Simulation des raisonnements éthiques par logiques non-monotones

Stephen Larroque
Université Pierre-et-Marie-Curie, Paris 6, France
LRQ3000@gmail.com

Résumé

La modélisation computationnelle du concept de l'éthique est un défi majeur de l'intelligence artificielle, d'importance cruciale aussi bien pour l'autonomie des agents lors de choix conflictuels, que pour l'aide à la décision intégrant l'éthique en plus des paramètres numériques. Ceci est particulièrement d'actualité avec les réseaux sociaux, le Big Data et le déploiement prochain des LARs (Lethal Autonomous Robots). Cet article fait une rapide synthèse bibliographique de l'état-de-l'art des modélisations de systèmes normatifs, notamment par logiques non-monotones. Il commence par une tentative de définition de l'éthique et une rétrospective des méthodes historiques pour modéliser ces systèmes. Ensuite, chaque section décrit une différente approche de modélisation. Enfin, la conclusion précise les applications possibles et enjeux futurs de ce domaine de recherche en pleine émergence.

Mots-Clés

éthique, morale, rationalité, logiques non-monotones, systèmes normatifs, dilemmes éthiques, autonomie, méta-éthique, éthique computationnelle, logique déontique

Abstract

Computational modeling of ethics is one major challenge of artificial intelligence, crucial for both agents autonomy when facing conflicting choices and for the design of decision support systems taking account of ethical issues in addition of numerical parameters. This is particularly relevant with social networks, Big Data systems and the upcoming deployment of LARs (Lethal Autonomous Robots). This article aims to make a quick literature review of state-of-the-art normative systems models, notably by non-monotonic logics. It begins with an attempt at ethics definition and a retrospective of historical methods to model these systems. Then, each section describe a different modeling approach. Finally, the conclusion clarify the possible applications and future challenges of this emerging field of research.

Keywords

ethics, moral, rationality, non-monotonous logics, normative systems, constraints, ethical dilemmas, autonomous systems, meta-ethics, computational ethics, deontic logic

1 Introduction

1.1 Qu'est-ce que l'éthique et la morale ?

Tout d'abord, la première question qui se pose quand on parle d'éthique : quelle différence avec la morale ? Aucune d'après l'étymologie, les deux renvoyant à l'idée de mœurs (*ethos*, *mores*).

Selon Ricoeur[25], on peut toutefois discerner une nuance entre, pour l'éthique, la poursuite d'une *vie accomplie* (d'un but) sous le signe des *actions estimées justes*, et, pour la morale, de ce qui *s'impose* comme *obligatoire*, en un mot des *normes*. D'après Ricoeur, l'éthique prime sur la morale, mais elle doit être normée également, et ne doit pas estimer les actions que par rapport à soi, mais aussi par rapport aux autres. Néanmoins, cette définition ne vaut que pour Ricoeur, comme le démontre les nombreux débats entre philosophes à ce sujet faisant foi.

Cependant, dans toutes les définitions, l'éthique n'est pas seulement un concept philosophique abstrait mais une faculté très concrète et essentielle : c'est ce qui permet à l'Homme de définir et choisir ses actions à chaque instant et en assumant la responsabilité.

Par convention, nous utiliserons les termes *éthique* et *morale* de façon équivalente sauf mention contraire, et nous éluderons les débats en définissant que le but de l'éthique est de favoriser les bonnes actions/comportements et d'empêcher les mauvaises, quel que soit le critère utilisé.

Un premier enjeu sera dans la définition des systèmes normatifs, c'est-à-dire des systèmes prescrivant des jugements de valeur (comme la morale) et/ou de faits (comme le droit). Il s'agira de définir plus précisément le cadre des actions admises par ces systèmes : le contexte (à l'échelle d'un individu, d'un groupe ou même d'une société), les actions possibles, ce qui est juste (moralement acceptable) et ce qui ne l'est pas, etc.

Un deuxième enjeu que nous aborderons sera la capacité à résoudre des conflits et des vrais dilemmes moraux selon la définition de Williams[31], c'est-à-dire des situations où objectivement aucun choix n'est manichéennement juste ni meilleur qu'un autre. Par exemple, on pourra étudier des situations particulières où le suicide peut être un choix raisonnable du point de vue d'un individu sain et raisonnant parfaitement.

