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. Il travaille à l’université de Montréal et fait partie du Centre de recherche en éthique et se spécialise dans l’éthique de l’intelligence artificielle. |
| **Genre de l’ouvrage ou de l’article**  ***0.5 pt*** | Livre, genre morale et philosophie. |
| **Objectif du texte(1pt), problématique (1pt) et thèses soutenues (1pt)**  ***3 pts*** | **Objectif**: L’auteur vise à sensibiliser sur l’approche et les choix de programmation quant aux normes morales dans le développement des technologies autonomes, tel que les intelligences artificielles par exemple.  **Problématique** : Il existe plusieurs règles morales et une multitude de façons différentes de programmer, sans compter toutes les sources de biais possible. Ainsi, comment effectuer les bons choix permettant d’assurer le bien pour la société aujourd’hui et dans le futur ?  **Thèses** : Dans la conception des systèmes intelligent, il faut tenir compte des besoins de tous, incluant les groupes marginaux.  Dans la conception, il faut valoriser une approche de morale normative afin de bien justifier nos actions plutôt que de suivre nos propres intuitions morales qui sont des sources de biais.  Il n’y a pas de connexion entre intelligence et le bien fondé des objectifs d’une AI, il est donc préférable d’utiliser une approche inductive dans sa programmation avec ce que nous aurions aimé qu’elle fasse afin de lui laisser choisir elle-même ses objectifs et s’adapter en cours de route. |
| **Références théoriques clés[[1]](#footnote-1)**  ***2 pts*** | Site web Moral Machine experiment, MIT (p.22)  Livre La voiture qui en savait trop, Jean-François Bonnefon et al (2019)  Moral Machines : Teaching Robots Right from Wrong, Oxford University Press (2010)  Déclaration de Montréal pour un développement responsable de l’IA (2018)  Homo Deus : une brève histoire du futur, Albin Michel (2017)  Federal Ministry of Transport and Digital Infrastructure, Ethics Commission: Automated and Connected Driving (2017)  Les robots et le mal, Desclée et Brouwer (2018)  Ethics for Robots, Routledge (2018) (p.41)  Technology and the Virtues, Shannon Vallor (2016) (p.45)  Superintelligence, Nick Bostrom (p.48)  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.**  ***3 pts*** | Éthique des algorithmes (p.12), normes conventionnelles, prudentielles et morales (p.14), utilitarisme (p.17), déontologisme (p.21), psychologie morale et éthique normative (p.25), apprentissage automatique (p.30), systèmes experts(p.30), apprentissage par renforcement et apprentissage supervisé (p.31), agents moraux artificiels (p.32), patient moral (p.34), éthique de la vertu (p.44), AI étroites et 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 le sens commun (p.55). |
| **Résumé analytique[[2]](#footnote-2)**  ***10 pts (environ 300-400 mots pour ce travail)*** | L’auteur présente, à travers ses cinq premiers chapitres, un cheminement sur les différentes problématiques que constitut la question : quelles normes morales utilisées et quelle approche de programmation utilisée dans le développement des intelligences artificielles ?  Tout d’abord, le premier chapitre (Introduction :le bus des jours fériés) nous introduit aux normes morales qui permettent de justifier une façon d’agir en appliquant une conclusion juste et équitable pour tous, plutôt que de se limiter à un groupe d’individus. Il fait un parallèle avec son propre passage dans le bus un jour férié où il observe que la composition des gens n’est pas celle qu’il a l’habitude de voir et où il prend note de tenir compte de ses personnes marginales dans ses conceptions.  Dans le second chapitre (Le vieillard ou l’enfant), il discute de la manière dont il donne ses cours d’éthique. Il vient ainsi confronter ses étudiants à leurs propres intuitions personnelles grâce aux deux variantes du dilemme du tramway. Il démontre également que l’ambigüité à répondre proprement à ses dilemmes prend toute son importance dans le développement des futures technologies comme la programmation des voitures autonomes. Il montre également à travers une étude à l’échelle mondiale du MIT que les gens ont des sources de biais (intuitions personnelles) et qu’il est nécessaire d’utiliser une approche d’éthique normative dans la conception. Cette dernière favorise l’utilisation de normes morales pour justifier nos actions et le questionnement sur leur bien fondé ainsi que sur leur portée.  Dans le troisième chapitre (AristotleMD et l’intelligence artificielle), il commence avec l’exemple d’Aristotle, une technologie d’IA utilisée comme assistant personnel pour les enfants. Cela lui permet de démontrer la portée d’influence de cette technologie à travers la société. En nous rappelant que l’intelligence de cette technologie découle de la technique humaine, il fait ensuite l’historique des différents moyens utilisés dans le développement des AI. Ainsi, il est possible de voir que nous avons commencé avec des définitions beaucoup plus fonctionnalistes de l’intelligence, comme les capacités prédictives et la capacité d’atteindre des objectifs. Puis, il décrit l’évolution des méthodes de programmation des systèmes, les premiers étant d’avantage des séries de règles à respecter (systèmes experts), puis que nous sommes passé du côté d’une programmation plus inductive où l’IA apprend par elle-même. Dans ce deuxième mode de programmation, elle peut apprendre en étant récompensé par renforcement ou elle peut faire ses propres associations de règles à partir d’exemples, cette dernière approche étant la plus utilisée de nos jours. Cela engendre ainsi la possibilité de certains AI à se classer dans la catégorie des agents moraux artificiels, soit d’avoir une capacité de prise de décision pouvant occasionner du mal. Bien que le facteur d’imputabilité reste encore vague, il n’en demeure pas moins que les AI ont différents degré d’agentivité morale artificiel en fonction de leur niveau d’autonomie et de leur niveau d’incidence sur la polulation. L’auteur rappel l’importance de discerner la conscience de l’inteligence, cette première étant la capacité de ressentir et d’avoir des expériences agréables ou désagréables alors que la seconde est une capacité d’atteindre des objectifs qui peuvent être bien ou mal. Il est dès lors possible de voir l’importance de développer la conscience des AI, puisque les systèmes actuels n’en possèdent pas.  