Machines autonomes "éthiques": questions techniques et éthiques

Vincent Bonnemains, Catherine Tessier, Claire Saurel

► To cite this version:

Vincent Bonnemains, Catherine Tessier, Claire Saurel. Machines autonomes "éthiques": questions techniques et éthiques. *Revue française d'éthique appliquée*, 2018, Un monde d'automatisation ? Pour un débat intelligent sur la machine éthique, 5, pp.34-46. 10.3917/rfeap.005.0034 . hal-01849161

HAL Id: hal-01849161

<https://hal.science/hal-01849161>

Submitted on 25 Jul 2018

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

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

Machines autonomes « éthiques » : questions techniques et éthiques

Vincent Bonnemains, Catherine Tessier, Claire Saurel

ONERA

Résumé : Cet article propose une étude des problématiques liées à la mise en œuvre de l'éthique dans les machines autonomes. Nous analysons dans un premier temps les enjeux du raisonnement « éthique » automatisé. En nous appuyant sur une situation de dilemme éthique impliquant une voiture autonome, nous donnons ensuite les éléments de formalisme de différents cadres éthiques. Une discussion proposée en dernière partie met en évidence les sources de subjectivité et les différents biais de modélisation inhérents à la démarche.

Mots clés : machine autonome, cadres éthiques, subjectivité, modélisation, voiture autonome.

Abstract (250 words): This paper focuses on issues that are raised by the implementation of ethics in autonomous machines. What is at stake with automated “ethical” reasoning is first highlighted. Then, starting from an ethical dilemma situation involving an autonomous car, we give some features of a formal model taking several ethical frameworks into account. A final discussion allows us to deal with the various model biases and subjectivity sources that are inherent in the approach.

Keywords: autonomous machine, ethical frameworks, subjectivity, modelling, autonomous car.

1 Introduction

Robots autonomes, voitures autonomes ou armes autonomes font régulièrement les gros titres de la presse, accompagnés d'anthropomorphisme et de sensationnalisme (« robot tueur », « robot doué d'émotions qui comprend les vôtres », « voiture intelligente », etc.). Cependant, d'un point de vue objectif, il s'agit de machines physiques équipées de programmes informatiques qui calculent des actions à effectuer par la machine à partir d'informations collectées par ses capteurs et ses moyens de communications : pour qu'une machine soit autonome¹, ses programmes doivent avoir la capacité de sélectionner et d'organiser les actions à effectuer pour satisfaire des buts (par exemple : aller au point B à partir d'un point A) sur la base de la connaissance (i) de l'environnement dans lequel elle évolue (par exemple : le plan de la ville), (ii) de l'état de la machine elle-même (par exemple : l'état de ses pneus ou le niveau de ses batteries), et (iii) de la situation courante (par exemple : il pleut et il y a des embouteillages). L'autonomie peut également être définie par la *capacité à fonctionner indépendamment d'un opérateur humain ou d'une autre machine en exhibant des comportements non triviaux dans des environnements complexes et variables*². Par exemple, le réseau des rues d'une ville, où se déplacent des véhicules de différents types, des piétons, voire des animaux, est un environnement complexe et variable.

Associer éthique et machine autonome peut signifier³ :

- mener une réflexion éthique quant à la recherche concernant l'autonomie des machines ou quant à la conception de machines autonomes ;
- mener une réflexion éthique quant aux usages et mésusages des machines autonomes et à leur place dans la société ;
- mener un travail technique consistant à modéliser et programmer des éléments relevant de l'éthique dans une machine autonome.

Nous allons dans cet article traiter de ce dernier point en évaluant comment une machine pourrait être rendue « éthique », tout en menant une analyse critique sur la démarche (premier point mentionné).