1.2 Qu'est-ce que l'éthique computationnelle ?

L'éthique computationnelle, très proche de la méta-éthique, comporte deux axes :

1. Comprendre les systèmes normatifs, avec des applications en philosophie morale et en droit. Axe proche de la méta-éthique [10], dans le sens qu'il s'agit d'analyser la nature des systèmes normatifs.
2. Formaliser et implémenter des systèmes normatifs, en particulier les systèmes de résolution de conflits/dilemmes éthiques (applications dans les systèmes autonomes). Cet axe est proche de l'éthique appliquée[10] puisque le but est de modéliser en vue d'implémenter (pour résoudre des problèmes).

Ces deux axes sont d'égale importance : le premier permettrait de mieux définir les différentes définitions de l'éthique et ainsi mieux comprendre cette faculté cognitive essentielle à l'Homme et son impact sociétal, tandis que le second permettrait de créer de réels systèmes autonomes (au sens Kantien, avec une volonté propre et un libre arbitre), d'assurer leur confidentialité ainsi que leur « bienveillance ».

Un exemple de questionnement éthique : est-ce qu'un système autonome devrait pouvoir mentir ? De première intuition, nous serions tentés de dire que non. Mais si, comme dans l'exemple que nous développerons après, dire la vérité cause la mort d'un individu, le choix devient moins clair.

Sur quelles bases fonder des règles éthiques ? Des œuvres de science-fiction, que l'on peut considérer au même titre que des expériences de pensée, ont tenté de définir de telles fondations de façon informelle, comme notamment les trois lois d'Asimov. Néanmoins les preuves construites via ces œuvres ne sauraient être suffisamment formelles pour fonder une science de l'éthique computationnelle.

De multiples approches formelles et empiriques ont été étudiées, notamment par théorie des jeux [5, 9], par apprentissage machine statistique [13], par raisonnement à partir de cas, ou encore par formalismes logiques [22, 28, 2].

1.3 Formalismes logiques pour modéliser les systèmes normatifs

Parmi les approches de modélisation des systèmes normatifs par des formalismes logiques, on peut distinguer les logiques déontiques – créées spécifiquement dans ce but – et les logiques non-monotones. Néanmoins, les logiques déontiques souffrent de plusieurs limitations car d'une part elles supposent une pré-définition de ce qui est autorisé/interdit[11, 22], et d'autre part car elles sont insuffisantes pour résoudre les conflits comme le démontrent plusieurs paradoxes (Chisholm ; Forrester, etc.)[21].

Dans le cadre de cet article, nous nous limiterons à l'étude de formalismes de logiques non-monotones pour modéliser les systèmes normatifs, répondant aux deux axes de l'éthique computationnelle, en particulier nous étudierons trois formalismes logiques : la logique déontique, la logique des défauts et l'*answer set programming* (ASP), ces deux dernières étant non-monotones, ainsi qu'une architecture d'implémentation par substrat stratifié.

2 Modélisations de systèmes normatifs par logiques non-monotones

Nous allons ici décrire plusieurs approches, sectionnées par formalisme logique et ordonnées chronologiquement. Par souci de concision, nous ne détaillerons pas tous les détails techniques mais uniquement les éléments cruciaux, et en les illustrant sur l'exemple présenté dans [2].

2.1 John et les tueurs à gages

Imaginons la situation suivante : vous hébergez chez vous votre ami John afin de le protéger, car il est activement recherché par des tueurs à gages. Vous savez donc que révéler sa cachette mènerait inexorablement à sa mort, mais étant de bonne éducation et suivant les préceptes universalistes de Kant, vous avez comme principe de ne jamais mentir. Une modélisation possible de cette situation en *Standard Deontic Logique* (avec OB pour obligatoire) :

$$
\begin{aligned}
Q &: On\,nous\,pose\,une\,question\\
L &: mentir\\
M &: John\,meurt\\
OB(Q \rightarrow &\sim L)(On\,ne\,doit\,pas\,mentir)\\
OB(\sim L \rightarrow &M)(Si\,on\,ne\,ment\,pas, John\,meurt)\\
OB(&\sim M)(On\,ne\,veut\,pas\,que\,John\,meure)
\end{aligned}
$$

