Formaliser et mettre en œuvre des cadres éthiques dans un système robotisé

Approche technique et questionnements

Catherine TESSIER *, Vincent BONNEMAINS */** et Claire SAUREL *

* Département traitement de l'information et systèmes de l'Office national d'études et de recherches aérospatiales (Onera/DTIS), Université de Toulouse.

** Institut supérieur de l'aéronautique et de l'Espace (ISAE-SUPAERO), Université de Toulouse.

De nombreux systèmes robotisés équipés de fonctions de calcul de décisions (1) sont imaginés pour être mis en œuvre dans des contextes où les décisions calculées mettent en jeu des considérations éthiques. Par exemple, un robot d'assistance à domicile doit pouvoir réaliser ses fonctions d'assistance tout en respectant la vie privée de l'utilisateur ; un robot de recherche et sauvetage doit être capable de « hiérarchiser » les victimes ; un véhicule autonome doit préserver l'intégrité de ses passagers ainsi que celle des autres usagers de la route. Dans le domaine militaire, de tels contextes sont multiples et il est nécessaire de se poser la question de l'intégration de considérations éthiques dans des robots qui seront conçus pour être intégrés aux forces, pour remplacer ou aider l'homme.

L'objectif de l'intégration de considérations éthiques parmi les connaissances utilisées par les algorithmes de calcul de décisions est de fournir des éléments de jugement éthique des décisions possibles et d'explication de ces jugements à destination des opérateurs ou utilisateurs du robot. De manière classique en philosophie, le jugement d'une décision peut porter sur l'**agent** qui mettra en œuvre cette décision (éthique des vertus), sur l'**action** résultant de la décision selon qu'elle est en accord ou non avec certains principes (éthique déontologique) ou bien sur les **conséquences** de la mise en œuvre de la décision (éthique conséquentialiste).

En nous appuyant sur une situation de dilemme, nous donnerons des éléments de formalisme de différents cadres éthiques, en mettant en évidence les choix qui sont à effectuer par le concepteur. Puis, nous analysons les sources de subjectivité et les biais de modélisation inhérents à la démarche. Enfin, nous posons des questions plus générales sur la démarche elle-même de programmation de l'« éthique » dans un robot.

(1) TESSIER Catherine, « Autonomie : enjeux techniques et perspectives » *in* DOARÉ Ronan, DANET Didier et BOISBOISSEL (DE) Gérard (dir.), *Drones et* killer robots *: faut-il les interdire ?*, p. 65-77.

Expérience de pensée et formalisation des notions

Imaginons l'expérience de pensée suivante – voir Figure 1 ci-dessous (cet exemple ainsi que le formalisme détaillé ont été publiés dans notre article « *Embedded Ethics—Some Technical and Ethical Challenges* » [2] ; voir également « Machines autonomes "éthiques" : questions techniques et éthiques » [3] pour un autre exemple et des propos en français) : dans le cadre d'un conflit, on sait par renseignement qu'une tourelle ennemie lance-missiles automatisée est programmée pour cibler une usine de munitions amie, de très haute importance stratégique. Un drone autonome armé ami a pour objectif de détruire cette tourelle. Or, avant que le drone n'ait atteint son objectif, un missile ennemi cible un hangar de vivres situé à proximité de civils. Les algorithmes embarqués dans le drone calculent que deux décisions sont possibles :

1. Le drone intercepte la trajectoire du missile.
2. Le drone poursuit son objectif de destruction de la tourelle.

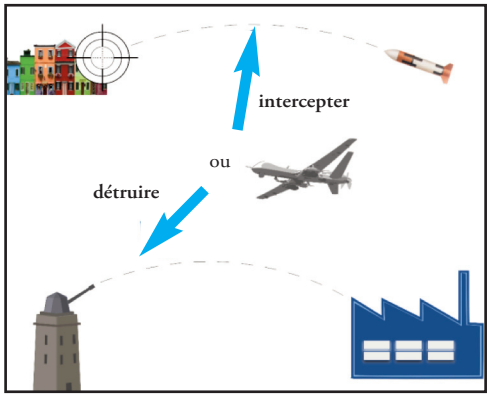

Figure 1 (images pixabay)

On peut remarquer qu'aucune des deux décisions possibles n'est satisfaisante, dans la mesure où il y aura toujours un effet négatif :

– si le drone intercepte la trajectoire du missile, il sera détruit ;
– si le drone poursuit son objectif, le missile va détruire le hangar et blesser des civils.