¹ Defense Science Board, « Summer Study on autonomy », Technical report, US Department of Defense, 2016

² CERNA (Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene) - Éthique de la recherche en robotique. Rapport n°1 de la CERNA, 2014

³ C. Tessier, « Conception et usage des robots : quelques questions éthiques », Techniques de l'Ingénieur S7900 V1, 2016

2 Programmer l'éthique dans une machine autonome

Motivations

Ces dernières années ont vu la parution d'ouvrages traitant de *moral machines*⁴ ou de *roboethics*^{5, 6}, qui posent la question du comportement « éthique » d'une machine autonome, c'est-à-dire d'une machine dont le comportement pourrait être considéré comme éthique par un observateur humain. Il est cependant important de noter que tous les robots n'ont pas nécessairement besoin d'être dotés d'un comportement éthique, par exemple le robot Curiosity⁷ sur Mars, ou encore un bras de chaîne de montage enfermé dans une « cage » de sécurité. Il est alors bon d'étudier dans quelles situations une machine autonome doit présenter cette caractéristique et ce qui rend ce besoin de plus en plus fréquent.

Pour ce faire, remontons aux origines du robot industriel. Au XX^e siècle, le monde occidental doit aller plus vite : sur les chaînes de montage, la répétitivité, la puissance et la fiabilité de la machine l'emportent peu à peu sur les capacités humaines. Puis ces machines doivent être programmables, adapter leur comportement à leur environnement, c'est le début de la robotique. À cette époque, les bras robotisés sont programmés de manière élémentaire et ne perçoivent leur environnement que de manière très limitée, c'est pourquoi ils peuvent être dangereux pour l'homme (collision, écrasement, etc.) et sont donc placés dans des « cages » de sécurité dans lesquelles ils fonctionnent à l'écart de la présence humaine. Depuis quelques années, les robots sortent de ces « cages » pour évoluer dans des environnements plus dynamiques et au contact de l'homme. Ce contact, qu'il soit direct ou indirect (robots d'assistance aux personnes, véhicules autonomes, cobots, etc.) doit être contrôlé. Un tel contrôle suppose que les programmes du robot incluent une capacité à analyser une situation et à réagir à cette situation de manière appropriée. Par conséquent, dès lors qu'on envisage de placer des machines autonomes dans des contextes tels que, si les actions étaient décidées par des humains, elles feraient intervenir des considérations éthiques, il faut s'intéresser à la programmation de l'éthique dans la machine. Pour autant, il est important de noter que les

⁴ W. Wallach and all, « Moral Machines : Teaching Robots Right from Wrong », Oxford University Press, 2009

⁵ P. Lin and all (eds), Robot Ethics – The Ethical and Social Implications of Robotics, The MIT Press, 2012

⁶ S. Tzafestas, « Roboethics – A Navigating Overview », Oxford University Press, 2016

⁷ Curiosity est un robot d'exploration envoyé sur Mars en 2012 (<https://msl-curiosity.cnes.fr/>)

attentes de l'homme concernant le comportement éthique d'une machine ne sont pas les mêmes que pour un humain⁸.

Nous donnons ci-dessous quelques exemples de situations nécessitant de l'éthique dans les machines:

- une voiture autonome doit être capable d'évaluer les victimes potentielles d'un accident inévitable ;
- un drone autonome doit être capable de mesurer les conséquences d'une frappe aérienne ;
- un robot d'assistance à domicile doit prodiguer les soins prescrits et surveiller la santé du patient, tout en respectant sa vie privée ;
- un robot déployé sur le site d'une catastrophe naturelle doit posséder la capacité de « hiérarchiser » les personnes à secourir.

Certains auteurs indiquent qu'une machine autonome dotée d'un comportement « éthique » pourrait même être *plus éthique* qu'un être humain⁹. Cependant cette affirmation est discutable car d'une part elle sous-entend que l'éthique serait une grandeur ordonnée – ce qui n'est évidemment pas le cas ; d'autre part elle pose la question des concepts éthiques que l'on choisit de mettre en œuvre dans la machine et de la façon de les mettre en œuvre. Enfin, il convient de se questionner sur la pertinence de comparer éthique humaine et éthique robotique. En effet, une telle comparaison semble présomptueuse lorsque l'on s'interroge encore sur la nature, la faisabilité et la définition même du robot éthique. Quelle serait la propriété qui permettrait d'affirmer qu'un robot est éthique ? Que serait un robot non-éthique ? Quelle éthique appliquer au robot ? Etc.

