

Résumé de lecture 2 du livre *Faire la morale aux robots*

Par  
Frédérik Boutin

Travail présenté à M. Yacine Benahmed

Dans le cadre du cours Enjeux professionnels et sociétéINF30007

28 novembre 2023

Résumé de lecture /

|  |  |
| --- | --- |
| **Référence bibliographique**  ***1 pt*** | Gibert, M. (2020). *Faire la morale aux robots : une introduction à l’éthique des algorithmes*. Atelier 10. 978-2-89759-516-6 |
| **Date de lecture**  ***0.5 pt*** | 22 novembre 2023 |
| **Présentation de/des auteur(s)**  ***1 pt*** | L’auteur, Martin Gibert, est un philosophe et chercheur qui se spécialise dans l’éthique de l’intelligence artificielle. Il travaille à l’Université de Montréal et il fait partie du Centre de recherche en éthique. |
| **Genre de l’ouvrage ou de l’article**  ***0.5 pt*** | Essai philosophique et métaphysique. |
| **Objectif du texte(1pt), problématique (1pt) et thèses soutenues (1pt)**  ***3 pts*** | **Objectif**: Sensibiliser sur l’approche et les choix d’utilisation de normes morales dans le développement des technologies autonomes (intelligence artificielle).  **Problématique** : Il existe plusieurs normes morales et une multitude de façons différentes de programmer, sans compter toutes les sources de biais possible. Ainsi, comment effectuer les bons choix de conception de ces technologies, permettant d’assurer le bien pour la société aujourd’hui et du futur ?  **Thèses** :  Dans la conception des systèmes intelligents, il faut tenir compte des besoins de toute la population, incluant les groupes marginaux.  Dans la conception, il faut valoriser une approche de morale normative, afin de justifier nos actions sur le bien, plutôt que de suivre nos propres intuitions morales qui sont source de biais.  L’utilisation du principe moral de l’éthique de la vertu serait préférable dans la programmation de systèmes d’IA, puisqu’il permet à un système d’apprendre par association à partir d’exemples d’un modèle jugé vertueux et une meilleure flexibilité.  Il n’y a pas de connexion entre l’intelligence et le bien-fondé des objectifs d’une IA, il est donc préférable d’utiliser une approche inductive pour son apprentissage, afin de permettre une adaptation des objectifs par celle-ci aux situations non envisagées actuellement. |
| **Références théoriques clés[[1]](#footnote-1)**  ***2 pts*** | Roman *Les robots et l’empire* (Issac Asimov, 1985) [p.58],  Roman *L’homme bicentenaire* (Issac Asimov, 1976) [p.58],  Roman *Le sorcier de Terremer* (Ursula Le Guin, 1976) [p.60],  Documentaire *Worlds of Ursula K. Le Guin* (2018) [p.61],  « *The magician* » (The Guardian, 2005) [p.61],  *Le langage de la nuit* (Livre de Poche, 2018) [p.62],  Roman *Les dépossédés* (Ursula Le Guin, 1974) [p.63],  « *Social science fiction* », dans Reginald Bretnor (dir.) [p.63],  *Moderne Science Fiction: Its Meaning and Its Future* (Coward-McCann, 1953) [p.63],  Livre *Brotopia : Breaking Up the Boys’ Club of Silicon Valley* (Emily Chang, 2018) [p.64-65],  « *Judging a book by its description: Analyzing gender stereotypes in the Man Bookers Prize winning fiction* » (arXiv, 2018) [p.66],  *De la marge au centre* (Éditions Cambourakis, 2017) [p.68],  *Qui peut sauver la morale? Essai de métaéthique* (Ithaque, 2019) [p.69],  « *What do philosophers believe?* », *Philosophical Studies*, vol 170, n° 4, 2014 [p.70],  Livre *Émotions et valeurs* (Christine Tappolet, 2000) [p.75],  *Moral Machines : Teaching Robots Right from Wrong* (Wendell Wallach, Colin Allen, 2009) [p.77],  Roman *Sérotonine* (Michel Houellebecq, 2009) [p.78],  Livre *Deep Learning* (Yoshua Bengio, 2016) [p.80],  « *The moral behavior of ethicists* », dans Justin Sysma et Wesley Buckwalter (dir.) [p.85],  *A Companion of Experimental Philosophy* (Blackwll, 2016) [p.85],  *Toward a Code of Ethics for Artificial Intelligence* (Springer, 2017) [p.85],  « *Higher social class predicts increased increased unethical behavior* », PNAS, vol. 109, n° 11, 2012 [p.86],  « *Ceux qui partent d’Omelas* »(Ursula Le Guin, 1973) [p.91-92],  *Aux douze vents du monde* (Le Bélial, 2018) [p.91] |
| **Concept clés (notions définies, organismes clés) - indiquer les définitions importantes avec la page.**  ***3 pts*** | Relativisme moral (p.70),  Relativisme de la locutrice (p.70),  Relativisme culturel (p.71),  Réalisme (p.71),  Voile d’ignorance (p.73),  Perception morale (p.76),  Apprentissage profond (p.80)  \*\*Voire en annexe de ce document pour mes définitions (qui résument les définitions du livre) des notions présentées ici.\*\* |