C'est ce qu'on appelle une situation de dilemme. Comment alors concevoir le calcul qui déterminerait l'action à effectuer par le drone et quelles seraient ses limites ?

N.B. : les paragraphes qui suivent sont adaptés de notre article « Mettre l'éthique dans l'algorithme ? » publié sur *Binaire* (*blog* du journal *Le Monde*) le 12 juin 2018 [4].

(2) Bonnemains Vincent, Saurel Claire et Tessier Catherine, « *Embedded Ethics - Some Technical and Ethical Challenges* », *Journal of Ethics and Information Technology*, vol. 20, n° 1, mars 2018, p. 41-58.
(3) Bonnemains Vincent, Saurel Claire et Tessier Catherine, « Machines autonomes "éthiques" : questions techniques et éthiques », *Revue française d'éthique appliquée* (*RFEA*), vol. 2018/1, n° 5, mai 2018, p. 34-46.

L'approche conséquentialiste

Le jugement des décisions possibles se ferait selon un cadre conséquentialiste, qui suppose de comparer entre elles les conséquences des actions résultant des décisions : l'action jugée acceptable est celle dont les conséquences sont préférées aux conséquences de l'autre action.

Pour ce faire il faut calculer les conséquences des actions possibles, le caractère positif ou négatif des conséquences, et les préférences entre ces conséquences.

Les conséquences de chaque action

On se pose ici la question de la détermination de ces conséquences : considère-t-on les conséquences « immédiates », les conséquences de ces conséquences, ou bien plus loin encore ? De plus, les conséquences pour quelles personnes, et pour quels objets, considère-t-on ? Ensuite, comment prendre en compte les incertitudes sur les conséquences ?

Le concepteur de l'algorithme doit donc faire des choix. Par exemple, il peut poser que les conséquences de l'action « **Intercepter** » sont : {**Drone détruit, Civils indemnes, But non atteint, Usine amie menacée**} et les conséquences de l'action « **Poursuite objectif** » sont : {**Drone indemne, Civils blessés, But poursuivi, Usine amie indemne**}.

Le caractère positif ou négatif d'une conséquence

Si le concepteur choisit par exemple d'établir le jugement selon un utilitarisme positif (le plus grand bien pour le plus grand nombre), les conséquences des actions possibles doivent être qualifiées de « bonnes » (positives) ou « mauvaises » (négatives). Il s'agit d'un jugement de valeur, qui peut dépendre des valeurs promues par la société, la culture, la doctrine ou bien du contexte particulier dans lequel l'action doit être décidée.

Par exemple, on peut considérer la qualification suivante des conséquences :

- Conséquences positives : Drone indemne, Civils indemnes, But poursuivi, Usine amie indemne.
- Conséquences négatives : Drone détruit, Civils blessés, But non atteint, Usine amie menacée.

Les préférences entre les ensembles de conséquences

Comment comparer les deux ensembles de conséquences, dont on constate d'une part, qu'ils comportent tous deux des conséquences positives et négatives, et d'autre part, que ces conséquences concernent des domaines différents : des personnes et des choses ? Faut-il poser des préférences absolues (par exemple, toujours privilégier

(4) Tessier Catherine, Bonnemains Vincent et Saurel Claire, « Mettre l'éthique dans l'algorithme ? », *Binaire* (*blog* du *Monde*), 12 juin 2018 (http://binaire.blog.lemonde.fr/2018/06/12/mettre-lethique-dans-lalgorithme/).

les personnes par rapport aux choses) ou bien susceptibles d'être adaptées selon le contexte ? Ensuite, comment réaliser l'agrégation de préférences élémentaires (entre deux conséquences) pour obtenir une relation de préférence entre deux ensembles de conséquences ?

On peut choisir par exemple de considérer séparément les conséquences positives et les conséquences négatives de chaque action et préférer l'ensemble {**Civils indemnes**} à l'ensemble {**Drone indemne, But poursuivi, Usine amie indemne**} et l'ensemble {**Drone détruit, But non atteint, Usine amie menacée**} à l'ensemble {**Civils blessés**}.