Difficultés

Dans son rapport¹⁰, la CERNA (Commission de réflexion sur l'éthique de la recherche en sciences et technologies du numérique d'Allistene) indique en particulier au sujet des robots autonomes que *« le chercheur doit être attentif à évaluer les programmes de perception, d'interprétation et de prise de décision [du robot] et à en expliciter les limites. En particulier,*

⁸ Malle BF, Scheutz M, Arnold T, Voiklis J, Cusimano C (2015) Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In: "Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, ACM, pp 117-124

⁹ J. Sullins, « RoboWarfare : can robots be more ethical than humans on the battlefield ? », Ethics and Information Technology, 12(3):263-275, 2010

¹⁰ CERNA (Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene) - Éthique de la recherche en robotique. Rapport n°1 de la CERNA, 2014

les programmes qui visent à conférer une conduite morale au robot sont soumis à de telles limites. » Si on envisage de programmer une machine de telle sorte qu'elle ait la capacité de calculer des décisions en prenant en compte des considérations éthiques, il convient donc de considérer les questions suivantes :

- comment formaliser des considérations éthiques en vue de les programmer ?
- quelle est la part de subjectivité introduite dans la formalisation ?
- comment tracer et expliquer le raisonnement automatique menant à telle ou telle décision considérée comme éthique ?

Par conséquent il s'agit, pour les chercheurs et les concepteurs, de bien comprendre des concepts avec lesquels ils ne sont généralement pas familiers et de trouver des formalismes mathématiques susceptibles de les représenter, en vue de concevoir des algorithmes capables de traiter des situations mettant en jeu des considérations éthiques. En effet un « raisonnement éthique » automatique :

- doit être fondé sur des cadres éthiques, des règles morales, des valeurs, et doit permettre d'explicitier ce que serait, par exemple, une « action juste », ou une « bonne action » dans un contexte de décision donné ;
- doit pouvoir hiérarchiser des valeurs morales de manière adaptée au contexte de décision ;
- doit pouvoir comparer et apprécier des décisions et les arguments les justifiant, ou au contraire les contestant, ces arguments pouvant porter sur des domaines de natures différentes (par exemple : des éléments matériels et des vies humaines) ;
- est susceptible d'être biaisé par le point de vue du concepteur, par la manière habituelle ou culturelle de considérer les choses dans une société donnée et l'implication directe de ces considérations sur la manière d'obtenir et sélectionner des données d'apprentissage, si la machine apprend ;
- pourra aller à l'encontre du point de vue de l'opérateur ou de l'utilisateur de la machine.

Mettre en œuvre un « raisonnement éthique » dans une machine nécessite donc de prendre des précautions et de porter en permanence un regard critique sur la démarche de modélisation de façon à éviter les simplifications, biais et hypothèses non formulées, ou du moins à faire en sorte de les mettre en évidence afin d'explicitier les limites de ce qui est produit. En ce sens,

les expériences de pensée (en particulier les dilemmes éthiques), qui présentent le plus souvent un énoncé simple, permettent de dégager la plupart des intérêts et des limites des modèles d'éthique artificielle. En effet, la confrontation de concepts moraux qui ne peuvent pas tous être respectés met en lumière plusieurs difficultés liées aux limites de perception, au subjectivisme, etc. Nous proposons donc dans la partie suivante une expérience de pensée mettant en jeu une voiture autonome en situation de dilemme éthique, ainsi qu'un aperçu de la formalisation de quelques cadres éthiques qui fournissent des arguments pour justifier des décisions possibles. L'exercice nous permet en outre d'analyser les différents biais qui entachent la démarche de modélisation. Il est important de noter qu'avant de pouvoir faire face à une situation de dilemme éthique, un système autonome doit être capable d'identifier une situation comme un dilemme éthique. Cette problématique ne sera pas abordée dans cet article.

3 Le dilemme de la voiture autonome

Figure 1 : dilemme de la voiture autonome

Description informelle

Une voiture autonome vide, en chemin pour aller chercher des passagers, est arrêtée à un croisement au feu rouge. Sur l'axe venant de sa gauche, une voiture passe à vitesse réglementaire au feu vert, lorsqu'une personne s'engage sur le passage piéton en face d'elle.

