29,5

UQAR

Rimouski | Lévis

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 société
INF30007

2 décembre 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) et se questionner sur la société idéale. **Problématique** : Il existe plusieurs normes morales et une multitude de façons différentes de programmer. Ainsi, comment effectuer les bons choix de conception de ces technologies, permettant d'assurer le bien pour la société aujourd'hui et du futur ? Et dans quelle société désirons-nous vivre ? Quels sont les discriminants de notre société actuelle ? **Thèses** : Dans la conception des systèmes intelligents, il faut tenir compte des besoins de toute la population, incluant particulièrement ceux des groupes marginaux et sous-représentés. Dans cette optique, changer les « paramètres par défaut » de notre vision et de nos systèmes afin de mettre de l'avant leur représentation est une mesure concrète pour améliorer la société et la rendre plus juste. 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 possède donc une meilleure adaptabilité.

1

0,5

1

0,5

3

Références théoriques clés[1] ***2 pts***	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],** *Le langage de la nuit* (Livre de Poche, 2018) **[p.62],** Roman *Les dépossédés* (Ursula Le Guin, 1974) **[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],** *Qui peut sauver la morale? Essai de métaéthique* (Ithaque, 2019) **[p.69],** Livre *Émotions et valeurs* (Christine Tappolet, 2000) **[p.75],** *Moral Machines : Teaching Robots Right from Wrong* (Wendell Wallach, Colin Allen, 2009) **[p.77],** Livre *Deep Learning* (Yoshua Bengio, 2016) **[p.80],** « *The moral behavior of ethicists* », dans Justin Sysma et Wesley Buckwalter (dir.) **[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],** Apprentissage profond **[p.80]** **Voir en annexe de ce document pour les définitions (du livre) des notions présentées ici.**
Résumé analytique[2] ***10 pts***	Le chapitre ***Isaac et Ursula*** commence avec la description de l'auteur de science-fiction Issac Asimov qui a su présenter dans ses œuvres un côté obéissant des robots. Dans la section ***Imaginer entre hommes***, l'auteure Ursula Le Guin nous est présentée, celle-ci écrit également dans le genre de la science-fiction. La vision d'Ursula de la science-fiction est plus axée sur la recherche d'un idéal de société à travers les conditions humaines, alors que celle

[1] Références de quelques-uns des ouvrages clés qui sont cités par l'auteur dans le texte

[2] Limitez-vous à une page et demie max

d'Asimov se veut plus prédictive et d'anticipation des problèmes du futur. Le cas d'Asimov nous permet ainsi de comprendre l'impact d'une vision partielle d'un auteur, puisqu'il essaie de résoudre des problèmes techniques, sans réaliser qu'il transmet les mêmes préjugés et discriminants dans ses œuvres. La section ***Quel genre de programmation?*** vient faire un parallèle avec la disproportion féminine des membres du comité d'éthique allemand au sujet des voitures autonomes, discuté dans la partie un du livre. L'auteur se questionne alors sur les répercussions de cette disproportion. Il nous rappelle également que le domaine informatique s'est beaucoup restreint à un profil type, celui des hommes. La section ***Choisir le réglage par défaut*** présente le portrait actuel de société où les données traduisent souvent un historique de discrimination raciale, classisme et de genre. L'auteur rappelle qu'il est parfois préférable de changer nos réglages par défaut afin d'afficher la diversité, c'est pourquoi il se met à écrire en employant le féminin.

Le chapitre ***Métaéthique pour programmeuses*** présente des interrogations métaéthiques, soit si une programmation morale est meilleure qu'une autre, ou encore s'il est possible de construire un bon robot ? C'est dans la section ***Le défi du relativisme moral*** qu'on distingue les 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, par le fait que deux choses opposées ne peuvent être simultanément vraies. Le sujet du voile d'ignorance nous est ensuite présenté comme un bon moyen de prise de décisions, puisqu'il nous impose de réfléchir à une question en excluant notre position dans la société. La section ***Le défi de la perception morale*** discute du rôle de nos émotions au sein de nos processus moraux. Comment celles-ci influencent notre perception d'une situation et nos décisions. Un lien est finalement réalisé avec le potentiel d'améliorer le processus décisionnel des AMA par l'apprentissage émotionnel de personnes vertueuses.