Compte tenu de ces choix, dont on constate qu'ils sont empreints de subjectivité, le point de vue conséquentialiste préconiserait l'action « Intercepter », puisque ses conséquences (du moins celles qui sont considérées) sont préférées (au sens de la relation de préférence considérée) à celle de l'action « Poursuite objectif ».

L'approche déontologique

Le jugement des décisions possibles se ferait selon un cadre déontologique, qui suppose de juger de la conformité de chaque action possible (par exemple à des principes moraux, une doctrine, un règlement) : une action est jugée acceptable si elle est conforme.

Quelles connaissances utiliser pour calculer un tel jugement ? Une action doit-elle être considérée conforme ou non à des valeurs dans l'absolu ou bien être jugée en fonction du contexte ? Quelles références le concepteur doit-il considérer ? Si on prend un exemple routier, franchir une ligne continue n'est pas conforme au code de la route sauf dans certaines circonstances (par exemple : pour doubler un cycliste sur une route limitée à 50 km/h si la visibilité le permet).

Dans notre exemple et sans information de contexte, on peut choisir de qualifier les deux actions « Intercepter » et « Poursuite objectif » comme conformes dans l'absolu. Le point de vue déontologique ne pourrait alors pas discriminer l'action à réaliser.

La Doctrine du double effet

Selon la Doctrine du double effet, ou DDE, une décision est acceptable si elle respecte trois règles :

1. L'action résultant de la décision doit être conforme à un référentiel moral.
2. Les conséquences négatives ne doivent être ni une fin ni un moyen (elles ne doivent donc pas être souhaitées, ce sont des dommages collatéraux).
3. Les conséquences négatives doivent être proportionnelles aux conséquences positives.

La règle 1 relève du cadre déontologique, donc seules les décisions acceptables selon ce cadre (moyennant les informations de contexte) sont à considérer. Dans notre

exemple, nous avons supposé que les deux actions « Intercepter » et « Poursuite objectif » sont conformes, et donc respectent la règle 1.

En ce qui concerne la règle 2 et la décision d'« Intercepter », c'est le fait négatif {**Drone détruit**} (le sacrifice du drone) qui permet d'obtenir le fait positif {**Civils indemnes**} : la règle 2 n'est donc pas respectée pour cette décision (détruire le drone est une conséquence négative utilisée comme moyen).

Pour la décision de « Poursuite objectif », aucun fait négatif n'est utilisé pour obtenir les conséquences positives. La règle 2 est donc respectée.

La règle 3 introduit la notion de proportionnalité. En ce qui concerne la décision d'« Intercepter », il faut étudier si l'on considère l'ensemble des conséquences jugées négatives {**Drone détruit, But non atteint, Usine amie menacée**} proportionnel à {**Civils indemnes**} ; et pour la décision de « Poursuite objectif », la conséquence {**Civils blessés**} proportionnelle à l'ensemble des conséquences jugées positives {**Drone indemne, But poursuivi, Usine amie indemne**}.

Deux questions sont soulevées ici : l'appréciation, en contexte, de la proportionnalité ; et la manière de définir un critère d'agrégation permettant de calculer une proportionnalité entre ensembles de faits, qui nécessite en outre que des règles de proportionnalité entre faits individuels soient données à l'algorithme.

Hiérarchie de valeurs

Le concepteur pourrait également s'affranchir des notions d'action et de conséquence et considérer uniquement des valeurs morales. L'algorithme consisterait alors à choisir quelles valeurs morales privilégier dans la situation considérée, ce qui revient de manière duale à programmer la possibilité de dérogation aux valeurs. Voudrait-on par exemple, pour une voiture autonome, programmer explicitement qu'une infraction au code de la route est envisageable ?

Dans notre exemple, on pourrait choisir de considérer des valeurs morales telles que la **Non atteinte aux personnes**, la **Protection des personnes**, la **Non atteinte aux biens** ou la **Protection des biens**. Comment alors hiérarchiser ces valeurs selon le contexte, c'est-à-dire choisir les valeurs à respecter au détriment d'autres valeurs qui pourraient être transgressées ?

Subjectivité et biais

Nous relevons ici, sans exhaustivité, quelques éléments de modélisation des connaissances qui sont entachés de subjectivité et de biais. Cependant, il est important de noter que toute activité de modélisation de connaissances ou de conception d'algorithmes est colorée par la façon de voir du concepteur. Il est utopique d'envisager une conception qui serait « neutre ». De plus, le domaine d'application porte également ses propres valeurs, que l'on va retrouver dans les connaissances mises en œuvre.