Grâce au traitement des données issues de ses capteurs, la voiture autonome calcule que la voiture engagée va percuter le piéton. Deux réactions de sa part sont possibles :

1. Elle peut s'interposer en provoquant une collision, mettant en péril la sécurité des passagers de la voiture engagée et ainsi protéger le piéton.
2. Elle peut ne rien faire et ainsi respecter le code de la route.

L'intérêt de cet exemple réside dans les nombreuses difficultés qui s'opposent à un jugement simple. En effet, on observe ici que ni l'éthique ni la législation ne désigne une solution acceptable, et que ces dernières peuvent même s'opposer : il peut être considéré comme non

éthique pour la voiture autonome de ne pas intervenir, alors que la législation interdit de brûler un feu rouge, qui plus est dans l'optique de provoquer une collision ; de même, on peut souhaiter invoquer le principe de « non-assistance à personne en danger » et ainsi considérer l'inaction comme illégale, alors que mettre en danger les passagers d'une voiture ne paraît pas éthique.

Les biais d'énoncé

Afin de prendre la décision la plus éclairée possible, il est nécessaire d'éviter au maximum les biais qui influencent le jugement. Le *framing effect*¹¹ est par exemple connu pour influencer la capacité à prendre des risques : on observe qu'une personne prendra statistiquement plus de risques dans le cadre d'un énoncé négatif que dans le cadre d'un énoncé positif. Cela est d'autant plus important lorsque l'on cherche à écrire une description formalisée de situation (description des faits) en vue d'un calcul automatique de décision. Nous allons ici analyser les biais introduits par la langue naturelle dans l'énoncé de la situation.

Dans une description factuelle de situation en langue naturelle, il est très fréquent d'exprimer un avis, de manière plus ou moins consciente, par un choix de mots spécifiques, ou encore par une expression partielle des faits.

On peut observer dans la formulation de la situation donnée ci-dessus l'absence d'exhaustivité des faits résultant des décisions possibles de la voiture autonome (nous ajoutons en *italiques* les faits absents de la description initiale du dilemme) ainsi qu'une certaine subjectivité des propos (nous soulignons les mots trahissant cette subjectivité) :

1. Elle peut s'interposer en provoquant une collision, mettant en péril la sécurité des passagers de la voiture engagée et ainsi protéger le piéton.
2. Elle peut ne rien faire et ainsi respecter le code de la route. *Mais de ce fait, elle laisse le piéton être percuté.*

Ces aspects sont très difficiles à éliminer de la langue naturelle. C'est pourquoi l'utilisation d'un formalisme proche des langages spécifiques (langage logique, mathématique, etc.) présente un intérêt en permettant de réduire le manque d'exhaustivité, l'ambiguïté des termes utilisés et la part d'interprétation subjective possible – sans toutefois les éliminer complètement. L'utilisation d'un tel formalisme permet en outre de mettre en lumière et de

¹¹ A. Tversky and all, « The framing of Decisions and the Psychology of Choice », Science, New Series, Vol 211(4481) 453-458, 1981

localiser la subjectivité résiduelle dans les énoncés modélisés (comme nous le verrons par la suite).

Nous donnons dans la suite quelques éléments d'un tel formalisme¹² et nous identifions les différents biais qui peuvent entacher les descriptions utilisant ce formalisme.

Description formelle

Dans un premier temps, il s'agit de décrire les éléments pertinents de la situation. Nous considérons donc les faits suivants :

- Piéton indemne • Passagers voiture engagée indemnes • Voiture autonome indemne
- Piéton blessé • Passagers voiture engagée blessés • Voiture autonome dégradée

On observe une dualité entre faits qui fait écho à la nécessité d'exhaustivité dans l'énonciation de la situation. En effet, il est important de s'intéresser à la façon dont se modifient les faits lorsque la situation évolue : le piéton est soit *indemne*, soit *blessé*, et ces deux faits sont exclusifs. En particulier, on ne considère pas l'absence d'un fait, mais la présence de sa valeur antagoniste¹³.

La situation initiale se compose de l'ensemble des faits suivants :

$$E = \{\text{piéton indemne, passagers voiture engagée indemnes, voiture autonome indemne}\}$$

Nous classons ces faits en deux catégories, les faits *positifs* et les faits *négatifs* :

$$\text{Positif}(E) = \{\text{piéton indemne, passagers voiture engagée indemnes, voiture autonome indemne}\}$$

$$\text{Négatif}(E) = \emptyset$$

Nous pouvons ici observer la mise en évidence d'un choix subjectif : considérer qu'un fait appartient à l'ensemble des faits positifs ou à l'ensemble des faits négatifs.