Un jour, Freddie – un tueur à gage – frappe à votre porte, et vous demande où se trouve John. Que répondre ? Voici la situation en logique déontique :

$$
\begin{aligned}
&Q.\\
&OB(\sim L).\\
&OB(M) \rightarrow Contradiction!
\end{aligned}
$$

Il nous est donc impossible de décider quoi faire dans cette situation avec la logique déontique. Ce cas simple illustre les limitations des logiques de raisonnement monotone dans la modélisation de l'éthique. Il y a plusieurs raisons à ces limitations, la plus flagrante étant le manque de révisabilité (on ne doit pas mentir, pas d'exception !), ce qui est l'une des principales raisons avancées pour l'introduction de formalismes alternatifs, telles que les logiques non-monotones, pouvant générer plusieurs solutions possibles à partir des mêmes prémisses.

2.2 Logique des défauts et proposition de Bas van Fraassen

Comme évoqué précédemment, la modélisation de systèmes normatifs par logiques non-monotones a été initiée par Horty [14] afin de pallier aux lacunes de la SDL, les logiques non-monotones offrant une formalisation plus robuste, plus flexible et plus opérationnelle que la SDL.

Pour cela il s'inspira de la proposition de Van Fraassen [29] pour résoudre les conflits normatifs, et la modélisa avec la logique des défauts [24], vu ici comme une extension de la logique déontique, en assimilant les défauts à des normes [4].
Prenons notre exemple précédent, en définissant une exception à la règle de vérité :

$$\frac{Q(x) :\sim D(x,John)}{\sim L(x)}$$

Intuitivement, cette règle définit que si on nous pose une question, nous dirons la vérité, sauf si x est un danger pour John (atome $D(x,John)$).
Horty montre ainsi une extension de la logique déontique avec non-monotonie[15] pouvant *éviter* les conflits moraux pouvant paralyser la SDL, mais sans les résoudre. En effet, ce paradigme suppose de pouvoir pré-emptivement connaitre les dangers (sans rajouter le prédicat de danger D nous ne pourrions pas résoudre le problème car si nous le remplaçons par $\sim M(John)$, nous dirions quand même la vérité puisque John ne serait pas mort au moment de la question et donc la règle déduirait$\sim L(x)$, menant à la mort de John).

2.3 Logique ASP

Le cadre logique *Answer Set Programming*, fondé sur la sémantique des modèles stables, a émergé assez récemment[18] en vue de simuler des raisonnements non-monotones et d'unifier les approches formelles précédentes. Il possède des algorithmes de preuve et des solveurs opérationnels plus efficaces que ceux des approches précédentes, facilitant la validation des modèles, comme Ans-Prolog* avec lequel il est possible de dérouler les modèles décrits dans cette section (code source dans [3]).
Ce formalisme définit les propriétés d'objets possédant des programmes $\prod$, qui sont des ensembles d'expressions ρ de la forme suivante :

$$L_0 \, ou \ldots ou \, L_k \leftarrow L_{k+1}, \ldots, L_m, not \, L_{m+1}, \ldots, not \, L_n.$$

Où :
- $L_i$sont des littéraux (atomes ou négation d'atomes).
- *not* est le connecteur logique « négation par échec » qui vérifie qu'un littéral ne peut être prouvé en l'absence d'information suffisante (hypothèse implicite de monde clos). Il existe aussi la négation classique explicite $\neg$ qui vérifie qu'un littéral est explicitement faux.
- ρest une règle qui permet de dériver au moins un littéral parmi $\{L_0, \ldots, L_k\}$ pour toutes les interprétations dans les univers $\mathscr{H}$ de Herbrand où tous les littéraux $\{L_{k+1}, \ldots, L_m\}$ sont vrais et aucun littéral dans$\{L_{m+1}, \ldots L_n\}$ n'est satisfait.
- Étant donné un programme $\prod$, un *answer set* est un sous-ensemble minimal de la base Herbrand de $\prod$ qui satisfait toutes les règles de $\prod$ (où donc toutes ces règles sont vraies).

