standardisation au sein de l'association IEEE[23] au travers de son initiative « Ethically Aligned Design » pour les systèmes autonomes et intelligents. Étant donné les travaux réalisés dans le projet ETHICAA, ces problématiques générales sur l'Intelligence Artificielle concernent les systèmes d'agents autonomes ciblés par le projet. Elles doivent cependant être complétées par la certification de l'éthique qui pourrait être intégrée au sein de tels systèmes. Reprenant la définition de certification du début de ce paragraphe, les questions à considérer sont :

Quel référentiel d'exigences considérer ? Un seul ou plusieurs ?

Il s'agit de conduire une étude d'impact sur les risques éthiques de ces systèmes et d'identifier un ensemble de critères et donc de mesures pour vérifier, certifier ces systèmes vis-à-vis de ces risques éthiques. Au vu de la diversité des éthiques et des utilisations, il est difficile d'envisager un seul et unique référentiel et qui, de plus, soit immuable. De telles études pourraient s'appuyer sur des approches de conception sensible aux valeurs (identification des valeurs que le système intègre et défend, validation et certification que la mise en œuvre du système défend ces valeurs).

Ne pas se limiter aux exigences mais englober aussi le contexte dans lequel elles doivent être satisfaites.

L'éthique d'un comportement ne peut se concevoir qu'en contexte. Il s'agit ici de capter les particularités du contexte et de la situation dans

Qualifier de manière reproductible, répétable et interprétable les solutions et systèmes artificiels éthiques.

lesquels le comportement doit être mis en œuvre et pour lesquels son éthique doit pouvoir être évaluée. Il est donc nécessaire de consigner les situations caractéristiques dans lesquelles les exigences éthiques référencées doivent pouvoir être satisfaites, et quelle sémantique leur donner, car une même exigence éthique peut avoir des interprétations différentes selon le contexte.

Des exigences allant au delà de valeurs et principes éthiques.

Les systèmes artificiels considérés sont des systèmes à forte autonomie décisionnelle. Il est donc essentiel que les référentiels incluent également des exigences non fonctionnelles telles que validité, robustesse, passage à l'échelle, etc. L'éthique des comportements mis en œuvre par ces systèmes doit en effet pouvoir être certifiée vis-à-vis de telles

23 http://standards.ieee.org/news/2017/ead_v2.html

exigences. Que ferions nous en effet d'un système ayant un comportement dont l'éthique ne résiste pas au passage à l'échelle ?

Auditer, tester et vérifier que ces systèmes artificiels satisfont aux critères.

Tout comme pour les systèmes d'Intelligence Artificielle, l'estimation de la qualité de l'éthique de systèmes d'agents autonomes devra pouvoir s'effectuer à l'aide de protocoles de mesure dans un cadre commun fondé sur la reproductibilité, la répétabilité, et l'estimation de la justesse. Il s'agit de comparer des approches dans les mêmes conditions sur une tâche précise via des tests et des expérimentations, des simulations. Il s'agit de définir des cadres permettant de qualifier de manière reproductible, répétable et interprétable les solutions et systèmes artificiels éthiques.

Accéder au code source des agents autonomes ?

Dans le cadre du projet ETHICAA, nous avons montré qu'une approche formelle permettait de vérifier effectivement si le comportement d'agents autonomes, dans un contexte donné, vérifiait une éthique particulière. Cela impose toutefois de devoir connaître exactement le comportement des agents. Il faudra donc, comme dans la plupart des autres processus de certification, que les organismes de certification puissent avoir accès à ce genre de données, soit grâce à des clauses de confidentialité protégeant le travail des concepteurs, soit, et c'est certainement ce qui apportera le plus de confiance en de tels systèmes, en rendant le code des agents en question « open source ». Cette solution, qui a par exemple été retenue pour donner plus de crédit au logiciel Parcoursup qu'à son prédécesseur APB, ne pourra cependant pas toujours l'être, notamment pour des raisons liées au secret industriel.