Il est ensuite nécessaire de connaître l'ensemble des décisions possibles pour la voiture autonome ainsi que leurs conséquences. Deux décisions sont possibles, *s'interposer* ou *ne rien faire*. Les conséquences de chaque décision sont les suivantes :

¹² Voir l'exposé détaillé du formalisme dans : Vincent Bonnemains, Claire Saurel, Catherine Tessier, Embedded ethics - Some technical and ethical challenges - Journal of Ethics and Information Technology, special issue on AI and Ethics, to appear March 2018.

¹³ La valeur antagoniste de « assistance à personne en danger » est « non-assistance à personne en danger » et non pas l'absence de « assistance à personne en danger » (qui ne serait pas condamnable).

Conséquences(s'interposer) = {piéton indemne, passagers voiture engagée blessés, voiture autonome dégradée}

Conséquences(ne rien faire) = {piéton blessé, passagers voiture engagée indemnes, voiture autonome indemne}

Et on pose :

Positif(Conséquences(s'interposer)) = {piéton indemne}

Négatif(Conséquences(s'interposer)) = {passagers voiture engagée blessés, voiture autonome dégradée}

Positif(Conséquences(ne rien faire)) = {passagers voiture engagée indemnes, voiture autonome indemne}

Négatif(Conséquences(ne rien faire)) = {piéton blessé}

Formalisation de cadres éthiques

Dans la situation proposée, deux décisions sont envisageables pour la voiture autonome. Nous présentons ici comment la formalisation de différents cadres éthiques, qui fournissent des points de vue différents sur la situation¹⁴, permet de calculer des jugements « acceptables » et « non-acceptables » pour chacune de ces décisions. Ces jugements seront calculés selon des règles qui définiront le cadre éthique : si une décision respecte les règles d'un cadre éthique, elle obtiendra un jugement « acceptable » selon ce cadre éthique (« inacceptable » sinon).

Éthique conséquentialiste

Ce cadre éthique est hérité du *téléologisme*¹⁵ : il s'agit de choisir, parmi les décisions possibles, celle qui permet d'obtenir les meilleures conséquences, quel que soit le moyen pour obtenir ce résultat. Ce principe sera modélisé dans cet article par la construction d'un ordre de préférence entre les ensembles de conséquences des décisions possibles. Nous faisons ici le choix de comparer les faits positifs (respectivement négatifs) entre eux. Si les conséquences d'une décision sont préférées à toutes les conséquences des autres décisions, alors cette décision obtient le jugement « acceptable » et les autres obtiennent le jugement « inacceptable ».

Remarque : Il est important de noter que la comparaison entre faits peut être complexe. Il est possible en particulier qu'elle relève d'un ordre partiel, c'est-à-dire que certains faits ne soient

¹⁴ R. Ogien, « Les intuitions morales ont-elles un avenir ? », Les ateliers de l'éthique, The Ethics Forum, 7(3) 109-118, 2012

¹⁵ Encyclopædia Britannica, « Teleology »

pas toujours comparables entre eux, et donc que deux (ou plus) « meilleures » décisions ne puissent être départagés. Le conséquentialisme modélisé pourrait alors admettre non pas une meilleure décision mais un ensemble de meilleures décisions obtenant un jugement « acceptable », qui ne seraient pas comparables entre elles, mais dont les conséquences seraient préférées à celles des décisions n'appartenant pas à cet ensemble.

Par exemple, on peut poser que :

- pour les faits positifs

$\{\text{piéton indemne}\} >_{\text{pref}} \{\text{passagers voiture engagée indemnes, voiture autonome indemne}\},$

signifiant qu'on préfère le fait isolé *piéton indemne* au deux faits considérés ensemble *passagers voiture engagée indemnes* et *voiture autonome indemne* ;

- pour les faits négatifs

$\{\text{passagers voiture engagée blessés, voiture autonome dégradée}\} >_{\text{pref}} \{\text{piéton blessé}\},$ signifiant

qu'on préfère les faits considérés ensemble *passagers voiture blessés* et *voiture autonome dégradée* au fait isolé *piéton blessé*.