Le quatrième chapitre (Les trois robots) commence avec un résumé du ministère des transport Allemand qui prend position sur les objectifs des voitures autonomes et oû aucune discrimination n’est possible entre les gens quand aux priorités accordés en cas d’accident. Il est possible de voir que c’est une programmation avec des bases déontologiques qui permettrait d’expliquer l’orientation du comité des transport, puisque cette dernière respecte des normes morales peut importe les conséquences finales et qu’elle pourrait utiliser le hasard comme prise de déscision. On explique que l’implémentation de ce système nécessiterait la définition de plusieurs normes et d’exception et que cela viendra rapidement laborieux. Il s’en suit l’approche de programmation avec des bases utilitaristes, qui semble plus simple en favorisant l’approche d’un calcul de bien-être et en remettant chaque avenue possible en nombre d’unité de bienêtre, ce qui évite la définition de plusieurs règles et l’utilisation d’une programmation autonome par apprentissage. Cependant, en pratique cette option devient aussi complexe, puisqu’il est impossible de prévoir les conséquences futures de notre prise de position et que cette méthode permettrait de prendre des déscisions défavorable sur un humain du moment que la valeur d’une vie humaine est chiffrée en termes de bienfait. Finalement, l’auteur discute de l’approche de programmation avec des bases vertueuses, c’est-à-dire d’utiliser l’apprentissage autonome supervisé avec des exemples d’un modèle considéré comme positif, cette approche se distingue des deux autre par son caractère adaptatif aux situation et par une plus grande simplicité. Cependant, il est rappeler qu’alors que nous discutons de prises de déscisions en cas d’accident en cas exceptionnels, le défit d’avantage le mécanisme de prise de déscision du système, parce qu’assurer la sécurité absolu d’un système diminue grandement son efficacité et qu’il y a un seuil à choisir.  Le cinquième chapitre (Attention, superintelligence) oû l’auteur nos décrit une conférence d’un philosophe qui explique le développement à venir des superintelligences, qui auront la possibilité de surpasser les capacités d’évolution humaines en ayant la capacité de transférer leurs connaissances d’un domaine à un autre autre. Cette évolution des systèmes autonomes à venir apporte ainsi la problématique de contrôle et de dérapages possible en cas d’une 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 trombonnes finisant par exterminée l’humanité pour les utilisé comme matière première. Il devient donc évident que la définition de règles de départ concrètes au système peu induire une erreur qui aura de graves conséquences grave dans le futur. Il y a également la problématique d’incertitude sur nos concepts actuels en lien avec nos idées reçues et nos préjugés, ceux-ci risquant d’être incompatible avec les nouveaux concepts dans le futur. Ainsi, l’approche permettant de minimiser les risques sur les objectifs d’une superintéligence serait de suggérer des concepts de départ que nous jugeons actuellement moralement acceptable, en lui demandant de faire ce que nous aurions aimé qu’elle fasse et en lui laissant le soin de changer les objectifs pour les adaptés s’ils deviennent incomptable avec les situations futurs. |
| **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)  « Ils [les étudiants] constatent aussi qu’on a tous des intuitions contradictoires et qu’on peut toujours trouver des philosophes pour les justifier-ça, ce n’est pas forcément une bonne nouvelle. Certains éprouvent peut-être un petit vertige existentiel devant les limites de la rationalité humaine. » (p. 21)  « 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 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)  « 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, trois doctrines éthiques que son l’utilitarisme, le déontologisme et l’éthique de la vertu, l’importance de distinguer la psychologie morale de l’éthique normative ainsi que l’importance de la notion d’inclusion afin d’apporter une conclusion est neutre et équitable pour tous dans ses décisions.  J’ai apprécié la lecture de ce livre pour plusieurs raisons qui seront détaillées dans les phrases suivantes. Tout d’abord, je trouve que l’auteur a su transmettre l’information de manière logique et ordonnancée. En effet, le premier paragraphe (Introduction : le bus des jours fériés) permet l’introduction aux concepts de base tel que l’éthique des algorithmes et nécessaire à la compréhension de la problématique sur le choix des normes morales à utiliser pour le développement des nouvelles technologies. Ceci permet une transition naturelle au deuxième chapitre (Le vieillard ou l’enfant) qui vient ajouter des définitions supplémentaires telles que la confrontation entre l’utilitarisme et le déontologisme et une ouverture d’un cas concret avec les voitures autonomes. Le troisième chapitre (AristotleMD et l’intelligence artificielle) traite des différents types d’apprentissage dans des AI et la différence entre l’intelligence et conscience. Le chapitre 4 (Les trois robots) reprend les éléments des trois chapitres précédents alors qu’il vient faire un retour sur les doctrines éthiques appliquées au cas des voitures autonomes et fait le lien entre les doctrines éthiques et les principes de programmation (méthodes d’apprentissage) utilisés pour l’implémentation de chacune d’elle. Finalement, le chapitre 5 (Attention, superintelligence) est une gradation importante des chapitre 3 et 4, puisqu’il apporte une approche futuriste des systèmes qui auront une autonomie complètes et une indépendance complète des humains. C’est précisément cette approche qui permet de comprendre toute la portée des choix réalisé et la démarche développée dans les chapitres précédent.  J’ai grandement apprécié le souci de l’auteur à identifier les concepts importants au sein du texte. En effet, il a identifié ceux-ci avec un élément visuel (surlignage vert), par exemple avec *éthique des algorithmes* à la page 12. J’ai trouvé que cela facilitait la compréhension et permettait de mieux différencier les éléments importants du texte.  Finalement, j’aime la façon dont l’auteur a su transmettre les informations. Pour chaque problématique, il a procédé à des définition d’éléments clés à la compréhension (ex. : mots surlignés en vert), puis il a fait des liens entre ceux-ci et des cas concret (ex. : l’assistant Aristotle) ainsi qu’avec des analogies (ex. : la légende du roi Midas dans la problématique de contrôle des IA) pour une meilleure compréhension. |

