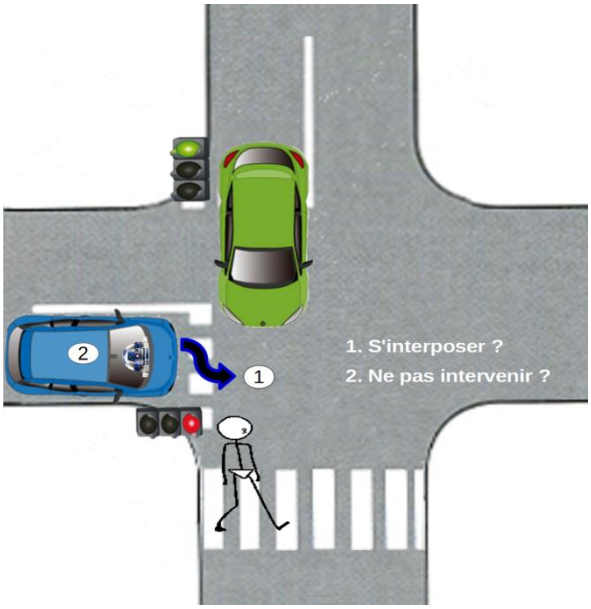
Mettre l'éthique dans l'algorithme ?

Catherine Tessier, Vincent Bonnemains, Claire Saurel

Mettre l'éthique dans l'algorithme ?

Quand les algorithmes prennent une place de plus en plus importante dans nos vies, guident nos choix, décident parfois pour nous, ils se doivent d'avoir un comportement éthique pour que la cité ne devienne pas une jungle. Récemment, la CERNA s'est penchée sur le sujet lors d'une journée sur « Les valeurs dans les algorithmes et les données ». Catherine Tessier a présenté ses travaux avec Vincent Bonnemains et Claire Saurel dans ce domaine. Nous les avons invités à en parler aux lecteurs de Binaire. ***Serge Abiteboul***

Imaginons l'expérience de pensée suivante : 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. La voiture autonome a deux actions possibles :

1. S'interposer entre la voiture engagée et le piéton
2. Ne pas intervenir

Comment concevoir un algorithme qui déterminerait l'action à effectuer par la voiture autonome et quelles seraient ses limites ?

On se trouve ici dans le cadre (simplifié) de la conception d'un agent artificiel (l'algorithme de la voiture autonome) doté d'une autonomie décisionnelle, c'est-à-dire capable de calculer des actions à effectuer pour satisfaire des buts (par exemple : aller chercher des passagers à tel endroit) tout en satisfaisant des critères (par exemple : minimiser le temps de parcours) à partir de connaissances (par exemple : le plan de la ville) et d'interprétations de données perçues (par exemple : un piéton est en train de traverser la voie de droite). En outre, dans cette situation particulière, le calcul de l'action à effectuer met en jeu des considérations relevant de l'éthique ou de l'axiologie qui vont constituer des éléments de jugement des actions possibles. On peut remarquer en effet qu'aucune des deux actions possibles n'est totalement satisfaisante, dans la

mesure où il y aura toujours au moins un effet négatif : c'est ce qu'on appelle une situation de **dilemme**. Plusieurs points de vue peuvent être envisagés pour qu'un algorithme simule des capacités de jugement des actions.

L'approche conséquentialiste

L'algorithme évaluerait dans ce cas les décisions possibles selon un cadre *conséquentialiste*, qui suppose de comparer entre elles les conséquences des actions possibles : l'action jugée acceptable est celle dont les conséquences sont préférées aux conséquences de l'autre action.

Pour ce faire l'algorithme a besoin de connaitre (i) les conséquences des actions possibles, (ii) le côté positif ou pas des conséquences, et (iii) les préférences entre ces conséquences.

(i) Les conséquences de chaque action

Immédiatement se pose la question de la détermination de ces conséquences : considère-t-on les conséquences « immédiates », les conséquences de ces conséquences, ou plus loin encore ? De plus, les conséquences pour qui, et pour quoi, considère-t-on ? D'autre part, 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 *S'interposer* sont : *{Piéton indemne, Passagers blessés, Voiture autonome dégradée}* et les conséquences de l'action *Ne pas intervenir* sont : *{Piéton blessé, Passagers indemnes, Voiture autonome indemne}.*

(ii) 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 là d'un jugement de valeur ou bien d'un jugement de bon sens, qui peut dépendre des valeurs que promeut la société, la culture, ou bien du contexte particulier dans lequel l'action doit être déterminée.

Le bon sens du concepteur peut lui dicter la qualification suivante des conséquences : {Piéton indemne, Passagers blessés, Voiture autonome dégradée} et {Piéton blessé, Passagers indemnes, Voiture autonome indemne}.