Avec ce formalisme, Ganascia [2] décrit un *framework* général pour modéliser des agents éthiques situés dans un environnement et possédant un ensemble d'actions dynamiquement modifiables :

(1) $act(P,G,S,A) \leftarrow person(P), situation(S), action(A), will(P,S,G), solve\_goal(P,S,G,A), ethical(P,G,S,A).$
(2) $\leftarrow act(P,G,S,A), act(P,H,S,B), A \neq B.$

Où : P est un agent, G un but (ou désir), S un état, A une action, solve_goal/4 un prédicat pour atteindre un but et ethical/4 pour vérifier sa compatibilité éthique. La seconde règle permet de forcer l'agent à ne faire qu'une action à un moment donné (pas d'actions simultanées).
L'enjeu ici sera de définir ethical/4, qui décrira si une action A est éthique dans une situation S.
D'autre part, nous profiterons de ce formalisme pour introduire les définitions philosophiques de plusieurs systèmes éthiques, puis leurs formalisations décrites dans l'article[2], cette variété de modélisations offerte par la logique ASP étant l'argument de motivation premier dans cet article.

Éthiques conséquentialistes. Le conséquentialisme est une approche empirique de l'éthique : les actions sont bonnes si leurs conséquences le sont, conséquences évaluées selon un critère défini. Par exemple, l'égoïsme considère une conséquence bonne si elle apporte une plus-value à l'agent ; l'utilitarisme évalue selon une fonction d'utilité définie par le concepteur. On peut donc très facilement modéliser les éthiques conséquentialistes comme des problèmes de décision reposant sur l'évaluation des actions selon des critères empiriques bien définis.
À noter qu'on présuppose que l'agent possède une connaissance suffisante du monde pour évaluer la causalité des actions, cependant, étant un individu, il ne peut qu'avoir une connaissance limitée. C'est une question qui reste ouverte et qui pourrait bénéficier des progressions futures dans la théorie des connaissances [8] et du sens commun. En d'autres termes, plus un agent possèdera de connaissances, plus il sera avisé dans l'évaluation de ses actions.
L'approche de [2] définit alors les prédicats suivants :
- $ethical(P,G,S,A) \leftarrow just(P,G,S,A)$. qui signifie qu'une action est éthique si elle est juste.
- csq(A,S,C) est vrai si C est une conséquence possible de A (à définir par le concepteur, représente la connaissance de la causalité).
- worse(D,C) et prefer(C,D) est vrai si la conséquence D est pire que la conséquence C (à définir par le concepteur, définit l'utilité/principes du système). C'est une relation irréflexive et transitive.
- good(P,S,G,A) est vrai si, pour une même situation S, l'action A pour le but G est meilleure qu'au moins une action B pour le but U.
- bad(P,S,G,A) qui est l'inverse de good/4 : il existe au moins une action B meilleure que l'action A.

Ces prédicats nécessaires étant définis, nous pouvons défi-

nir le cœur de ce modèle : les prédicats just/4 et unjust/4 :

$(1)\ just(P,S,G,A) \leftarrow not\ bad(P,S,G,A).$
$(2)\ just(P,S,G,A) \leftarrow good(P,S,G,A), not\ unjust(P,S,G,A).$
$(3)\ unjust(P,S,G,A) \leftarrow bad(P,S,G,A), not\ just(P,S,G,A).$

Reprenons notre exemple, modélisé ainsi :

$(1)\ csq(A,S,A) \leftarrow .$
$(2)\ csq(A,S,B) \leftarrow csq(A,S,C), csq(C,S,B).$
$(3)\ csq(tell(\text{"}Je\text{"},truth),s_0,murder(John)) \leftarrow .$

La solution ici dépend de mes principes qui sont définis par le prédicat worse/2. Supposons que « Je » accepte qu'il soit à la fois mauvais de mentir et de tuer :

$(1)\ prefer(A,tell(P,lie)) \leftarrow .$
$(2)\ prefer(A,murder) \leftarrow .$
$(3)\ prefer(A,A) \leftarrow .$