<https://fr.wikipedia.org/wiki/Martin_Gibert>

**Éthique des algorithmes** = Quelles règles à implanter aux machines pour favoriser le bien. (p.12)

**Normes conventionnelles** = Justifier la manière d’agir en fonction de la convention (groupe). (p.14)

**Normes prudentielles** = Justifier la manière d’agir en de nos préférences personnelles (individu). (p.14)

**Normes morales** = Justifier la manière d’agir lorsque la conclusion est neutre et équitable pour tous (tout le monde). (p.14)

**Utilitarisme** = Principe morale qui dicte l’impartialité (chacun compte pour un) et motiver les actions afin de favoriser le plus grand bien, ou à minimiser la souffrance. (p.17)

**Déontologisme** = Principe morale dont le motif d’action est dicté en fonction de normes ou de devoir à respecter sans regard des conséquences potentielles. (p.17)

**Psychologie morale** = Étude (science, donc descriptive et recherche de causes) des jugements moraux. (p.25)

**Éthique normative** = Morale (normative) recherche les raisons pour justifier nos actions, et que celles-ci justifient le bien. (p.25)

**Systèmes experts** = Modèle déductif. Maitrise parfaite d’un problème avec toutes ses solutions (Systèmes experts). (p.30)

**Apprentissage automatique** = Modèle inductif. Apprentissage, découverte des règles par elle-même. (p.30)

**Apprentissage par renforcement** = Apprentissage automatique qui utilise la récompense lors de l’atteinte des objectifs, par essaies. (p.31)

**Apprentissage supervisé** = Apprentissage automatique qui utilise une profusion de données (d’exemples) qui seront induite en lois générales. (p.31)

**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). (p.32)

**Patient moral** = Individu à qui il est possible de faire du bien ou du mal. (p.34)

**Éthique de la vertu** = Justifier le motif de ses actions en fonction d’un modèle. (p.44)

**AI étroites** = Elles peuvent accomplir une tâche précise, éventuellement mieux qu’un humain. (p.50)

**AI générales** = Elles ont la capacité de transférer des connaissances acquises d’un domaine à un autre. (p.50)

**Thèse de l’orthogonalité** = Pas de connexion entre intelligence et les buts qu’on se fixe (p.52)

**Problème de l’alignement** = S’assurer que les systèmes qu’on crée poursuivent les objectifs qu’on souhaite. (p.52)

**Risques existentiels** = Risques pouvant causer la fin de l’humanité. (p.53)

Spécification directe = Implantation de normes, valeurs définies à l’avance. (p.54)

Normativité indirecte = Demander ce que nous aurions aimé qu’elle fasse, on déplace les objectifs sur l’AI pour qu’elle s’adapte aux scénarios que nous n’aurions pas imaginé. (p.55)

Sens commun = Bon sens (connaissance de base des lois du monde) permettant de prendre des décisions. (p.55)

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)