Compatibilité des éthiques de l'agent, de l'utilisateur, du concepteur ?

Comment certifier et s'assurer de la conformité des valeurs et règles morales, des principes éthiques représentés au sein de l'éthique mise en œuvre par l'agent autonome, avec ceux souhaités par l'utilisateur et ceux que le concepteur du système souhaitait donner aux agents ? Est-ce que les valeurs et règles morales, les principes éthiques dont le concepteur a voulu doter le système sont explicites ? Sont-ils compatibles avec ceux de l'utilisateur ?

Certification dans un cadre d'évolution de l'agent autonome ou des conditions de déploiement, d'exécution, d'interaction.

La certification pourra-t-elle prendre en compte les conditions de déploiement du système avec notam-

ment les interactions avec d'autres produits, que ces interactions soient prévues ou non ? La certification pourra-elle prendre en compte l'évolution du système artificiel lorsque, par exemple, celui-ci est équipé de capacités d'apprentissage automatique ?

CONCLUSION

Au cours de notre parcours de recherche, le projet ETHICAA a produit un cadre de raisonnement éthique fondé sur une approche symbolique de l'Intelligence Artificielle pour permettre à un agent autonome d'évaluer son environnement, d'intégrer des principes éthiques et de déterminer, à partir de la mise en œuvre de ces principes, soit un plan d'actions, soit une évaluation du comportement d'autres agents. Au-delà de ces réalisations techniques et des recommandations que nous avons formulées, nous avons pu nous rendre compte qu'il importe de pouvoir s'entendre sur certaines valeurs afin de stimuler une évaluation des innovations technologiques dans les contextes où elles sont utilisées.

Aucune éthique ne peut s'élaborer indépendamment d'une discussion ouverte et contradictoire.

Dans une telle perspective, il convient de souligner la nature intrinsèquement intersubjective de l'engagement éthique : aucune éthique ne peut s'élaborer indépendamment d'une discussion ouverte et contradictoire (en l'occurrence entre chercheurs, concepteurs, développeurs, etc.) Il n'y a d'éthique qu'en assumant la confrontation des argumentations qui oblige chacun à se placer du point de vue de tous les autres.

Un tel espace de discussion nous est apparu essentiel dans l'exercice d'évaluation des questions d'éthique qui sont induites par les agents autonomes.

Mais le défi éthique est entier dans une époque hypermoderne où nous sommes, d'une part, dans une relation plutôt consumériste aux nouveautés technologiques (tout ce qui est nouveau est de ce fait massivement considéré comme allant dans le sens du progrès) et où, d'autre part, la majorité des usagers est loin de disposer de clés de compréhension nécessaires au déchiffrage de leurs environnements hyper-technologiques (objets connectés, agents autonomes, puces RFID, biométrie, etc.) qui apportent souvent dans la vie de tous les jours un plus grand confort. Or, comme l'exprimait déjà Herbert Marcuse, plus une certaine « commodité » nous est technologiquement garantie, plus elle se voit induite par la normalisation de nos pratiques technologiques, moins le champ des interrogations que nous serions en droit de formuler vis-à-vis de toute innovation technoscientifique (en termes de sens, de valeurs, de qualité du lien social, etc.) est intense et profond. Se donner pour tâche « d'orienter le présent vers un

avenir durable » suppose donc de travailler sans relâche à l'éveil d'une démarche critique constructive à l'ère hypermoderne.

Il convient de s'interroger sur le sens de l'agir humain, en tenant compte du fait que toute action échappe de plus en plus à la volonté de son auteur à mesure qu'elle s'inscrit dans le jeu de rétroactions de l'environnement où elle intervient. C'est ce qu'une écologie de l'action montre clairement : dans la pratique, l'intention risque le plus souvent de se traduire par un échec dans la mesure où les effets de l'action dépendent non seulement des intentions de celui qui agit, mais aussi des contextes où l'action se déroule. Or, c'est bel et bien la multitude (et la complexité) de ces interactions qu'une éthique des agents autonomes devra être en mesure d'embrasser.