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 un chercheur et des universitaires au sujet des sources de données à utiliser comme modèle de type vertueux pour l'apprentissage des IA. Cette section permet de comprendre la complexité derrière un apprentissage dit profond, où la compréhension est faite avec une hiérarchie de concepts, ayant plusieurs liens entre eux. La section ***Le triomphe modeste des robots vertueux*** se concentre sur la problématique d'identification

	des personnes vertueuses à utiliser comme modèle d'apprentissage. L'auteur rappelle les avantages des robots vertueux par rapport à leurs confrères utilitaristes et déontologistes, soit l'apprentissage par normativité indirecte, la représentation d'un compromis d'application, leur côté adaptatif et plus facilement réalisable. La section ***À la recherche d'une expertise morale*** discute des facteurs d'influence sur notre jugement, tels que la soumission à l'autorité ou encore nos humeurs. Ceci est d'ailleurs illustré par une expérimentation datant de 1960, où des volontaires ont infligé des chocs à des inconnus sous la pression de l'autorité. Nous constatons que le degré de scolarité ou de richesse ne produit pas de meilleurs candidats pour représenter le modèle vertueux et que c'est probablement l'intelligence collective qui pourrait trouver les meilleurs candidats. La section ***Éthique applicable*** discute de deux contextes applicatifs au modèle vertueux, soit le dilemme décisionnel des voitures autonomes et celui de la reconnaissance des situations de l'AMA Aristotle. Dans le premier cas, un processus décisionnel prenant source d'une proportionnelle et d'un critère de niveau de confiance d'une population de vertueux représenterait la solution optimisée, alors qu'un apprentissage utilisant des exemples émotionnels serait favorisé dans le second. Dans le chapitre ***Conclusion Une cité incroyable,*** l'auteur discute du livre « Ceux qui partent d'Omelas », comment ce dernier effectue une critique indirecte de l'utilitariste, puisqu'il permet de justifier une souffrance si la quantité de bien-être y est plus abondante. Finalement, l'auteur nous rappelle que l'importance n'est pas uniquement le type de robot que nous voulons programmer, mais davantage l'état de la société dans laquelle nous voulons vivre.
Citations clés ***3 pts***	« Isaac définissait la sf comme « cette branche de la littérature qui s'intéresse aux impacts du progrès scientifique sur les êtres humains[21] », cherchant notamment à anticiper les problèmes qu'on risque d'affronter dans le futur et à envisager des solutions. De son côté, Ursula insistait pour dire que la sf était moins prédictive que descriptive. Et ce que cette description révèle en creux, c'est combien la perspective adoptée par un auteur est partielle, et partiale. Dans les récits de robots, on oublie vite les femmes, les pauvres et les personnes racisées. On reproduit allègrement les hiérarchies et les discriminations sociales. Bref, s'il y a une chose à retenir d'Ursula, c'est qu'à trop s'inquiéter de la réplication des trombones, on néglige celle du patriarcat. » **[p. 63]** « Lorsque j'ai lu, adolescent, *Le sorcier de Terremer*, une chose m'a frappé. Comme plusieurs personnages centraux du livre, le héros imaginé par Ursula avait la peau rouge brun. J'ai pu, le temps d'un récit, associer l'héroïsme à autre chose qu'à des chevaliers blancs. » **[p. 68]**