| **Résumé analytique[[2]](#footnote-2)**  ***10 pts*** | Le chapitre ***Isaac et Ursula*** débute avec la description du docteur en biochimie et l’auteur de science-fiction Issac Asimov. Au cours de sa carrière, il a notamment su faire ressortir un côté obéissant chez les robots, les différenciants ainsi des robots purement monstrueux qui étaient présenté comme la norme. Dans la section ***Imaginer entre hommes***, l’auteur nous présente plutôt l’auteur Ursula Le Guin. Ce dernier écrit également dans le genre de la science-fiction. Un contraste est cependant souligné entre sa vision de la science-fiction, plus axée sur la recherche d’un idéal de société à travers les conditions humaines par rapport à celle d’Asimov, se voulant plus prédictive et d’anticipation des problèmes potentiels du développement technologique. Ceci nous permet de comprendre l’impacte d’une vision partielle d’auteur. En effet, alors que les auteurs comme Asimov essais de résoudre des problèmes techniques, ils ne réalisent pas qu’ils vont transmettent les mêmes préjugés et discriminants de leur société actuelle. La section ***Quel genre de programmation?*** vient faire le parallèle avec la composition des membres du comité d’éthique du ministère allemand des Transports au sujet des voitures autonomes, discuté dans la partie 1 du livre. En effet, l’auteur se questionne sur la répercussion de la disproportion des genres (1 femme seulement sur 14 membres) dans la prise de déscions. Il nous rappelle également que le domaine informatique c’est beaucoup restreint à un profil type, les hommes. La section ***Choisir le réglage par défaut*** présente le portrait actuel de société où les données traduisent souvent un historique de discrimination raciste, classiste et de genre. L’auteur rappel l’importance de ne pas répéter nos billets dans notre développement et qu’il est parfois préférable de changer les réglages par défaut afin d’afficher la diversité. C’est pourquoi il se met à écrire en employant le féminin par défaut pour la suite.  Le chapitre ***Métaéthique pour programmeuses*** présente les interrogations métaéthique que sont à savoir si une programmation morale est meilleure qu’une autre et s’il est possible de construire un bon robot ? C’est dans la section ***Le défi du relativisme moral*** nous présente donc les différentes théories morales rencontrées en programmation, soit le relativisme moral, le relativisme de la locutrice, le relativisme culturel et le réalisme. À la suite de cette comparaison, il nous est possible de réaliser que nous avons tous des intuitions réalistes, de par le fait que deux choses opposées ne peuvent être simultanément vrai et que nous avons tous une forme d’objectivité morale. Le sujet du voile d’ignorance nous est ensuite introduit comme un bon moyen de détermination de principes raisonnables de décision, puisque cette méthode nous permet de réfléchir à une question donnée en excluant (ignorant) notre position dans la société. Elle permet donc d’inclure les autres dans nos prises de décision. La section ***Le défi de la perception morale*** discute du rôle de nos émotions au sein de nos processus moraux. Comment nos émotions influent sur notre perception d’une situation et donc, par le fait même sur la prise de décision. Un lien est finalement réalisé avec l’amélioration décisionnelle des AMA par la détection des émotions.  Le chapitre ***Faire des robots vertueux*** discute de l’approche vertueuse dans la programmation des robots. La section ***Prendre le bien en photo*** débute par la description d’une rencontre de *brainstorm* entre le chercheur en intelligence artificielle Yoshua Bengio et des universitaires par rapport aux sources de données potentielles à utiliser comme modèle d’apprentissage des distinctions morales aux IA. Cette section permet de comprendre toute la complexité dans l’apprentissage morale, c’est-à-dire un apprentissage dit profond, où la compréhension est faite avec une hiérarchie de concepts et ayant plusieurs liens entre ceux-ci. La section ***Le triomphe modeste des robots vertueux*** se concentre