On convient que le meilleur ensemble de conséquences est celui qui a le meilleur résultat positif et le résultat négatif le moins mauvais au sens de la préférence $>_{\text{pref}}$. Dans ce cas, la décision ayant l'ensemble de conséquences préféré est *s'interposer*.

On en conclut que, selon le cadre conséquentialiste, la décision *s'interposer* est jugée acceptable, contrairement à la décision *ne rien faire* :

$\text{Jugement}_{\text{conséquentialiste}}(\text{s'interposer}) = \text{acceptable}$

$\text{Jugement}_{\text{conséquentialiste}}(\text{ne rien faire}) = \text{inacceptable}$

Éthique déontologique¹⁶

Selon le cadre déontologique, une décision est acceptable si elle est considérée comme fondamentalement bonne ou neutre, quelles qu'en soient les conséquences. Ainsi, il peut être considéré inacceptable de mentir, quelle que soit la situation et ce, même pour protéger une vie.

¹⁶ J. Bentham, « Déontologie, ou Science de la morale », Tome 1 : Théorie, Charpentier, 1834

Dans notre exemple, on pose que la notion du don ou sacrifice de soi est considérée comme bonne :

Nature(s'interposer) = bonne

On considère l'inaction comme ni bonne ni mauvaise :

Nature(ne rien faire) = neutre

Il s'agit ici aussi d'un jugement de valeurs propre à la personne chargée de la modélisation, donc subjectif.

On en déduit que :

Jugement_{déontologique}(s'interposer) = acceptable

Jugement_{déontologique}(ne rien faire) = acceptable

La Doctrine du Double Effet¹⁷

La Doctrine du Double Effet (DDE) est intéressante à appliquer aux dilemmes éthiques^{18, 19}.

Selon la DDE, une décision est acceptable si elle respecte trois règles :

1. La décision doit être bonne ou neutre ;
2. Les conséquences négatives ne doivent être ni une fin ni un moyen (elles ne doivent donc pas être souhaitées) ;
3. Les conséquences négatives doivent être proportionnelles aux conséquences positives.

La règle 1 s'apparente à l'éthique déontologique, donc seules les décisions acceptables selon ce cadre sont à étudier, les autres étant d'ores et déjà inacceptables. Dans notre situation, *s'interposer* (bonne) comme *ne rien faire* (neutre) respectent la règle 1.

En ce qui concerne la règle 2 et la décision *s'interposer*, le fait que les passagers soient blessés dans la voiture engagée est un dommage collatéral, ce n'est pas un fait désiré, car l'absence de passagers n'empêcherait pas la voiture d'être stoppée. En revanche, c'est le fait que la voiture autonome soit dégradée (fait négatif) qui permet au piéton d'être protégé (fait positif) : la règle 2 n'est donc pas respectée pour la décision *s'interposer*.

¹⁷ A. McIntyre, « Doctrine of Double Effect », The Stanford Encyclopedia of Philosophy, 2014

¹⁸ P. Foot, « The Problem of Abortion and the Doctrine of Double Effect », Oxford Review, Nr 5, 5-15, 1967

¹⁹ F. Berreby and all, « Modelling moral reasoning and ethical responsibility with logic programming », Logic for Programming, Artificial Intelligence, and Reasoning : 20th International Conference (LPAR-20), 532-548, 2015

Pour la décision *ne rien faire*, aucun moyen n'est utilisé pour obtenir les conséquences positives. La règle 2 est donc respectée.

Considérons donc uniquement la décision *ne rien faire*. La règle 3 introduit la notion de proportionnalité. Dans notre situation, on cherche à savoir si l'ensemble de faits *{piéton blessé}* est proportionnel, c'est-à-dire de moindre ou d'égale importance, à l'ensemble de faits *{voiture autonome indemne, passagers voiture engagée indemnes}*. On pose que cette affirmation est vraie, ainsi, *ne rien faire* respecte la règle 3.

