ONERA
THE FRENCH AEROSPACE LAB

www.onera.fr

PROPOSITION DE SUJET DE THESE

Intitulé : Elaboration de justifications de décisions pour un agent autonome en situation de dilemme moral

Référence : **TIS-DDS-2019-06**
(à rappeler dans toute correspondance)

Laboratoire d'accueil à l'ONERA :

Domaine : Intelligence artificielle et décision — Lieu (centre ONERA) : Toulouse

Département : DTIS

Unité : RIME — Tél. : 05 62 25 29 14

Responsable ONERA : Catherine TESSIER — Email : catherine.tessier@onera.fr

Directeur de thèse envisagé :

Nom : Catherine TESSIER et Claire SAUREL

Adresse : ONERA Toulouse

Tél. : 05 62 25 29 14 Email : catherine.tessier@onera.fr claire.saurel@onera.fr

Sujet : Dans de nombreux domaines civils ou militaires, il est envisagé de confier à des agents autonomes ou des robots des tâches réalisées habituellement par des personnes : robots d'assistance au soin de personnes en situation de fragilité, véhicules autonomes, drones de combat… Lorsqu'un robot est confronté à une situation où des choix sont envisageables entre plusieurs actions, son comportement est géré par des logiciels conçus pour lui faire calculer une décision. Lorsque chacune des décisions possibles froisse au moins une des valeurs morales d'un être humain qui devrait prendre la décision dans le même contexte que le robot, le robot est placé dans une situation de dilemme moral. Ce problème soulève des débats et des craintes dans la société car :

- il touche à des questions philosophiques et peut donner lieu à des avis divergents ;
- il suscite une peur que la décision ne soit pas maîtrisée, et qu'elle échappe complètement au bon sens en vigueur dans le domaine applicatif dans lequel doit évoluer le robot ;

Les éléments permettant de comprendre les enjeux éthiques associés à chacune des décisions possibles ne sont pas forcément explicités. Dans de telles situations, la responsabilité du choix de la décision revient au concepteur de l'algorithme qui anime le robot s'il est complètement autonome, ou bien à l'opérateur du robot en cas de partage de l'autorité. Il est donc utile sinon nécessaire de pouvoir produire une justification accompagnant et caractérisant le calcul des décisions possibles selon différents points de vue éthiques, afin de déterminer des choix « justes » à cet égard - que ce soit en aide à la conception de l'algorithme qui gérera le robot, ou bien en aide à la résolution de conflit sur le choix de la décision en situation, si la décision est partagée avec un opérateur.

Une thèse précédente [1] a consisté à proposer un formalisme pour identifier et décrire une situation de dilemme moral et calculer des jugements des décisions possibles en situation de dilemme ; ces jugements sont élaborés en référence à un ensemble choisi de cadres éthiques, inspirés de travaux en philosophie. Pour pouvoir donner un sens aux jugements ainsi déterminés à un opérateur ou à un concepteur d'algorithme, et l'aider dans sa réflexion sur les décisions à faire calculer au robot, il est indispensable de pouvoir argumenter ou justifier ces jugements [2,3] de manière suffisamment

pertinente pour l'humain. En effet, la compréhension par l'humain de ces jugements peut lui permettre de les comparer ; cela est particulièrement indispensable si une des décisions possibles donne lieu à des jugements différents selon les cadres éthiques considérés, ou si plusieurs décisions donnent lieu à des jugements identiques selon un même cadre éthique.

L'objet principal de la thèse est de construire, pour les différentes décisions possibles, des justifications axées sur les points de vue éthiques modélisés, qui soient pertinentes pour un concepteur ou un opérateur humain.

Programme de la thèse

- Etat de l'art sur les cadres éthiques, en vue de compléter les points de vue éthiques selon lesquels des décisions peuvent être jugées ;
- Etude de travaux sur la génération de justifications et d'argumentation ;
- Définition des concepts à ajouter dans le formalisme existant pour intégrer de nouveaux points de vue éthiques d'une part, et produire des justifications des jugements des décisions possibles d'autre part ;
- Définition d'une méthode d'élaboration et de présentation de ces justifications ;
- Expérimentations en vue d'éprouver les concepts proposés avec des participants en situation d'opérateurs de robot.

Méthodes et outils envisagés :

Logique, graphes, programmation et outils graphiques

Références

[1] 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

[2] Katie Atkinson, Trevor Bench-Capon - Taking account of the actions of others in value-based reasoning. Artificial Intelligence: Volume 254 Issue C, January 2018

[3] Francesco Olivieri, Regis Riveret, Antonino Rotolo, Guido Governatori, Serena Villata - Dialogues on Moral Theories. DEON 2018

Collaborations extérieures :

PROFIL DU CANDIDAT

Formation : Ecoles d'ingénieurs ou Université

Spécificités souhaitées : Intelligence Artificielle, Logique, Programmation, Expérimentation avec des participants