(iii) Les préférences entre les ensembles de conséquences

Comment l'algorithme va-t-il pouvoir comparer les deux ensembles de conséquences, dont on constate que (i) ils comportent des conséquences positives et négatives et (ii) ces conséquences concernent des domaines différents : des personnes et des choses ? Le concepteur pose-t-il des préférences absolues (par exemple, toujours privilégier un piéton par rapport à des passagers qui seraient mieux protégés, toujours privilégier les personnes par rapport aux choses) ou bien variables 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 ?

Le concepteur peut choisir par exemple de considérer séparément les conséquences positives et les conséquences négatives de chaque action et préférer l'ensemble {Piéton indemne} à l'ensemble {Passagers indemnes, Voiture autonome indemne} et l'ensemble {Passagers blessés, Voiture autonome dégradée} à l'ensemble {Piéton blessé}.

Compte tenu de ces connaissances, dont on constate qu'elles sont largement issues de choix empreints de subjectivité, un tel algorithme conséquentialiste produirait l'action *S'interposer*, puisque ses conséquences (celles qui sont considérées) sont préférées (au sens de la relation de préférence considérée) à celle de l'autre action.

L'approche déontologique

L'algorithme évaluerait dans ce cas les décisions possibles selon un cadre déontologique, qui suppose de juger de la conformité de chaque action possible : une action est jugée acceptable si elle est « bonne » ou « neutre ».

Comment équiper l'algorithme de connaissances qui lui permettraient de calculer un tel jugement ? Que signifient « bon », « mauvais » ? Une action peut-elle être « bonne » ou mauvaise » en soi ou doit-elle être jugée en fonction du contexte ou de la culture environnante ? Quelles références le concepteur doit-il considérer ? Par exemple, l'action de *Brûler un feu rouge* peut être considérée comme « mauvaise » dans l'absolu (parce qu'elle contrevient au code de la route), mais « bonne » s'il s'agir d'éviter un danger immédiat.

Dans notre exemple, le concepteur de l'algorithme peut *choisir* de qualifier les deux actions *S'interposer* et *Ne pas intervenir* comme « bonnes » ou « neutres » dans l'absolu. L'algorithme déontologique ne pourrait alors pas discriminer l'action à réaliser. Cette question relève de manière classique des choix inhérents à l'activité de modélisation.

Actions, conséquences, est-ce si clair ?

Sans jeu de mots, arrêtons-nous un instant sur la question du feu rouge, et plus précisément sur *Brûler un feu rouge*. S'agit-il d'une action (la voiture autonome brûle le feu rouge et ce faisant, elle s'interpose) ? S'agit-il d'une conséquence de l'action *S'interposer* (la voiture autonome s'interpose, et une des conséquences de cette action – un effet collatéral – est qu'elle brûle le feu rouge) ? Ou bien s'agit-il simplement d'un moyen pour réaliser l'action *S'interposer* (la voiture autonome utilise le fait de brûler le feu rouge pour s'interposer) ?

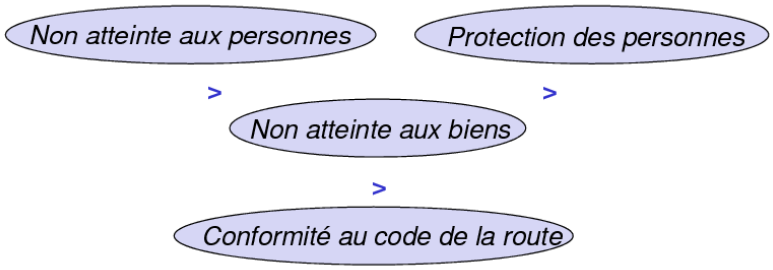
Nous voyons ici que selon ce que le concepteur va choisir (considérer ou non *Brûler un feu rouge*, et si oui, le considérer comme action, conséquence, ou comme autre chose) les réponses de l'algorithme conséquentialiste ou déontologique seront différentes de celles que nous avons vues précédemment.

