ÉLÉMENT DE PORTFOLIO 03

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : A Declarative Modular Framework for Representing and Applying Ethical Principles

URL de l'élément : `https://hal.sorbonne-universite.fr/hal-01564675`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Il s'agit d'un article publié à AAMAS 2017 présentant le cadre de raisonnement éthique développé dans l'équipe à l'occasion de la thèse de Fiona Berreby. Il est représentatif des travaux de l'équipe sur la modélisation de raisonnement éthique, posant les bases d'un cadre qui a été développé par nos travaux ultérieurs. Cet article a été depuis cité par un certain nombre d'acteurs de la communauté d'éthique computationnelle (une trentaine de citations externes, dont une dizaine données en référence plus bas).

3 PRÉSENTATION DE CET ÉLÉMENT

Cet article examine l'utilisation de langages de haut niveau dans la conception d'agents autonomes éthiques. Il propose un cadre logique nouveau et modulaire pour représenter et raisonner sur une variété de théories éthiques, sur la base d'une version modifiée du calcul des évènements, implémentée en Answer Set Programming.

Le processus de prise de décision éthique est conçu comme une procédure en plusieurs étapes, identifiant les différents composants qu'il est nécessaire de représenter pour permettre un raisonnement complet.

Un modèle d'action, fondé sur des mécanismes classiques de représentation de l'action et du changement, permet à l'agent d'évaluer son environnement et d'anticiper le déroulé des évènements résultant de ses choix, obtenant ainsi pour chaque scénario envisagé (choix possible d'actions par l'agent) une trace des évènements qui se déclenchent dans le système et des états résultants.

Un modèle de causalité permet ensuite d'analyser cette trace pour identifier les relations causales entre les actions de l'agents et les différents évènements pouvant survenir, lui permettant de raisonner sur sa responsabilité, d'identifier conséquences de ses actes et d'estimer si certains effets sont utilisés comme moyens d'arriver à d'autres.

Enfin, le modèle éthique à proprement parler, séparé en modèles du Bien et modèles du Juste déterminent à partir des informations précédentes quels sont les choix éthiquement acceptables selon différents principes éthiques. L'article en présente un certain nombre, tirés de la littérature en éthique normative, intégrant dans un même cadre des principes conséquentialistes et d'autres plus déontologiques.

L'ambition de cette approche est double. Tout d'abord, elle est de permettre la représentation systématique d'un nombre illimité de processus de raisonnements éthiques, à travers un cadre adaptable et extensible. Deuxièmement, elle est d'éviter l'écueil trop courant d'intégrer directement l'information morale dans de raisonnement général sans l'expliciter, alimentant ainsi les agents avec des réponses atomiques qui ne représentent pas la dynamique sous-jacente. En séparant clairement les différentes problématiques de représentation (action, causalité, éthique) et en identifiant explicitement à chaque étapes les entrées du modèle spécifiques à un domaine donné, nous visons à déplacer de manière globale le processus de raisonnement moral du programmeur vers le programme lui-même.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Jugement éthique dans le processus de décision d'un agent bdi. *Rev. d'Intelligence Artif.*, 31(4) :471–499, 2017.

[2] Louise A Dennis, Martin Mose Bentzen, Felix Lindner, and Michael Fisher. Verifiable machine ethics in changing contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11470–11478, 2021.

[3] Louise A Dennis and Cristina Perea del Olmo. A defeasible logic implementation of ethical reasoning. In *First International Workshop on Computational Machine Ethics (CME-2021)*, 2021.

[4] Abeer Dyoub, Stefania Costantini, Francesca Alessandra Lisi, and Ivan Letteri. Logic-based machine learning for transparent ethical agents. In *CILC*, pages 169–183, 2020.

[5] David Fuenmayor and Christoph Benzmüller. Normative reasoning with expressive logic combinations. In *ECAI 2020*, pages 2903–2904. IOS Press, 2020.

[6] Umberto Grandi, Emiliano Lorini, Timothy Parker, and Rachid Alami. Logic-based ethical planning. *arXiv preprint arXiv :2206.00595*, 2022.

[7] Martin Jedwabny, Pierre Bisquert, and Madalina Croitoru. Generating preferred plans with ethical features. In *Florida Artificial Intelligence Research Society*, volume 34, 2021.

[8] Emiliano Lorini. A logic of evaluation. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, pages 827–835, 2021.

[9] Sylvie Michel, Sylvie Gerbaix, and Marc Bidan. A plea for choosing ex ante an ethical theorical position for a relevant response to ethical issues posed by algorithmic systems. In *2022 3rd International Conference on Next Generation Computing Applications (NextComp)*, pages 1–6. IEEE, 2022.

[10] Emery A Neufeld, Ezio Bartocci, Agata Ciabattoni, and Guido Governatori. Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4) :43, 2022.

[11] Maurice Pagnucco, David Rajaratnam, Raynaldio Limarga, Abhaya Nayak, and Yang Song. Epistemic reasoning for machine ethics with situation calculus. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 814–821, 2021.

[12] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics : A survey. *ACM Computing Surveys (CSUR)*, 53(6) :1–38, 2020.

[13] John Zoshak and Kristin Dew. Beyond kant and bentham : How ethical theories are being used in artificial moral agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.