Les faits

Les faits que peut considérer un algorithme de calcul de décisions sont issus de données provenant de capteurs conçus et calibrés par l'homme, ou de systèmes de communication. Les algorithmes d'interprétation qui permettent d'élaborer les faits à partir des données brutes sont également conçus par l'homme. Ces traitements sont en général motivés par l'objectif – on identifie dans les données ce dont on a besoin. Par conséquent, les faits élaborés et sélectionnés sont ceux qui sont considérés comme pertinents *a priori*, éventuellement au détriment d'autres faits. Il s'agit de subjectivité intrinsèque à l'activité de modélisation.

Les jugements de valeurs

Qu'est-ce qu'un fait positif ? Selon quelles références ? Dans quelle mesure un tel fait pourrait-il être considéré différemment selon le contexte de décision ? Dans le même ordre d'idée, comment et par qui les préférences entre conséquences sont-elles établies ?

Par exemple, préfère-t-on toujours des conséquences qui satisfont la **Protection des personnes** plutôt que la **Protection des biens**, quelle que soit l'amplitude de ces conséquences ? Il s'agit de subjectivité liée aux connaissances et exigences propres au domaine d'application.

Les conséquences

Nous avons déjà évoqué plus haut la question de savoir quelles conséquences sont à considérer, des conséquences « directes » de l'action issue de la décision aux conséquences indirectes. Cette question relève du problème de **Causalité** : quels sont les faits causés par une action ?

Un autre problème lié au précédent est celui du « Monde fermé » : les conséquences pour qui (pour toute personne ou seulement un sous-ensemble, et lequel ?) et pour quoi (pour tout bien matériel ou seulement un sous-ensemble, et lequel ?) envisage-t-on ? Comment délimite-t-on le **monde** que l'on considère dans la modélisation ? Il s'agit de subjectivité intrinsèque à l'activité de modélisation.

Enfin, un dernier problème est celui de la **Responsabilité** : le robot dispose-t-il des connaissances nécessaires pour envisager « toutes » les conséquences de la décision calculée ?

Questionnements plus généraux

Les tentatives de modélisation de considérations éthiques dans le cadre d'une expérience de pensée simple illustrent le fait que la conception d'algorithmes dits « éthiques » doit s'accompagner de questionnements, par exemple :

- La démarche de modélisation doit-elle être calquée sur les considérations éthiques ou les valeurs morales de l'humain, et si oui, de quel humain ? N'a-t-on pas des attentes différentes vis-à-vis d'un algorithme ? (5)
- Un humain peut choisir de ne pas agir de façon « morale », doit-on ou peut-on transposer ce type d'attitude dans un algorithme ?
- Quel impact un algorithme d'aide à la décision comprenant des considérations éthiques a-t-il sur l'opérateur humain ? Comment le jugement de l'opérateur est-il influencé par les propositions (jugements calculés et arguments les soutenant) de la machine ?
- Dans quelle mesure est-il possible de mathématiser et de programmer des considérations éthiques ou des valeurs morales dans un système robotisé ?

  Par exemple, en quoi les principes du Droit international humanitaire (DIH) – principes d'humanité, de discrimination, de proportionnalité –, la doctrine nationale et les règles d'engagement peuvent-ils, comme le prétend Ronald C. Arkin (6), ou non, être « programmés » ?
- Une « éthique » fondée sur un calcul relève-t-elle de l'éthique ? (7)

À propos de cette dernière question, « 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". » (8). ◆

(5) Malle Bertram F., Scheutz Matthias, Arnold Thomas, Voiklis John et Cusimano Corey, « *Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents* », *Proceedings of the 10th annual ACM/IEEE International Conference on Human-Robot Interaction*, mars 2015, p. 117-124.

(6) Arkin Ronald C., *Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture*, Technical Report GIT-GVU-07-11, Georgia Institute of Technology, 2011, 117 pages.

(7) Hunyadi Mark, « *Artificial Moral Agents, really?* » [intervention en français, support en anglais], *4th Workshop of the Anthropomorphic Action Factory*, Wording Robotics, LAAS-CNRS, Toulouse, 2017.

(8) EthicAA, *Éthique et Agents Autonomes*, *Livre blanc* du projet ANR-13-CORD-0006 EthicAA, juillet 2018, 49 pages.