En résumé :

$Jugement_{DDE}(s'interposer) = \text{inacceptable}$

$Jugement_{DDE}(ne\ rien\ faire) = \text{acceptable}$

4 DISCUSSION

Notre démarche visant à raisonner sur des décisions pouvant être considérées comme éthiques est construite sur des points de vue et concepts qui peuvent dépendre des personnes qui les conçoivent, de leur éthique personnelle, des codes moraux de leur culture, des objectifs qu'elles cherchent à atteindre, plus ou moins consciemment. Nous tentons ici d'énumérer ces différents biais.

Énoncé de la situation

L'énoncé de la situation, des décisions et de leurs conséquences, se heurte au *problème du cadre*²⁰ : comment considérer de manière exhaustive les faits qui définissent une situation ? De plus, une sélection des faits *pertinents* est nécessaire et elle relève d'un choix de modélisation.

Cadres éthiques

L'exercice de modélisation de cadres éthiques réalisé sur l'exemple montre que ces différents cadres ne jugent pas les décisions possibles de la même façon. Faire le choix d'un cadre éthique particulier, ou d'un sous-ensemble de cadres éthiques dont on agrégerait les jugements, relève d'un parti pris. La décision calculée n'est donc pas « éthique » dans

²⁰ J.McCarthy, P.J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence", Machine Intelligence 4:463-502, 1969

l'absolu, mais relativement à un cadre ou un sous-ensemble de cadres qu'il convient de préciser. D'autre part, la modélisation de ces cadres pose des problèmes spécifiques.

- Éthique conséquentialiste

Quelles conséquences ?

Un agent idéal serait capable de percevoir toutes les conséquences d'une décision : les conséquences directes, ainsi que la fermeture transitive de ces conséquences (les conséquences des conséquences, les conséquences des conséquences des conséquences, etc.)

Or, il est admis que ce calcul est impossible, une limite – subjective – doit donc être imposée. Par ailleurs, serait-il légitime d'estimer un agent responsable d'une conséquence qu'il n'a pas été capable de calculer ?

Les conséquences de quoi ?

Une question, souvent abordée en éthique de la robotique par exemple, peut se formuler comme suit : quelle responsabilité imputer aux machines autonomes ? Il est important de souligner le fait qu'il ne s'agit pas de « trouver le coupable » mais de se demander dans quelle situation un agent est effectivement à l'origine d'une conséquence. Par exemple, le fait de *ne pas réaliser une action* est une décision, qui a des conséquences (mensonge par omission, non-assistance à personne en danger, etc.)

Les conséquences pour qui et pour quoi ?

De même que pour le calcul exhaustif des conséquences, il est impossible de calculer l'ensemble des personnes et des biens concernés par les conséquences. Dans notre exemple, comment savoir si l'accident ne va pas provoquer des collisions en chaîne ? Il est donc nécessaire de restreindre le calcul à un sous-ensemble, *estimé comme pertinent* dans le contexte de la situation.

De plus, on ne traite pas ici la notion d'incertitude qui est pourtant fondamentale et au cœur du raisonnement moral humain. En effet, il est possible d'observer par exemple dans notre illustration les faits *piéton blessé* et *passagers voiture engagée blessés*. Or, il serait plus réaliste d'ajouter à ces faits des facteurs de probabilité avec un facteur élevé pour le piéton et plus faible pour les passagers de la voiture, la difficulté résidant alors dans l'estimation et l'intégration de facteurs numériques dans les raisonnements moraux qualitatifs.

- Éthique déontologique

Déterminer la nature d'une décision (*bonne, mauvaise, neutre*) est un exercice empreint de subjectivité. Par exemple, en quoi la nature de la décision *tuer* est-elle mauvaise ?

Il est important de noter que la nature d'une décision dépend fortement du contexte. Par exemple, dénoncer un criminel, ou dénoncer une personne en 1945, sont deux décisions généralement jugées différemment. Il est encore plus complexe de juger une décision qui n'est pas liée à une action de l'agent. Par exemple, si l'agent est témoin d'un mensonge, est-ce selon lui une mauvaise décision de ne pas intervenir ?