sur la problématique de l’identification des personnes vertueuses à utiliser comme modèle pour l’apprentissage des systèmes d’IA. Il y a d’abord un rappel sur les avantages des robots vertueux par rapports à leurs confrères utilitaristes et déontologiques, soit qu’ils utilisent un apprentissage par normativité indirecte, ils représentent un bon compromis d’application et qu’ils sont adaptatifs et plus facilement réalisable. La section ***À la recherche d’une expertise morale*** discute du lien d’influence entre les agissements et des facteurs tels que la soumission à l’autorité ou encore nos humeurs. Ceci est d’ailleurs illustré par une expérimentation datant de 1960 oû des volontaires ont infligés des chocs à des inconnus sous la pression de l’autorité. L’auteur nous démontre que n’y les personnes scolarisées, n’y les personnes prospères sont de meilleurs candidats pour représenter des modèles vertueux et que c’est probablement par l’intelligence collective qu’on pourrait trouver les meilleurs candidats. La section ***Éthique applicable*** discute de deux contextes applicatifs du modèle vertueux, le premier étant le dilemme de l’accident de la voiture autonome avec le choix de la personne à sauver (le vieillard ou l’enfant), le deuxième étant le retour sur l’AMA conversationnel Aristote et la reconnaissance des situations. Une prise de décision en fonction d’une proportionnelle et d’un critère de niveau de confiance d’une population de vertueux représenterait la solution optimisé dans le premier cas, alors que l’AMA apprendrait la reconnaissance des situations conversationnelles d’après les exemples des situations émotionnelles des personnes vertueuses. |
| **Citations clés**  ***3 pts*** | « […] je regarde les gens dans le bus et je me demande ce que le développement de l’intelligence artificielle changera à leur quotidien. J’entrevois aussi le problème difficile, mais pas insurmontable : comment programmer les robots en fonction de principes moraux qui puissent satisfaire tout le monde ? » (p. 2) |
| **Idées importantes en lien avec le cours et appréciation personnelle**  ***7 pts*** | Le livre discute de plusieurs concepts en lien avec notre cours, tels que la différence entre les normes morales et nos intuitions personnelles, les trois doctrines éthiques que sont l’utilitarisme, le déontologisme et l’éthique de la vertu, ainsi que l’importance de la notion d’inclusion de tous dans la conception.  D’abord, je trouve que l’auteur a su transmettre l’information de manière logique et cohérente à travers les cinq chapitres. En effet, le premier chapitre ***Introduction :le bus des jours fériés*** permet une introduction aux concepts de base tel que l’éthique des algorithmes. Ils sont tous nécessaires à la compréhension de la problématique sur le choix des normes morales dans la conception de technologies. Le second chapitre ***Le vieillard ou l’enfant*** vient ajouter un second niveau de définitions telles que l’utilitarisme et le déontologisme. De plus, il permet l’application de la problématique au cas concret des voitures autonomes. Le troisième chapitre ***AristotleMD et l’intelligence artificielle*** apporte des informations sur les différents modes d’apprentissage des IA, qui seront utilisés par la suite dans le chapitre 4. Il présente la différence entre l’intelligence et la conscience, qui sera reprise au chapitre 5. Le chapitre 4 ***Les trois robots*** fait des liens entre les doctrines du chapitre 2 et les modes d’apprentissage du chapitre 3. Finalement, le chapitre 5 ***Attention, superintelligence*** représente la portée de l’évolution des systèmes qui ont été introduits dans les chapitres précédents.  J’ai également apprécié le souci de l’auteur dans l’identification des concepts importants du texte (groupes de mots surlignés en vert), par exemple avec *éthique des algorithmes* à la page 12. J’ai trouvé que cela facilitait la compréhension et permettait de mieux retenir les éléments importants.  Finalement, j’ai aimé la façon dont l’auteur a structuré l’information. Pour chaque problématique, il a procédé à la définition d’éléments clés. Puis, il a fait des liens entre ces éléments et des cas concrets (ex. : l’assistant *Aristotle*), ainsi qu’avec des analogies (ex. : la légende du roi Midas) pour une meilleure compréhension. |

