Qualification et quantification du Bien et du Mal dans un modèle de raisonnement éthique

Projet de Recherche Doctoral EDITE

Directeur de thèse : Gauvain Bourgne (MCF HDR)
Courriel : Gauvain.Bourgne@lip6.fr
Adresse professionnelle : Sorbonne Université - LIP6 BC169 - 4, place Jussieu 7505 Paris
Doctorant actuellement encadré : Yousef Taheri Sojasi - 2021 - 50%
Laboratoire d'accueil : LIP6, UMR 7606
Ecole doctorale de rattachement : EDITE ED130

Mots-clé associés au sujet : éthique computationelle, agrégation, échelle bipolaire, représentation des connaissances, raisonnement moral

Contexte et motivation

L'étude de la morale d'un point de vue computationnel a attiré l'intérêt croissant de chercheurs en intelligence artificielle [AA11]. L'autonomie croissante des agents artificiels et l'augmentation du nombre de tâches qui leur sont déléguées nous incitent à aborder leur capacité à traiter les restrictions et les préférences éthiques, que ce soit dans leur propre structure interne ou pour des interactions avec des utilisateurs humains. Des domaines aussi variés que la santé ou le transport posent des problèmes éthiques qui sont en ce sens particulièrement pressants, car ils peuvent exiger de la part des agents des prises de décision dont les conséquences sont immédiates ou lourdes. L'éthique computationnelle peut aussi nous aider à mieux comprendre la morale et raisonner plus clairement sur les concepts éthiques qui sont employés dans des domaines philosophiques, juridiques et technologiques.

En philosophie, dans le cadre de l'éthique normative, qui s'attache à définir les principes devant régler une conduite juste, on distingue différents courants selon ce qui est mis en avant dans ces prises de décisions [SA21, AM21]. En particulier, les approches déontologiques s'appuient sur des notions de Devoir tandis que les approches conséquentialistes se basent sur une promotion du bonheur ou du Bien général qui résulte de nos actes.

Différents cadres de raisonnement éthique ont été proposés pour modéliser ces principes éthiques[TKS+21]. Si un grand nombre d'entre eux se focalisent sur la modélisation d'un principe éthique en particulier [GB17, BT12, PS07], certaines approches essaient de fournir un modèle permettant de représenter différents principes, en se fondant sur une représentation logique du contexte de la décision (décrite par une théorie d'action et/ou une théorie causale) [LBN17, LPSN20, BSG21]. Tout ces cadres partagent néanmoins le besoin, pour une application donnée, de qualifier et/ou quantifier le Bien ou le Mal associés à des actes ou leurs conséquences, de manière absolue ou relative à des modalités ou personnes. Ce besoin d'évaluer des situations complexes pouvant affecter différents agents selon différents critères se retrouve dans d'autres domaines de l'intelligence artificielle tels que les problèmes d'allocation de ressources, et plus généralement le choix social computationnel (voir [Con18] pour une étude du lien entre choix social computationnel et questions éthiques), ou l'étude des fonctions d'agrégation [Det00, CKKM02].

Dans ce contexte, au sein de l'équipe d'accueil (ACASA), nous avons développé une architecture modulaire qui permet la représentation systématique et adaptable de principes éthiques [BSG21, BBG18]. Les quatre composantes de cette architecture sont le modèle d'action, qui permet de modéliser la succession d'états et d'événements déroulant d'un ensemble d'actions, le modèle causal, qui détermine à partir de cette trace quelles sont les relations causales liant ces événements afin de permettre d'estimer toutes les conséquences effectives d'une action, le modèle du Bien, qui modélise ce qui est peut être considéré comme Bien ou Mauvais dans le

contexte d'application et le modèle du Juste qui peut utiliser ces évaluations locales du Bien pour estimer si une action est ou non juste selon différents principes éthiques.