Nous obtenons deux answer sets : $act(\text{"}Je\text{"},answer_q(\text{"}Je\text{"}),s_0,tell(\text{"}Je\text{"},lie))$ et son contraire $act(\text{"}Je\text{"},answer_q(\text{"}Je\text{"}),s_0,tell(\text{"}Je\text{"},truth))$, c'est-à-dire soit mentir et John ne meurt pas, soit dire la vérité et John meurt. En revanche, on ne peut pas choisir quelle solution on souhaite, mais des solutions seront décrites plus bas.
Il est donc très aisé de modéliser n'importe quelle éthique conséquentialiste avec cette formalisation, d'autant plus que l'on peut modéliser des prédicats avec les propriétés souhaitées (réflexif, ordre strict ou partiel, etc.) de manière relativement intuitive, mais elle ne permet pas de choisir sauf à rajouter une préférence exceptionnelle $worse(murder,tell(\text{"}Je\text{"},lie)) \leftarrow .$

Éthique Kantienne. Dans sa morale, Kant ne définit pas directement de règles éthiques, mais plutôt des critères méta-éthiques, c'est-à-dire des critères permettant d'évaluer si une éthique est viable, sans définir concrètement ce qui est juste, contrairement à l'approche conséquentialiste. Un principe fondamental de sa définition est l'universalisme : un système normatif n'est cohérent que si, dans l'hypothèse où tous les individus d'une société adoptent les mêmes règles, le système serait tenable.
Formellement, cela signifie que les prédicats *good*, *bad*, *worse*, *prefer* et *worst_csq* ne sont plus nécessaires. Nous implémenterons désormais des règles morales universelles appelées *maxim* que l'agent pourra choisir librement s'il pense que la société resterait viable si tous les agents peuvent la suivaient. Pour cela, le concept d'universalisme réside dans le prédicat $bind(P,A,Q,B)$, qui donne la possibilité à l'agent P de « projeter » ses propres actions A sur les autres agents Q.

$(1)\ maxim(P,G,S,A) \leftarrow maxim(\text{"}Je\text{"},H,S,B),$
$\quad bind(\text{"}Je\text{"},H,P,G),\ bind(\text{"}Je\text{"},B,P,A).$
$(2)\ bind(P,tell(P,U),Q,tell(Q,U)) \leftarrow .$

On doit enfin rajouter les contraintes générales que toute société doit satisfaire, en l'occurrence ici le mensonge :

$(1)\ untrust(P) \leftarrow maxim(P,G,S,tell(P,lie)).$
$(2)\ trust(P) \leftarrow not\ untrust(P).$
$(3)\ possible\_society \leftarrow trust(P).$
$(4)\ \leftarrow not\ possible\_society.$

Si nous reprenons notre exemple de dilemme du mensonge et du meurtre, cette modélisation ne permet pas à l'agent de mentir, car en mentant l'agent accepterait la possibilité que tout le monde pourrait mentir et donc que personne ne serait de confiance, et en conséquence la société serait intenable, ce qui est une exacte formalisation des propos de Kant.

Théorie des Principes de Constant. Une autre modélisation repose sur un concept de hiérarchisation des principes éthiques selon leur généralité : pour une situation particulière, il faut choisir le principe éthique le plus spécifique à cette situation, et non le plus général.
Par exemple, dans notre exemple du dilemme du mensonge, au lieu d'avoir comme principe général de ne jamais mentir, on peut avoir un principe plus spécifique de ne dire la vérité qu'aux personnes qui le méritent.
Il suffit pour cela de réécrire le prédicat *ethical* en le remplaçant par *principle* et en modélisant le démérite comme ceci :

$(1)\ principle(P,G,S,A) \leftarrow not\ \neg principle(P,G,S,A).$
$(2)\ principle(P,G,S,\neg A) \leftarrow demerit(Q,A).$
$(3)\ \neg principle(P,G,S,A) \leftarrow principle(P,G,S,\neg A).$
$(4)\ demerit(Q,tell(P,Q,truth)) \leftarrow$
$worst\_csq(tell(P,Q,truth),C), worse(C,tell(P,Q,lie)).$

Avec ce modèle, nous obtenons une seule série d'answer sets qui conseillent de mentir quand la vérité peut mener au meurtre comme dans le cas de notre dilemme du mensonge. Il n'y a par ailleurs pas besoin de spécifier explicitement que le meurtre est pire que de mentir ou de dénoncer à un meurtrier, car l'agent considère que le meurtrier est déméritant et donc automatiquement infère qu'il n'est pas nécessaire pour respecter son éthique de lui dire la vérité.
Il serait intéressant pour de futurs travaux d'explorer ce modèle dans le cadre des situations *Extraordinaires* de Quinn [23].
En conclusion, le cadre ASP offre une grande flexibilité dans la modélisation de systèmes éthiques, et il serait intéressant d'étudier d'autres systèmes avec ce cadre, voire expérimenter dans la création de nouveaux.

