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 (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** : La programmation xxx ne |
| **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)*** |  |
| **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*** |  |

<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)