LETTRES SORBONNE UNIVERSITÉ

2024

Ethique de l'Intelligence Artificielle en sciences affectives

Kouamé Kouassi
M2 Langue & Informatique

Professeure
Laurence Devillers
Sorbonne Université Paris IV

1 Introduction

L'idée d'une intelligence artificielle naît des philosophes tels que Leibniz avec sa machine de calcul théorique appelée *Calculus ratiocinator* en 1666. C'est une machine permettant de démeler le vrai du faux dans une discussion dans un langage philosophique universel appelé *Caractéristique Universelle* [Leibniz,1666].Il a donc imaginé un procédé automatique couplant ce langage et un algorithme qui est en mesure de vérifier la véracité d'une déclaration et aussi d'énoncer des théories. Cela représente donc l'une des prémisses de ce qu'on nomme la machine à raisonner. Dès le 20e siècle, les scientifiques commencent à explorer l'idée d'un cerveau électronique pour faire allusion à l'ordinateur doté d'un processeur. Au cours de l'évolution de l'informatique, les scientifiques sont parvenus à inventer des interfaces hommes-machines ou *IHM* permettant l'interaction de l'homme et de la machine au travers d'un ordinateur à l'écrit, ou d'un robot dans le domaine de la robotique industrielle. C'est dans cette optique que le *chatbot* ou dialogueur ou encore agent conversationnel est développé. Les agents conversationnels sont des systèmes programmés pour échanger de façon autonome avec un humain. Aux prémisses de cette création, les agents conversationnels étaient programmés sur la base d'un système de règle (système expert). De nos jours, des agents conversationnels comme *ChatGPT* (OpenAI 2023) sont boostés par l'intelligence artificielle.

2 Bref aperçu du fonctionnement de l'IA

L'Intelligence Artificielle est un ensemble de théories et de techniques mises en oeuvres en vue de réaliser des machines capables de simuler les capacités cognitives des hommes. Il existe une IA faible et une IA forte. De nos jours, nous disposons essentiellement d'IA faible qui est soit symbolique, c'est-à-dire qui exploite des connaissances par l'application de relations logiques entre des propositions unitaires, soit connexionniste ou neuronale en utilisant des techniques basées sur les réseaux de neurones, l'apprentissage profond et les statistiques pour mieux prédire. Il faut noter que l'IA symbolique présente un problème ; elle ne permet pas de mieux modéliser les subtilités de la pensée humaine parce qu'il fonctionne sur un système à base de règles et que le raisonnement humain ne suit pas la logique d'un raisonnement formel. En d'autres termes, aucun langage formel ne pourrait interpréter précisément le raisonnement d'un humain parce que l'intelligence humaine suit une règle d'inférence non démonstrative (Jean Luc Roulin, 2006). Il y a aussi le fait que les langages formels ne savent pas reconnaître de nouvelles informations. Typiquement un cas de dialogue avec un *chatbot* créé avec un système symbolique ne saurait interpréter de nouvelles informations qui ne sont pas mentionnées dans le contexte de la discussion. L'humain lorsqu'il s'exprime avec son semblable est capable de traiter les nouveaux thèmes adjoints à une discussion quand ceux-ci sont utiles et qu'ils concourrent à sa discussion. Ce que ne pourrait faire une IA symbolique (principe d'explosion). Dans un tel contexte, l'IA symbolique aura des difficultés à sélectionner l'information et générer des réponses inadaptées.Tout cela continuerait donc avec l'IA connexionniste qui se trouve toujours sous l'aile de l'IA faible. Pour améliorer donc le *chatbot* il faut intégrer la notion de pensée dans le langage et aussi le fait que cette IA soit capable de ressentir de vrais sentiments et une compréhension d'elle-même. Cela serait possible avec une IA forte.

3 Des problèmes éthiques apparaissent

Les applications de l'IA dans nos vies ne sont pas sans conséquences négatives notamment en ce qui concerne l'éthique. Il existe une éthique ancienne et une éthique moderne. Dans l'éthique moderne on retrouve les notions de norme, d'obligations morales, de devoir. On définirait en philosophie ce terme de la manière suivante: "Science qui traite des principes régulateurs de

l'action et de la conduite morale". Selon l'éthique ancienne, l'éthique se fonde sur le terme *eudemonia* qui est une doctrine selon laquelle le but de l'action est le bonheur procuré par l'activité de la raison découvrant et contemple la vérité. C'est à dire que le bonheur est le but ultime de l'agir de l'homme.