Ce modèle du Bien, particulièrement crucial pour les principes conséquentialistes visant à maximiser le bien général, doit ainsi d'abord qualifier chaque conséquence bonne ou mauvaise en précisant la modalité selon laquelle elle peut être considérée comme telle (respect de la dignité humaine, respect de la vie, honnêteté...) et la personne ou le groupe affecté par cette conséquence. Ainsi mentir à X pour sauver Y serait à la fois mauvais du point de vue de l'honnêteté vis-à-vis de X et bon du point de vue du respect de la vie vis-à-vis de Y. Il est alors nécessaire pour la plupart des principes utilitaristes de quantifier le Bien de façon globale. Cela peut alors se faire par agrégation de toutes les conséquences, bonnes ou mauvaises, d'une action.

Dans l'architecture modulaire de [Ber18], l'agrégation est effectuée par une simple somme (éventuellement pondérée) de chacun des éléments qualifiés, ce qui peut avoir des effets indésirables : cela peut par exemple amener à considérer comme *juste* une solution consistant à concentrer tout le malheur sur un faible nombre de personnes. Un premier travail commun a déjà été réalisé avec d'autres équipes du Lip6 travaillant sur les outils d'intelligence computationnelle (LFI), incluant notamment les questions d'agrégation, et sur les problématiques de choix social ou de décisions collectives dans les système multi-agent (SMA). Cela a permis de faire un premier état des problématiques d'agrégations qui se posent dans ce cadre et des outils disponibles, avec en particulier la question de la prise en compte différenciée du Bien et du Mal à travers des échelles bipolaires [RL15, MP19]. Il est en effet souvent considéré comme pire de faire le mal plutôt que de ne pas faire le bien et ce type de nuance ne peut être exprimé par une échelle unique continue. Ces travaux constituent un premier socle sur lequel s'appuyer pour proposer différents mécanismes de quantification du Bien et du Mal adaptés aux différents principes éthiques et pleinement intégrés dans le cadre complet de raisonnement. Pour illustrer, évaluer et contraster ces mécanismes, il est aussi indispensable de construire une plateforme de test avec une base d'exemples et de situations concrètes permettant la diffusion de ces outils.

Objectifs scientifiques

Cette thèse doit étudier tous les différents mécanismes de quantification du Bien ou du Mal mis en oeuvre dans la modélisation d'un raisonnement éthique. Dans un premier temps, il est indispensable d'identifier les données irréductibles (ce qu'il est nécessaire d'indiquer et de quantifier comme entrée atomique) et de caractériser les différentes étapes d'agrégation (par cible, par modalité, sur les différentes conséquences plus ou moins directes d'une actions, sur les différents actions d'un plan...). Un objectif majeur de cette thèse est alors de concevoir un cadre générique formel couvrant et intégrant tous ces aspects, de l'implémenter dans une plateforme ouverte et de le mettre en oeuvre sur une série d'exemples qu'il s'agira de construire avec un souci d'ancrage dans des situations réalistes, concrètes et illustratives.

Il s'agit ainsi d'une part d'identifier les choix importants qui gouvernent ce processus pour axiomatiser les propriétés de ces mécanismes correspondant à différents principes éthiques, et d'autre part de mettre en place un cadre formel de raisonnement et un format de description complet des paramètres d'une situation pour construire une plateforme de comparaison de ces propositions. Cette plateforme devra proposer une base d'exemples ouverte, facilement enrichissable, qui puisse faire référence pour la communauté d'éthique computationelle. La mise en place d'un cadre suffisamment général pour représenter de façon unifiée des situations éthiques diverses et s'accommoder d'approches variées, couplée à la constitution d'un repertoire de situations ainsi formalisées, est en effet un enjeu majeur pour ouvrir le dialogue entre des approches de la modélisation éthique issues de domaines différents (programmation logique, choix social computationnel, décision multicritère...).

