

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

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. |
| **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],  Apprentissage profond [p.80]  \*\*Voir en annexe de ce document pour 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 optimisée dans le premier cas, alors que l’AMA apprendrait la reconnaissance des situations conversationnelles d’après les exemples des situations émotionnelles des personnes vertueuses.  Dans le chapitre ***Conclusion Une cité incroyable,*** l’auteur discute du livre « Ceux qui partent d’Omelas », comment celui-ci effectue une critique indirecte de l’utilitariste, puisque celui-ci permet de justifier une souffrance si la quantité de bien-être y est plus abondante. Le livre peut donner sur plusieurs interprétations, et l’auteur choisi de croire que c’est les vertueux qui représente le contingent à cette société utilitariste qu’est Omelas. Finalement, il nous rappelle que l’importance n’est pas uniquement le type de robot que nous voulons programmer, mais d’avantage l’état de la société dans laquelle nous voulons vivre qui est important. |
| **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 humains21 », 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 en lien avec notre cours déjà décrit 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 plus en détail l’éthique de la vertu qu’en première partie. Également, les concepts d’idéale de société passant par plus d’inclusion et par la mise en avant des gens marginalisés dans nos modèles de société.  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 d’avantage une introduction en matière avec la mise en contexte, les définitions clés, les problématiques et des exemples concerts à celles-ci, cette deuxième partie est d’avantage une introspection de la solution, donc d’une programmation d’IA utilisant des modèles relatant l’éthique de la vertu, ainsi qu’une ouverture afin de ne pas oublier les problématiques discriminatoires de notre société actuelle afin de ne pas répéter nos mêmes erreurs dans le futur. En effet, le chapitre ***Isaac et Ursula*** démontre bien le contraste idéologique entre deux auteurs, comment la société fait la promotion de biais discriminatoires et comment il est important de changer notre vision par défaut de la société pour le bien de tous. Le chapitre ***Métaéthique pour programmeuses*** discute de théories morales rencontrées en programmation et des bienfaits que peut apporter les émotions dans notre processus de prise de décisions morales. Le chapitre ***Faire des robots vertueux*** discute du contexte implémentatif de système d’IA utilisant l’approche d’apprentissage par modèles virtueux, par exemple en utilisant l’intéligence collective pour trouver les modèles, les bienfaits de ce genre de programmation par rapport aux deux autres (utilitaire, déontologique) et nous explique comment les personnes virtueuses se distingues par rapport aux autres. Finalement, le chapitre ***Conclusion Une cité incroyable*** nous rappelle la leçcon importante qu’en focusant uniquement notre attention sur le développement de technologies, nous risqu’on de perdre de vue l’objectif primordiale qu’est de développer d’abord une société plus juste et équitable, favorisant le bienêtre collectif.  Pour conclure, j’ai aimé comment l’auteur a su apporter des situations externes afin de facilité la démonstration de ses concepts, par exemple en utilisant la situation de la société utopique, mais égoïste, décrit dans le livre « Ceux qui partent d’Omelas », pour illustrer l’importance de garder la recherche d’idéale de société comme 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)**

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)