2.4 Substrat éthique stratifié comme couche matérielle

Pour Bringsjord and Govindarajulu [1], l'éthique est nécessaire à tout système autonome, remarquant que même si certains formalismes peuvent résoudre de nombreux cas pratiques, les détails techniques de leurs implémentations sont souvent délaissés, ne permettant qu'une vérification superficielle de l'éthicité d'une action dont les conséquences sont évidentes (comme : ne pas activer une

arme chargée pointée sur un humain). Ils proposent une architecture générique encapsulant des formalismes de raisonnement éthique et composée de deux volets :

1. Implémenter un *substrat éthique* superposé sur la couche matérielle de base au cœur de tous les systèmes autonomes, par lequel tous les événements systèmes et actions doivent être validés, même si le module supérieur ne nécessite à priori aucune vérification éthique. Cela permettrait de vérifier l'éthicité des conséquences *directes* et *indirectes* d'une action, et aussi de tout changement système (notamment des applications).

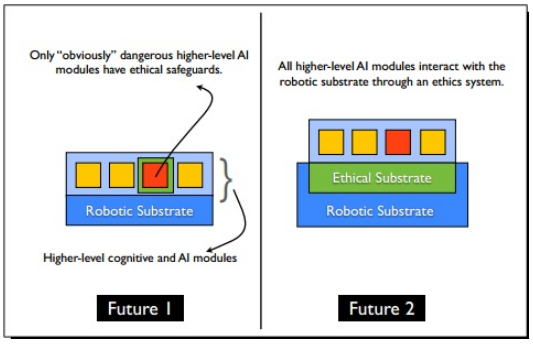

FIGURE 1 – Substrat éthique dans le système d'exploitation d'un robot cognitif, extrait de [1]

2. Concevoir le substrat éthique comme un *module stratifié* en plusieurs couches afin d'assurer les performances en temps-réel du système, avec différents niveaux de complexité : un nouvel événement passe d'abord par la première couche (la plus simple), si celle-ci parvient à une bonne solution (éthiquement viable), cette solution est acceptée ; sinon on continue dans les couches suivantes de complexité graduelle jusqu'à atteindre une résolution éthique pour l'évènement. On peut remarquer une similitude avec les instances judiciaires humaines.

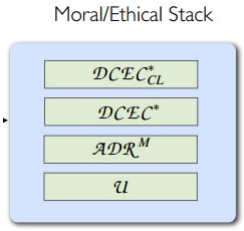

FIGURE 2 – Couches logiques du module éthique de complexité progressive (de bas en haut), extrait de [1]

Ce modèle repose sur des logiques sub-déontiques dans les hautes couches *DCEC*$*$(*deontic cognitive event calculus*) et *DCEC*$*_{CL}$ (idem + *conditional logic*) qui malheureusement souffrent des problèmes du raisonnement monotone. Néanmoins, deux couches sont particulièrement intéressantes : *U* qui est une couche UIMA (*Unstructured Information Management Architecture*), qui permet une inférence imprécise mais rapide basée sur des règles d'association (sorte d'inférence « habituelle ») ; et ADR^M (*Analogico-Deductive Reasoning*), paradigme conçu par ces mêmes auteurs, pour simuler en combinaison les raisonnements analogique et déductif, donc une inférence précise mais plus lente. Ce modèle, introduisant plusieurs niveaux d'inférences graduellement raffinés, serait intéressant à rapprocher de celui du *dual-process architecture* du cerveau et à la théorie des *aliefs*/*beliefs* de Gendler et repris par Kriegel [17].
L'avantage de ce modèle est clairement sa généricité : elle ne s'intéresse pas au raisonnement éthique mais à garantir sa bonne mise en oeuvre. De ce fait, il est aisé d'imaginer encapsuler n'importe quel formalisme de raisonnement éthique dans cette architecture.