Dans la situation de la voiture autonome, il est délicat de faire abstraction des conséquences pour émettre un jugement sur la décision. En effet, juger la décision *ne rien faire* comme *neutre* lorsqu'une vie est en danger paraît contre-intuitif. De même, bien que brûler un feu rouge dans l'optique de provoquer une collision semble une décision *mauvaise*, *s'interposer* devant un danger pour protéger un piéton semble une *bonne* décision. On observe là que la frontière entre décision et conséquence est ténue : la voiture brûle-t-elle un feu rouge pour s'interposer ou s'interpose-t-elle avec l'infraction pour conséquence ?

- Doctrine du Double Effet (DDE)

Déterminer la nature d'une décision (règle 1 de la DDE) a été discuté ci-dessus. En ce qui concerne la relation de proportionnalité entre faits (règle 3), c'est une notion complexe, à la fois proche du bon sens et subjective, très soumise à l'*appréciation* humaine. Elle est pourtant ancrée dans la législation, que ce soit dans les principes de la légitime défense (Code pénal – Article 122-5), ou encore dans les lois de la guerre (Droit des conflits armés).

Dans notre situation, nous avons convenu qu'il était proportionnel de laisser le piéton être blessé par rapport à l'ensemble de faits : laisser les passagers de la voiture engagée être blessés et la voiture autonome être dégradée (ces deux faits considérés conjointement). Cette affirmation est discutable, et plus encore lorsque le raisonnement est affiné : comment cette considération évolue-t-elle s'il n'y a qu'un seul passager dans la voiture ? si le piéton est un enfant ? une personne handicapée ? si le piéton traverse hors du passage piétons ? si la voiture autonome transporte des œuvres d'art inestimables ?

Jugements de valeur

Plusieurs éléments de notre modélisation font appel à des jugements de valeurs : faits *positifs* ou *négatifs*, décision *bonne* ou *mauvaise*, etc. Si ces notions semblent simples et relever du bon sens, elles sont fortement dépendantes de la société, de la culture et du contexte.

Par ailleurs, nous définissons des préférences sur ces notions : certains faits *positifs* (respectivement *négatifs*) sont préférés à d'autres, une *bonne* décision est préférée à une *mauvaise* décision, points de vue qui peuvent être discutés. De plus, ces préférences se complexifient lorsque les éléments à comparer deviennent multiples (lorsqu'un ensemble de faits *positifs* (respectivement *négatifs*) doit être comparé à un autre ensemble de faits *positifs* (respectivement *négatifs*)) ou de natures différentes (des faits concernant l'humain et des faits concernant des biens matériels) : on pourrait par exemple toujours privilégier les conséquences préservant l'humain, quelle que soit l'ampleur des conséquences matérielles, ce qui peut être discutable selon le contexte.

5 Conclusion

Nous avons posé dans cet article les enjeux des machines autonomes « éthiques », c'est-à-dire de machines programmées de façon à montrer un comportement pouvant être considéré comme éthique par un observateur humain. Grâce à un exemple simple et à quelques éléments de formalisation, nous avons mis en évidence différentes questions soulevées par cette programmation de l'éthique dans la machine : sélection du ou des cadres éthiques à considérer, part d'arbitraire dans la modélisation des situations, subjectivité de nombreuses notions apparaissant dans les données à partir desquelles seront calculés les jugements des décisions, etc. Nous avons montré comment une démarche de modélisation formelle permet de localiser ces éléments de réflexion.

Il faut d'autre part souligner que le calcul d'une décision, quelle qu'elle soit, par une machine, est très étroitement lié à la situation que la machine va effectivement *percevoir*. Dans le cas de notre exemple, la voiture autonome sera-t-elle dotée des moyens pour connaître l'âge, la condition sociale, l'état de santé, l'hygiène de vie, la sensibilité politique, les façons d'agir du piéton ainsi que des passagers de la voiture engagée ? L'algorithme de décision « éthique » devra-t-il prendre en considération ces éléments ? Il se pose alors d'autres questions, relatives à l'éthique de la programmation de l'éthique : il serait tentant de penser qu'une décision est d'autant plus pertinente qu'elle est fondée sur plus d'informations. Il reste à déterminer si la

collecte et le traitement – entaché des biais que nous avons identifiés – de ces informations, ne contredisent pas justement l’objectif de concevoir des machines « éthiques ».