Il existe donc une éthique de l'IA, c'est-à-dire une éthique en relation avec l'intelligence artificielle. C'est une éthique qui amène à réfléchir sur les enjeux, les valeurs et les finalités poursuivies par l'action humaine lorsque ces IA sont déployées. Des problèmes d'ordre éthiques se posent en effet lorsque des technologies à base d'intelligence artificielle sont créées. C'est le cas par exemple des biais qui existent dans les systèmes de TAL. Si les systèmes de TAL produisent des résultats souvent biaisés c'est parce que ces systèmes se sont entraînés sur des jeux de données qui renferment des biais et des préjugés. Ce problème vient du fait que les créateurs de ces systèmes ont dû faire des choix à un moment donné à la conception du système, entraîner le système sur un corpus non équilibré et stéréotypé. Un autre exemple est celui des voitures autonomes en contexte de dilemme moral. Les voitures autonomes sont dotées d'IA leur permettant de conduire seules sur les routes en analysant de manière autonomes et d'agir sans intervention du conducteur. Les hommes quand ils sont face à un dilemme moral mettent en jeu leur système de croyance. Ce n'est pas forcément le cas pour une machine qui doit faire un choix dans un dilemme et qui n'a pas de système de valeurs intégré.

3.1 L'expérience éthique en ligne *Moral Machine* du MIT

L'expérience en ligne du MIT sur les voitures autonomes[Awad E et al., 2018] le montre clairement. C'est une expérience permettant de recueillir une perspective humaine sur les décisions prises par les machines quand un dilemme moral se présentent à elles. Dans cette expérience les auteurs essaient de mesurer les attentes sociales à propos des principes éthiques qui devraient guider les décisions faites par les machines. Avant de décrire brièvement l'expérience il faut rappeler les lois de la robotique d'Asimov [Asimov et John W., 1942] qui sont les prémisses de la création des machines morales dans ses oeuvres fictionnelles et qui montrent à quel point il est difficile de créer des machines morales :

- Un robot ne peut porter atteinte à un être humain ni, restant passif, laisser cet être humain exposé au danger
- Un robot doit obéir aux ordres donnés par les êtres humains, sauf si de tels ordres entrent en contradiction avec la première loi
- Un robot doit protéger son existence dans la mesure où cette protection n'entre pas en contradiction avec la première ou la deuxième loi
- Loi zéro : un robot ne peut pas porter atteinte à l'humanité, ni par son inaction, permettre que l'humanité soit exposée au danger

Même si les voitures autonomes ne sont pas des robots dans le premier sens du terme, un robot est une machine programmée de même qu'une voiture autonome est une machine qui est amenée à prendre seule des décisions. Ces lois s'appliquent donc à ces machines. L'expérience est multilingue et met en place un un dilemme moral en recueillant des données à grande échelle sur la manière dont les citoyens de pays différents souhaiteraient que les véhicules autonomes résolvent les dilemmes moraux dans un contexte d'accident inévitable. Il y a eu 39,61 millions de décisions provenant de 233 pays. Il n'y a que deux résultats possibles soit le véhicule autonome dévie soit il reste sur sa trajectoire. Les individus cliquent ensuite sur le résultat qu'il jugent préférable. Il y a neuf facteurs : épargner les humains (par rapport aux animaux de compagnie), rester sur la trajectoire (par rapport à dévier), épargner les passagers (par rapport aux piétons), épargner plus

de vies (par rapport à moins de vies), épargner les hommes (par rapport aux femmes), épargner les jeunes (par rapport aux personnes âgées), épargner les piétons qui traversent légalement (par rapport à ceux qui traversent illégalement), épargner les aptes (par rapport aux moins aptes) et épargner ceux ayant un statut social plus élevé (par rapport à ceux ayant un statut social plus bas). L'utilité de cette expérience en ce qui concerne l'éthique des machines est que les résultats peuvent permettre de mettre en place une éthique universelle des machines, identifier des regroupements de pays avec des vecteurs homogènes de préférences morales. Comme résultats, on a une préférence à épargner les humains par rapport aux animaux, épargner plus de vies et épargner plus de jeunes par rapport aux personnes âgées. Aussi la règle éthique numéro 7 proposée par la commission allemande sur la conduite automatisée et connectée entre en accord avec les attentes sociales évaluées par le Moral Machine ou la conduite de la protection de la vie humaine devrait avoir la plus haute priorité sur la vie animale. La règle éthique allemande numéro 9 ne prend pas position clairement sur le fait de savoir si et quand les véhicules autonomes devraient être programmés pour sacrifier quelques-uns pour épargner beaucoup, mais laisse cette possibilité ouverte : il est donc important de savoir qu'il y aurait un accord public fort avec une telle programmation, même si elle n'est pas imposée par la réglementation. Cette règle stipule également que toute distinction basée sur des caractéristiques personnelles, telles que l'âge, devrait être interdite. Cela entre en conflit clairement avec la forte préférence pour épargner les jeunes.