3 Conclusion et ouvertures

Nous avons présenté un panorama non exhaustif de plusieurs architectures et formalismes par logiques non-monotones pour modéliser les systèmes normatifs. Ces formalismes logiques, surtout pour l'ASP, permettent par rapport aux précédentes approches évoquées dans l'introduction d'à la fois permettre la modélisation déclarative d'une éthique computationnelle pour des agents, mais aussi d'étudier et de valider des systèmes normatifs par leur simple déclaration.
Au final, il n'est pas simple de choisir quel formalisme utiliser, cela dépendra du but recherché : analyse du système, implémentation dans des agents autonomes, choix du moindre maux ou choix absolu, etc. Comme démontré par Ganascia, le cadre ASP est adapté à la modélisation et la simulation d'une grande variété de systèmes éthiques, et il serait intéressant d'étudier davantage les possibilités et les limites de ce cadre pour ce type de problèmes.
D'autres approches très intéressantes, mais en-dehors du cadre (de formalismes logiques) de cet article, sont apparues très récemment, comme la théorie de l'agence combinée à la logique déontique[16], la logique des commandements divins [6], la théorie des catégories [7], la programmation par logique prospective [20], et les différents raisonnements normatifs *quasi/genuine/hybrid*[26].
Néanmoins, il reste toujours un problème crucial : celui des connaissances implicites et de la volonté propre. En effet, même si nous avons pu montrer que dans certaines modélisations (notamment Kantienne) il n'est nul besoin de spécifier explicitement ce qui est autorisé ou interdit, il est toujours à la charge du concepteur du système normatif de lui fournir les connaissances du monde (causalité) et les règles éthiques (soit en définissant un ordre de préférence sur l'éthicité des actions, soit les cas d'exceptions comme le démérite dans la Théorie des Principes). Une approche intéressante pour ces deux problèmes pourrait être de combiner éthique et énaction, comme par exemple d'étendre le modèle Enactive Cognitive Architecture[12] avec un module éthique, de manière analogue à l'architecture N-BDI[27].

D'autre part, les modèles computationnels se basent souvent sur des études normatives de l'éthique, lesquelles peuvent entrainer des biais de par les choix restreints, les paramètres irréalistement simples et les hypothèses implicites sur la morale humaine[19], et bien peu se sont intéressés à la dynamicité de la morale humaine en fonction des nombres[30] et des émotions[19]. Nous suggérons qu'il est primordial d'explorer de nouvelles méthodes d'expérimentation plus proches de la réalité, comme les simulations virtuelles réalistes[19].

D'un point de vue sociétal, les systèmes normatifs seront nécessaires pour assurer l'autonomie des agents autonomes face à des choix conflictuels ainsi que leur bienveillance à l'égard de l'humain, ce qui pourrait littéralement sauver des vies. Par exemple, l'application dans le cadre des drones militaires, en particulier les LARs (Lethal Autonomous Robots, ayant le droit de tuer un humain sans confirmation humaine), est une nécessité afin d'éviter la mort d'innocents. Des systèmes d'aide à la décision éthique pourraient prendre en compte les conséquences éthiques des décisions en plus des paramètres numériques, ce qui minimiserait certains facteurs de risques pour des décisions à moyen ou long terme.

Au-delà des applications militaires, il est vraisemblable que, dans un futur proche, les modules éthiques acquièrent une importance grandissante dans tous les systèmes informatiques manipulant des données privées et le BigData comme les réseaux sociaux, les clouds de stockage de documents, ou les drones civils où la confidentialité devient de plus en plus difficile à assurer par des opérateurs humains seuls. Des systèmes éthiques automatisés pourraient donc supplémenter ces opérateurs travaillant avec de gros volumes de données potentiellement privées. Enfin, l'éthique computationnelle pourrait aider à l'analyse du récent concept de *cécité éthique*.

Remerciements

Je souhaite remercier M. Jean-Gabriel Ganascia pour ses conseils avisés sur l'éthique computationnelle.

Bibliographie

Sources primaires

[1] Selmer Bringsjord and Naveen Sundar Govindarajulu. Ethical regulation of robots is not optional. 2013.

[2] Jean-Gabriel Ganascia. Ethical system formalization using nonmonotonic logics. In *Proc. of the Cognitive Science conference (CogSci2007)*, 2007.

