30,5/31

UQAR

Rimouski | Lévis

Résumé de lecture 1 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

21 octobre 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***	13 octobre 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

1

0,5

1

0,5

3

	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] ***2 pts***	Site web *Moral Machine experiment*, (MIT) [p.22], *La voiture qui en savait trop*, (Humensciences, 2019) [p.25], *Moral Machines : Teaching Robots Right from Wrong*, (Oxford University Press, 2010) [p.32], *Déclaration de Montréal pour un développement responsable de l'IA* (2018) [p.34], *Homo Deus : une brève histoire du futur*, (Albin Michel, 2017) [p.36], *Federal Ministry of Transport and Digital Infrastructure, Ethics Commission: Automated and Connected Driving* (2017) [p.37], *Les robots et le mal*, (Desclée et Brouwer, 2018) [p.38], *Ethics for Robots*, (Routledge, 2018) [p.41], *Technology and the Virtues*, (Shannon Vallor, 2016) [p.45], *Déclaration d'Asilomar* (2017) [p.52], *Superintelligence*, (Dunod, 2017) [p.53], *The Cambridge Handbook of Artificial Intelligence*, (Cambridge University Press, 2014) [p.54], *Quand la machine apprend*, (Odile Jacob, 2019) [p.56]
Concept clés (notions définies, organismes clés) - indiquer les définitions importantes avec la page.	Éthique des algorithmes (p.12), Normes conventionnelles, Normes prudentielles et Normes morales (p.14), Utilitarisme (p.17), Déontologisme (p.21), Psychologie morale et Éthique normative (p.25), Apprentissage automatique et Systèmes experts (p.30), Apprentissage par renforcement et Apprentissage supervisé (p.31), Agents moraux artificiels (p.32), Patient moral (p.34), Éthique de la

2

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

3 pts	vertu (p.44), AI étroites et AI générales (p.50), Thèse de l'orthogonalité et Problème de l'alignement (p.52), Risques existentiels (p.53), Spécification directe(p.54), Normativité indirecte et Sens commun (p.55). **Voire en annexe de ce document pour mes définitions (qui résument les définitions du livre) des notions présentées ici.**
Résumé analytique[2] ***10 pts***	Tout d'abord, le premier chapitre ***Introduction :le bus des jours fériés*** nous introduit au concept des normes morales qui permettent de justifier une façon d'agir en appliquant une conclusion juste et équitable pour tous. L'auteur a fait un lien avec son passage dans le bus un jour férié, où il avait observé que la composition des gens était différente de l'habitude et où il s'était noté de prendre en considération ces personnes, plus marginales, dans ses conceptions. Le second chapitre ***Le vieillard ou l'enfant*** présente la manière dont l'auteur donne ses cours d'éthique. Il vient confronter ses étudiants à leurs propres intuitions personnelles, grâce à l'ambiguïté générée à travers les deux variantes du dilemme du tramway. Il vient ensuite montrer que cette ambiguïté prend toute son importance dans le développement des futures technologies comme dans la programmation des voitures autonomes. Il présente ensuite les sources de biais (intuitions personnelles) de gens ayant répondu à l' étude *Moral Machine experiment* du MIT sur des cas de choix moraux dans des scénarios d'accident de voiture. L'auteur explique ensuite qu'il est nécessaire d'utiliser une approche d'éthique normative dans la conception puisque cette dernière favorise l'utilisation de normes morales pour la justification de nos actions, plutôt que des biais. Il faut cependant se questionner sur le choix de la norme à utiliser, parce qu'elle aura de multiples conséquences, comme il est possible de l'imaginer dans les cas d'accidents de voitures autonomes. Le troisième chapitre ***Aristotle[MD] et l'intelligence artificielle*** débute avec l'exemple d'*Aristotle*, une technologie d'IA utilisée comme assistant personnel pour les enfants. L'auteur démontre que la portée d'influence d'une telle technologie peut représenter un grand

