Agents moraux autonomes et algorithmes éthiques: Pluralisme et argument à question ouverte

Clayton Peterson*[1] and Naïma Hamrouni[1]

[1]Département de philosophie et des arts – Université du Québec à Trois-Rivières, Canada

Résumé

Plusieurs auteurs au sein de la littérature scientifique en éthique des machines et en " intelligence artificielle éthique " (*ethical AI*) voient l'automatisation des raisonnements et des comportements éthiques comme une solution plausible pour la résolution des dilemmes moraux. Ainsi, l'idée véhiculée est non seulement que les agents artificiels autonomes moraux, comme les machines et les algorithmes, sont possibles, mais que ceux-ci pourraient éventuellement dépasser l'être humain quant à l'analyse des problèmes éthiques pratiques, à la prise de décision, ainsi qu'aux comportements éthiques. Partant de la distinction de Moor (2006) entre les agents moraux implicites (principes éthiques imposés lors de la programmation et de la conception), explicites (capacité à représenter l'éthique et à faire des choix basés sur cette représentation) et complets (capacité à faire des jugements moraux explicites et à les " justifier "), certains auteurs, comme Anderson et Anderson (2007), soutiennent qu'un objectif important en éthique des machines est la conception d'agents éthiques explicites ou complets dont les actions seraient *justifiées* d'un point de vue éthique. Or, en plus d'être inquiétante, cette position repose sur une mécompréhension de ce qu'est l'éthique, en assumant notamment que celle-ci puisse se définir de manière algorithmique et fonctionnelle. En se basant sur le pluralisme éthique et sur l'argument à question ouverte de Moore, nous argumenterons que cette prémisse est erronée et, incidemment, que les approches en éthique des machines devraient réévaluer leurs objectifs.

*Intervenant