	« Chez l'être humain, la prise de décision morale implique divers processus cognitifs plus ou moins conscients : on perçoit une situation, on évalue les raisons morales en présence, on se heurte à d'éventuels dilemmes, on détermine la bonne chose à faire, et on agit. Or, contrairement à l'humain qui peut manquer de volonté (acrasie), un robot n'a aucun mal à passer de la décision à l'action. En revanche, la première étape de ce processus, soit celle de la perception morale, semble une véritable gageüre pour un robot. » **[p. 75-76]** « En mettant les vertueuses à profit pour calibrer les paramètres éthiques des robots, on risque moins de faire subir aux générations futures nos aveuglements moraux. Les ama s'inspireront des vertueuses de leur temps. C'est même tout le projet de la programmation arétaïque : parvenir à harnacher l'expertise morale des gens. Ce qu'il faut, c'est choisir de bons exemples pour faire de bons robots. C'est apprécier la personnalité morale des gens et s'inspirer de leurs vertus. » **[p. 89]** « Avec quelle sorte de robots voulons-nous vivre, certes, mais surtout dans quel monde souhaitons-nous les côtoyer ? » **[p. 93]**
Idées importantes en lien avec le cours et appréciation personnelle ***7 pts***	En plus de la couverture, par le livre, des concepts déjà décrits dans le résumé de lecture 1, il couvre dans cette partie ceux de la discrimination entre les genres, les ethnicités et les classes. Il couvre également de façon plus détaillée l'éthique de la vertu et il introduit les concepts de société idéale passant par un changement de nos constructions sociétales. Je trouve encore une fois que la force de l'auteur est sa capacité à transmettre l'information de manière logique et cohérente à travers les chapitres. Alors que la première partie était davantage une ~~introduction~~ entrée en matière avec la mise en contexte, les définitions clés et des problématiques, cette deuxième partie est orientée sur une introspection de la solution qu'est l'utilisation d'un apprentissage utilisant comme modèle l'éthique de la vertu. Sans oublier son ouverture rappelant les problématiques discriminatoires de notre société actuelle, afin de ne pas répéter nos mêmes erreurs dans le futur. Cette structuration du livre par ses liens entre les chapitres ainsi que par son retour aux situations initiales (cas de la voiture autonome, de l'AMA Aristotle) permet une couverture complète de la problématique initiale tout en conservant l'intérêt du lecteur. Par exemple, le chapitre ***Isaac et Ursula***, qui démontre comment la société fait la promotion de biais discriminatoires et comment il est

6,5

	important de changer notre vision par défaut de celle-ci, est un bon complément à l'exemple initial du bus des jours fériés et de ses passagères. Le chapitre ***Métaéthique pour programmeuses***, qui discute des différentes théories morales rencontrées en programmation et des bienfaits des émotions dans notre processus de prise de décisions morales, est un bon complément aux chapitres précédents ***Les trois robots*** et ***Attention, superintelligence***, qui ont introduit l'éthique de la vertu et sa distinction. Il en va de même avec le chapitre ***Faire des robots vertueux***, qui discute du contexte d'implémentation de l'apprentissage par modèle vertueux en utilisant l'intelligence collective pour trouver les candidats et qui explique les bienfaits de ce genre de programmation par rapport aux autres (utilitaire, déontologique). Finalement, le chapitre ***Conclusion Une cité incroyable*** permet d'apporter une ouverture pertinente en soulignant l'importante de ne pas focaliser uniquement notre attention sur les principes de développement, mais de viser un objectif plus grand qu'est de développer d'abord une société plus juste et équitable, favorisant le bien-être collectif. Pour conclure, j'ai aimé comment l'auteur a su apporter des analogies afin de faciliter la démonstration de concepts. Par exemple, en utilisant la situation de la société utopique, mais égoïste, décrite dans le livre « Ceux qui partent d'Omelas », pour illustrer l'importance de garder la recherche d'un idéal de société en tant que priorité.

Annexe :

Relativisme moral : « [...] —une position métaéthique— estiment ainsi qu'il n'y a pas de normes morales universelles. » **[p.70]**

Relativisme de la locutrice : « [...] soutient que c'est l'attitude d'approbation ou de désapprobation d'un individu qui détermine la vérité d'un énoncé moral. » **[p.70]**

Relativisme culturel : « [...] soutient que c'est l'approbation de la communauté—avec ses normes conventionnelles—qui détermine la vérité d'un énoncé. » **[p.71]**

Réalisme : « [...] considèrent pour leur part qu'il existe bel et bien des vérités morales. » **[p.71]**

Voile d'ignorance : « Celle-ci consiste à se demander quel serait notre avis sur une question si on ignorait notre position dans la société. » **[p.73]**

Apprentissage profond : « [...] essaye de comprendre le monde d'après une hiérarchie de concepts, chacun d'entre eux se définissant par une multitude de relations avec des concepts plus simples. On parle d'apprentissage *profond* parce que l'information est traitée à l'aide d'algorithmes structurés selon de multiples couches de « neurones artificiels ». » **[p.80]**