3

9,5

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

	enjeu au sein de nos relations humaines. Il fait ensuite l'historique des différents moyens utilisés dans le développement des AI. Ainsi, il est possible de voir que nous avons débuté avec la programmation de systèmes experts nécessitant la définition de règles à respecter. Puis, nous nous sommes dirigés vers une programmation plus inductive où l'IA apprend par elle-même, soit en étant récompensée ou par association à partir d'exemples. Cette dernière approche étant la plus utilisée de nos jours. Cela engendre ainsi la possibilité pour certains systèmes d'IA à se classer dans la catégorie des agents moraux artificiels, soit d'avoir une capacité de prise de décision pouvant occasionner du mal. L'auteur vient donc souligner l'importance de discerner la conscience de l'intelligence, car c'est cette première qui permet de ressentir et de favoriser le bien, alors que la seconde est une capacité d'atteindre des objectifs pouvant favoriser le bien comme le mal. Il est dès lors possible de voir l'importance du développement d'un mécanisme de conscience chez les IA. Le quatrième chapitre ***Les trois robots*** commence avec un résumé des conclusions d'un rapport pour le ministère allemand des Transports, qui prend position sur les objectifs des voitures autonomes en cas d'accident. Le rapport ne permet aucune décision discriminatoire (ex. : favoriser les jeunes). À partir de ce cas, l'auteur présente le lien avec l'approche déontologique, puisque cette dernière permet le respect de normes morales sans considérer les conséquences. Il explique également que l'implémentation de ce principe viendra rapidement laborieuse, nécessiterait la définition de plusieurs règles afin de comprendre tous les cas possibles. Il s'en suit de l'approche utilitariste qui est elle aussi complexe, puisqu'elle nécessite de bien mesurer les conséquences futures des décisions possibles, ce qui est impossible. L'auteur nous introduit finalement à l'approche de l'éthique de la vertu, c'est-à-dire l'utilisation d'exemples d'un modèle considéré moral dans la programmation d'un système, afin qu'il les associe lui-même à des lois générales. On comprend que cette approche se distingue des deux autres par son caractère adaptatif et sa simplicité. Finalement, l'auteur indique que le défi est davantage dans la définition d'un mécanisme de prise de décisions du système, établissant un seuil entre la sécurité et l'efficacité. Le cinquième chapitre ***Attention,***

	superintelligence parle des principes soulevés à partir d'une conférence sur les futurs systèmes de superintelligences, qui auront la possibilité de surpasser les capacités humaines grâce au transfert de connaissances. Cette évolution des systèmes autonomes à venir apporte ainsi la problématique de leur contrôle et de dérapages potentiels en cas de mauvaise définition des objectifs. C'est ainsi qu'on fait une analogie avec la fausse bonne idée dans la légende du roi Midas et la superintelligence productive de trombones. Il est ainsi expliqué que la définition de règles de départ strictes peut induire une erreur qui aura de graves conséquences dans le futur. De plus, il y a toute la problématique d'incertitude sur nos concepts actuels en lien avec nos idées reçues et nos préjugés. Finalement, l'approche permettant de minimiser les risques sur les objectifs d'une superintelligence serait l'utilisation de la normativité indirecte qui laisserait le soin à cette superintelligence de changer les objectifs au besoin.
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) « Dans ce livre, la question n'est donc pas de savoir comment les gens pensent qu'on devrait faire la morale aux robots. Elle est bien différente, et plus abyssale. Comment faire la morale aux robots? » (p. 25) « Mais les IA semblent instaurer une rupture radicale: les robots peuvent être intelligents sans être conscients. Dès lors, du point de vue moral, on peut se demander ce qui compte le plus entre l'intelligence et la conscience. Si on définit l'intelligence comme la capacité d'atteindre un objectif, il est clair que sa valeur est instrumentale, puisqu'elle peut aussi bien se mettre au service du bien que du mal, [...] » (p. 35) « C'est, en définitive, ce qu'il faudrait enseigner aux

3

	robots. Qu'ils soient déontologiques, utilitaristes ou vertueux, les bons robots devraient toujours agir de la bonne manière au bon moment. » (p. 46) « Il s'agit plutôt d'établir un seuil: quand doubler? Quand la situation sera-t-elle *suffisamment* sécuritaire? C'est cet équilibre subtil entre sécurité et efficacité qu'il faut traduire en algorithme. Pour tout dire, à ce jour, autant les robots déontologiques, utilitaristes que vertueux semblent encore incapables d'offrir une solution applicable et satisfaisante.» (p. 47) « Il en profite pour en rajouter: «Pensez-y, l'IA est la dernière invention que les humains auront besoin de créer. Les machines seront alors de meilleurs inventeurs que l'on ne l'est.» » (p. 50) « D'où l'analogie proposée par Bostrom: une superintelligence pourrait très bien atteindre les objectifs qu'on lui aurait assignés, mais elle ne pourrait rien contre la stupidité de ces objectifs. » (p. 51) « Car si les superintelligences fonctionnent comme des miroirs grossissants, une leçon semble s'imposer: formuler une règle, c'est courir le risque de se tromper de règle. » (p. 56)
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

7

	déontologisme. De plus, il permet l'application de la problématique au cas concret des voitures autonomes. Le troisième chapitre ***Aristotle[MD] 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.