Une piste intéressante pour l'étude de cas réels est de s'appuyer sur les travaux en éthique médicale. Des critères et méthodologies ont été étudiés et utilisés par l'OMS pour évaluer des politiques de santé et ont généré des discussions et critiques très pertinentes pour notre

objet d'études. En particulier, on peut citer des mesures comme QALY (Quality-Adjusted Life Years) ou DALY (Disability-Adjusted Life Years) dont la mise en place et les critiques nous informent sur la complexité du problème et la variété des données à prendre en compte. Ce corpus devrait permettre d'extraire à la fois des critères généraux et des points de vigilance méthodologiques pour implémenter dans l'architecture d'évaluation éthique différentes variantes de ces propositions.

Profil de l'étudiant recherché

Nous recherchons un ou une candidate titulaire d'un master en IA, motivé/e, avec des compétences solides en représentation des connaissances et en logique formelle, ainsi que des capacités de programmation pour assurer l'implémentation et un intérêt appuyé pour les thématiques multi-disciplinaires dans lesquelles ce sujet s'inscrit (notamment en philosophie). Des compétences en programmation logique (notamment ASP) constituent un avantage.

References

[AA11] M Anderson and S Anderson. *Machine ethics*. Cambridge University Press, 2011.

[AM21] Larry Alexander and Michael Moore. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[BBG18] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. Cadre déclaratif modulaire d'évaluation d'actions selon différents principes éthiques. *Revue d'Intelligence Artificielle*, 32(4):479–518, 2018.

[Ber18] Fiona Berreby. *Models of ethical reasoning*. PhD thesis, Sorbonne Université, EDITE, LIP6, Paris, 9 2018.

[BSG21] Gauvain Bourgne, Camilo Sarmiento, and Jean-Gabriel Ganascia. ACE modular framework for computational ethics: dealing with multiple actions, concurrency and omission. In *1st International Workshop on Computational Machine Ethics*, Online event, 2021. CEUR-WS.org.

[BT12] Selmer Bringsjord and Joshua Taylor. The divine-command approach to robot ethics. *Robot Ethics: The Ethical and Social Implications of Robotics', MIT Press, Cambridge, MA*, pages 85–108, 2012.

[CKKM02] T. Calvo, A. Kolesarova, M. Komornikova, and R. Mesiar. Aggregation operators: properties, classes and construction methods. In *Aggregation operators: new trends and applications*, pages 3–104. Physica-Verlag, 2002.

[Con18] Vincent Conitzer. Computational social choice and moral artificial intelligence. Tutorial at IJCAI-2018, 2018.

[Det00] Marcin Detyniecki. *Mathematical Aggregation Operators and their Application to Video Querying*. PhD thesis, Université Pierre and Marie Curie, Paris, 2000.

[GB17] Naveen Sundar Govindarajulu and Selmer Bringsjord. On automating the doctrine of double effect. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 4722–4730, Melbourne, Australia, August 2017. AAAI Press.

[LBN17] Felix Lindner, Martin Mose Bentzen, and Bernhard Nebel. The HERA approach to morally competent robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6991–6997, Vancouver, Canada, September 2017. ISSN: 2153-0866.

[LPSN20] Raynaldio Limarga, Maurice Pagnucco, Yang Song, and Abhaya Nayak. Non-monotonic reasoning for machine ethics with situation calculus. In *AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020 Canberra, ACT, Australia, November 29–30, 2020, Proceedings*, pages 203–215, Canberra, ACT, Australia, 2020. Springer, Springer Nature.

[MP19] H. Martin and P. Perny. Biowa for preference aggregation with bipolar scales: application to fair optimization in combinatorial domains. In *Proc. of the Int. Joint Conf. on Artificial Intelligence, IJCAI19*, pages 1822–1828, 2019.

[PS07] L M Pereira and A Saptawijaya. Modelling morality with prospective logic. In *Progress in Artificial Intelligence*, pages 99–111. Springer, 2007.

[RL15] A. Revault d'Allonnes and M.-J. Lesot. Dynamics of trust building: models of information cross-checking in a multivalued logic framework. In *Proc. of the IEEE Int. Conf. on Fuzzy Systems, Fuzz-IEE 2015*, 2015.

[SA21] Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[TKS+21] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 53(6):132:1–132:38, December 2021.