4 De possibles solutions

Les décideurs et les fabricants doivent concevoir des systèmes d'IA en prêtant attention aux préférences morales des populations locales. Les valeurs culturelles, morales, éthiques doivent guider les processus de décision. La préférence pour épargner les individus jeunes plutôt que les individus âgés est beaucoup moins marquée dans les pays du groupe de l'est et beaucoup plus élevée dans ceux du groupe du sud. Il en va de même lorsqu'il s'agit d'épargner des personnes ayant un statut supérieur, une conception qui prévaut bien plus dans le groupe du sud. Par rapport aux deux autres groupes, les pays du groupe du sud montrent en outre une préférence beaucoup moins marquée pour l'épargne des humains que pour les animaux de compagnie. Seules la préférence (faible) pour l'épargne des piétons par rapport aux passagers et la préférence (modérée) donnée à l'épargne du légal par rapport à l'illégal semblent être réparties de la même manière dans tous les groupes (Awad et al, 2018). Enfin, quelques particularités marquantes peuvent être observées dans le groupe du sud, telles que la forte préférence pour l'épargne des femmes et la forte préférence pour l'épargne des personnes en bonne santé. Mais les prédicteurs culturels et économiques de préférence du Moral Machine (différences entre les cultures individualistes et collectivistes) peuvent être un frein à une éthique des machines universelle. L'inégalité économique économique à l'échelle d'un pays joue sur les choix moraux de chacun.

5 Les problèmes éthiques en Sciences affectives

Le domaine de l'*affective computing* est un domaine en informatique qui étudie et dans lequel on conçoit des technologies qui sont capables de reconnaître les émotions humaines. Cela va se voir typiquement avec des agents conversationnels émotionnels à l'écrit où à l'oral avec des robots qui sont capables de détecter et interpréter des émotions en analysant les indices émotionnels des personnes avec qui il interagit. Ces indices émotionnels comportent 3 dimensions prosodiques que sont le timbre, l'énergie et le rythme. Également d'autres phénomènes d'accentuation et d'intonation avec quatre dimensions sont analysés par les robots émotionnels ; il s'agit de la

fréquence fondamentale F0 et les formants, l'intensité, la durée (rythme et silences) et la qualité de la voix. Les 3 principales technologies sont la détection, l'interprétation des émotions et le raisonnement dialogique en utilisant des informations émotionnelles (Devillers L, 2021). Par raisonnement dialogique, on fait référence à la capacité à interpréter l'information dans le contexte d'un dialogue par la machine.

Le fait que les robots dotés d'intelligence artificielle puisse décoder nos émotions pose problème car il s'agit de capacités données à la machine qui sont intrusives dans la vie dans l'homme. Aussi, non seulement on donne les moyens à la machine de décoder nos émotions mais en plus on lui attribue parfois des traits humains. Le robot dialogueur va pouvoir simuler des émotions de l'humain comme le rire, la peur. Nous sommes donc dans le domaine de l'anthropomorphisme, c'est-à-dire attribuer des traits humains à des machines. Tout comme les voitures machines morales, les problèmes éthiques majeurs de ces technologies concernent les notions de responsabilité et d'autonomie. L'accent est surtout mise sur la responsabilité des machines. La notion d'agentivité est également liée à la responsabilité ; L'homme en tant qu'agent, est doté d'agir [Schlosser 2015]. Agir c'est instancier de bonnes relations causales entre les états et les évènements qui impliquent l'agent lui-même, c'est-à-dire l'homme. La responsabilité dans le sens juridique du terme se définit par l'obligation de réparer que l'on a causé. Par ces définitions, on comprend vite les difficultés d'application de ces notions à des systèmes d'intelligence artificielles de surcroît à des robots affectifs. En effet ces machines dotés d'une certaine intelligence émotionnelle quand elles maintiennent la canal de communication avec l'homme (back-channel) font du *nudge* c'est-à-dire utilisent des techniques subtiles pour influencer les individus. C'est de la manipulation est c'est problématique. Du fait que certains comportements de l'homme ne peuvent être directement interprétées par le robot, les créateurs de ces technologies manipulent donc les individus par le biais de ces robots qui utilisent des techniques douteuses. Certes il reste du travail à faire concernant les questions d'ordre éthiques pour ces technologies, mais des efforts sont quand même réalisés pour essayer d'encadrer l'intelligence artificielle.