**Annexe :**

**Éthique des algorithmes** : Déterminer quelles sont les règles ou les principes moraux à implanter aux technologies pour favoriser le bien.

**Normes conventionnelles** : Justifier la manière d’agir en fonction de la convention (groupe).

**Normes prudentielles** : Justifier la manière d’agir en fonction de nos préférences personnelles (individu).

**Normes morales** : Justifier la manière d’agir grâce à une conclusion neutre et équitable pour tous.

**Utilitarisme** : Principe moral qui dicte l’impartialité (chacun compte pour un) et qui motive ses actions afin de favoriser le plus grand bien (minimiser la souffrance).

**Déontologisme** : Principe moral dont le motif d’action est dicté en fonction de normes ou de devoirs à respecter, sans regard sur les conséquences potentielles.

**Psychologie morale** : Science qui étudie les causes des jugements moraux.

**Éthique normative** : Morale (normative) qui recherche les raisons pour justifier nos actions.

**Systèmes experts** : Modèle déductif qui nécessite une maitrise parfaite d’un problème avec toutes ses solutions.

**Apprentissage automatique** : Modèle inductif qui permet à un système l’apprentissage par lui-même, grâce à la découverte de règles.

**Apprentissage par renforcement** : Apprentissage automatique par essais, qui utilise la récompense lors de l’atteinte des objectifs.

**Apprentissage supervisé** : Apprentissage automatique qui utilise une profusion de données (exemples) qui seront induites en lois générales.

**Agents moraux artificiels** : Capacité de prendre ses décisions en sachant discerner le bien du mal, mais contrairement à l’agent moral, il n’est pas tenu responsable de ses actions (imputabilité).

**Patient moral** : Individu à qui il est possible de faire du bien ou du mal.

**Éthique de la vertu** : Justifier le motif de ses actions en fonction d’un modèle moral.

**AI étroites** : Elles peuvent accomplir une tâche précise, éventuellement mieux qu’un humain.

**AI générales** : Elles ont la capacité de transférer des connaissances acquises d’un domaine à un autre.

**Thèse de l’orthogonalité** : Il n’y a pas de connexion entre l’intelligence et le bien-fondé des buts qu’on se fixe.

**Problème de l’alignement** : S’assurer que les systèmes qu’on créer poursuivent les objectifs qu’on souhaite.

**Risques existentiels** : Risques pouvant causer la fin de l’humanité.

**Spécification directe** : Implantation de normes et de valeurs définies à l’avance.

**Normativité indirecte** : Demander ce que nous aurions aimé que le système fasse, ce qui donne les objectifs au système afin de s’adapter aux scénarios.

**Sens commun** : Bon sens (connaissance de base des lois du monde) permettant de prendre des décisions.

1. Références de quelques-uns des ouvrages clés qui sont cités par l’auteur dans le texte [↑](#footnote-ref-1)
2. Limitez-vous à une page et demie max [↑](#footnote-ref-2)