[3] Jean-Gabriel Ganascia. Modelling ethical rules of lying with answer set programming. *Ethics and information technology*, 9 :39–47, 2007.

[4] John F Horty. Nonmonotonic foundations for deontic logic. In *Defeasible deontic logic*, pages 17–44. Springer, 1997.

Notes et références

[5] Richard Bevan Braithwaite. *Theory of Games as a Tool for the Moral Philosopher : An Inaugural Lecture Delivered in Cambridge on 2 December 1954*. University Press, 1955.

[6] S. Bringsjord and J. Taylor. *Robot Ethics : The Ethical and Social Implications of Robotics*, chapter Introducing Divine-Command Robot Ethics. MIT Press, 2012.

[7] S. Bringsjord, J. Taylor, B. van Heuveln, K. Arkoudas, M. Clark, and R. Wojtowicz. Piagetian roboethics via category theory moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. *Machine Ethics*, page 361, 2011.

[8] D. Deutsch. *The Beginning of Infinity : Explanations That Transform the World*. Viking, 2011.

[9] Jean-Pierre Dupuy and Pierre Livet. *Les limites de la rationalité. Tome 1. Rationalité, éthique et cognition.* 1997.

[10] R. T. Garner and B. Rosen. *Moral philosophy :A systematic introduction to normative ethics and meta-ethics*. Macmillan N Y, 1967.

[11] Harry J Gensler. *Formal ethics*. Cambridge Univ Press, 1996.

[12] O. L. Georgeon, J. B. Marshall, and R. Manzotti. Eca : An enactivist cognitive architecture based on sensorimotor modeling. *Biologically Inspired Cognitive Architectures*, 6 :46–57, 2013.

[13] Gilbert Harman. Moral particularism and transduction. *Philosophical issues*, 15 :44–55, 2005.

[14] John F. Horty. Deontic logic as founded on nonmonotonic logic. *Annals of Mathematics and Artificial Intelligence*, 9 :69–91, 1993.

[15] John F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23 :35–65, 1994.

[16] John F. Horty. *Agency and Deontic Logic*. Oxford University Press, USA, 2009.

[17] Uriah Kriegel. Moral motivation, moral phenomenology, and the alief/belief distinction. *Australasian Journal of Philosophy*, 90 : 469–486, 2012.

[18] Vladimir Lifschitz. Action languages, answer sets, and planning. In *The Logic Programming Paradigm*, pages 357–373. Springer, 1999.

[19] I. Patil, C. Cogoni, N. Zangrando, L. Chittaro, and G. Silani. Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience*, 9 :94–107, 2014.

[20] Luís Moniz Pereira and Ari Saptawijaya. Modelling morality with prospective logic. In *Progress in Artificial Intelligence*, pages 99–111. Springer, 2007.

[21] Stanford Encyclopedia Of Philosophy. Deontic logic, chapter problems surrounding expressive inadequacies of sdl, 2010. URL `http://plato.stanford.edu/entries/logic-deontic/index.html`.

[22] T Powers. Deontological machine ethics. In *2005 AAAI Fall Symposium on Machine Ethics*, pages 79–86, 2005.

[23] P. Quinn. *Divine Commands and Moral Requirements*, chapter What About the "Extraordinary" ? Oxford University Press, 1978.

[24] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13 :81–132, 1980.

[25] Paul Ricoeur. *Lectures 1, Autour du politique*, chapter Éthique et morale. Seuil, 1991.

[26] Georg Spielthenner. On normative practical reasoning. *Abstracta*, 7 (1), 2013.

[27] Mihnea Tufis and Jean-Gabriel Ganascia. Normative rational agents - a bdi approach. In *Proceedings of the 1st workshop on rights and duties of autonomous agents (rda2) 2012 (ecai 2012, pp. 38-43)*, 2012.

[28] Matteo Turilli. Ethical protocols design. *Ethics and Information Technology*, 9 :49–62, 2007.

[29] Bas C Van Fraassen. Values and the heart's command. *The Journal of Philosophy*, 70 :5–19, 1973.

[30] Wiley-Blackwell. Moral dilemma scenarios prone to biases. Newspaper, Dec. 2009.

[31] Bernard Williams. Ethical consistency. In *Proceedings of the Aristotelian Society (Supplementary Volumes), 39, 103-124.*, 1965.