6 Les lignes directrices pour l'IA bénéfique : la conférence d'Asolimar

La conférence du Futur of Life Institute organisée à Asolimar en 2017 à réunie plusieurs chercheurs de l'intelligence artificielle pour améliorer la coordination et la réflexion collective sur l'avenir de l'IA. Les principes de l'IA d'Asolimar sont des efforts en matière d'éthique et de gouvernance de l'IA. La liste des principes est établie sur leur site. Mais on peut citer ici deux principes en matière d'éthique de l'IA :

- Responsabilité : Les concepteurs et les constructeurs de systèmes d'IA avancés sont des parties prenantes dans les implications morales de leur utilisation, de leur usage et de leurs actions, avec la responsabilité et l'opportunité de façonner ces implications.
- Alignement des valeurs : les systèmes d'IA hautement autonomes devraient être conçus de manière à ce que leurs objectifs et leurs comportements puissent être assurés de s'aligner sur les valeurs humaines tout au long de leur fonctionnement (concordance ce qui est dit plus haut, le système de valeurs des communautés doit être intégré dans les IA).

Il faut aussi mentionner l'avis du Comité National Pilote d'Éthique du Numérique (CNPEN) qui est un rapport intéressant sur des questions éthiques dans le domaine de l'IA générative.

7 L'avis du CNPEN : la question de la manipulation et les solutions proposées

Cette partie du rapport met l'accent sur les risques de manipulation quand on utilise les systèmes d'intelligence artificielles qui utilise des techniques de manipulation qui échappent aux individus qui interagissent avec les systèmes d'IA génératives. Les systèmes d'IA génératives sont basés sur des modèles qui permettent de générer du texte par exemple. Comme dit plus haut dans la section 3, les biais ou les fausses informations sont produits par les systèmes d'IA génératives parce qu'ils sont entraînés sur des modèles de langues qui comportent justement des biais et de fausses informations. L'utilisation de ces technologies peut conduire manipuler politiquement les individus. Il faut donc considérer les enjeux éthiques dès la conception jusqu'au déploiement du système. Nous avons parlé de responsabilité, en effet les IA génératives sont parfois utilisés comme des outils d'aide à la décision ce qui entraînent la responsabilité de l'utilisateur. Il faut donc évaluer les biais connus des modèles en utilisant des jeux d'essais standardisés.

Les systèmes d'IA génératives peuvent s'apparenter à des systèmes de dialogue surtout celles qui génèrent du texte, un système qui dialogue avec l'humain. Il existe quelques paradigmes d'évaluation qui peuvent être mises de plusieurs manières, soit par l'application d'ensemble de tests avec des jeux d'essais comme évoqué dans l'avis du CNPEN soit une évaluation découlant de l'interaction de la machine avec l'utilisateur.

8 L'évaluation des systèmes de dialogue

Dans la revue TAL intitulée "Dialogue", Laurence Devillers et al, proposent quelques axes selon lesquels il est possible de classer les méthodes d'évaluation présentés brièvement :

- Les objectifs de l'évaluation dans lesquels on a le diagnostic va permettre d'identifier le fonctionnalités du système à améliorer, le progrès d'un système pour comparer différentes versions d'un même système et l'amener au niveau de performance souhaité
- La taille du contexte qui caractérise le contexte dialogique pris en compte dans l'évaluation
- Les objets mesurés, il s'agit de distinguer l'évaluation de la réponse de l'évaluation de la représentation interne que produit le système
- la nature de l'évaluation, c'est-à-dire une évaluation avec l'utilisateur ou à partir de corpus de tests

Il faut préciser que ces axes d'évaluation étaient proposés dans les années 2000, même si les technologies ont évoluées depuis avec l'apprentissage machine, fondamentalement les techniques d'évaluation des systèmes d'IA restent les mêmes.

Bibliographie

- Ménissier, Thierry. "Quelle éthique pour l'IA?." Naissance et développements de l'intelligence artificielle à Grenoble. 2019.
- Devillers, Laurence. "Les dimensions affectives et sociales dans les interactions humain-robot." Interfaces numériques 2.1 (2018): 105-118.
- Devillers, Laurence, and Roddy Cowie. "Ethical Considerations on Affective Computing: An Overview." Proceedings of the IEEE (2023).

- Devillers, Laurence, Hélène Maynard, and Patrick Paroubek. "Méthodologies d'évaluation des systèmes de dialogue parlé: réflexions et expériences autour de la compréhension." Traitement automatique des langues 43.2 (2002): 155-184.
- Kohler, Arnaud. "Relation entre IA symbolique et IA forte." (2020).
- Peroli, Enrico. "Le bien de l'autre. Le rôle de la Philia dans l'éthique d'Aristote." Revue d'éthique et de théologie morale 5 (2006): 9-46.
- Systèmes d'intelligence artificielle générative : enjeux d'éthique. Avis 7 du CNPEN. 30 juin 2023.
- Commission allemande éthique Conduite automatisée et connectée (lien)
- PILLIER, Christiane. Les dilemmes moraux. Entre-vues, 2003 Jun, - pp. 110 - 113.
- Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I. The Moral Machine experiment. Nature. 2018 Nov;563(7729):59-64. doi: 10.1038/s41586-018-0637-6. Epub 2018 Oct 24. PMID: 30356211.