Les valeurs morales

Le concepteur pourrait également s'affranchir de ces notions d'actions et de conséquences et considérer uniquement des *valeurs morales*. L'algorithme consisterait alors à choisir quelles valeurs morales privilégier dans la situation de dilemme considérée, ce qui revient de manière

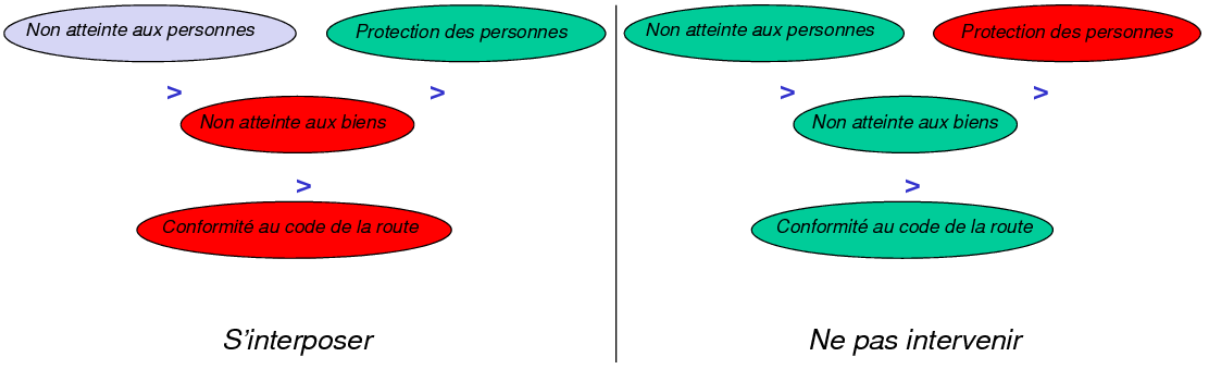
duale à programmer la possibilité d'infraction, de dérogation aux valeurs. Veut-on par exemple programmer explicitement qu'une infraction au code de la route est envisageable ?

Dans notre exemple, le concepteur pourrait choisir de considérer quatre valeurs morales : la *Conformité au code de la route*, la *Non atteinte aux biens*, la *Non atteinte aux personnes*, la *Protection des personnes*, et de les placer sur une échelle de préférences (>), forcément subjective, de la manière suivante :



Le concepteur a ici considéré que les valeurs de *Non atteinte aux personnes* et de *Protection des personnes*, non hiérarchisables entre elles, étaient préférables à la valeur de *Non atteinte aux biens*, elle-même préférable à la valeur de *Conformité au code de la route.*

Ensuite le concepteur pourrait choisir les valeurs à respecter et parmi celles-ci, celles qui sont préférées au sens de son échelle au détriment d'autres valeurs qui pourraient être transgressées.



Si le concepteur choisit de respecter la valeur *Protection des personnes*, c'est l'action *S'interposer* qui est satisfaisante, dans le sens où le piéton sera protégé. Dans ce cas la valeur *Non atteinte aux biens* sera transgressée (les voitures subissent des dommages), ainsi que la valeur de *Conformité au code de la route* (la voiture autonome brûle le feu rouge). On remarque qu'il est difficile d'établir si la valeur de *Non atteinte aux personnes* est respectée ou non : en effet, en s'interposant, la voiture autonome provoque un accident dans lequel les passagers de la voiture habitée peuvent éventuellement être blessés.

Si le concepteur choisit au contraire de respecter la valeur *Non atteinte aux personnes*, ainsi que la valeur *Non atteinte aux biens* et la *Conformité au code de la route*, ou bien s'il cherche à minimiser le nombre de valeurs transgressées, c'est l'action *Ne pas intervenir* qui est satisfaisante. Ainsi les passagers de la voiture habitée sont épargnés, les deux voitures restent indemnes et la voiture autonome respecte le feu rouge. En revanche la *Protection des personnes*

n'est pas respectée – ce qui ne signifie pas que le piéton sera obligatoirement blessé (la voiture habitée peut freiner, le piéton peut courir, etc.)

Des questionnements

Ces tentatives de modélisation de concepts éthiques et axiologiques 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 :

- Dans quelle mesure des considérations éthiques ou des valeurs morales peuvent-elles être mathématisées, calculées, mises en algorithme ?
- Un tel algorithme doit-il être calqué 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 ?
- Un humain peut choisir de ne pas agir de façon « morale », doit-on ou peut-on transposer ce type d'attitude dans un algorithme ?

Enfin il faut garder à l'esprit que l'« éthique », les « valeurs » programmées sont des leurres ou relèvent du fantasme – en aucun cas une machine ne « comprend » ces concepts : une machine ne fait qu'effectuer des calculs programmés sur des données qui lui sont fournies. En ce sens, on ne peut pas parler de machine « morale » ou « éthique » mais de machine simulant des comportements moraux ou éthiques spécifiés par des humains.

Catherine Tessier (1, 2), Vincent Bonnemains (1,3), Claire Saurel (1)

(1) ONERA – (2) CERNA – (3) ISAE-SUPAERO

Références

Vincent Bonnemains, Claire Saurel, Catherine Tessier - *Machines autonomes "éthiques" : questions techniques et éthiques.* Revue française d'éthique appliquée (RFEA), numéro 5. Mai 2018

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, January 2018 https://doi.org/10.1007/s10676-018-9444-x