Évaluer les défis en matière d'éthique de l'IA et de l'AM

En général, la vie privée, les préjugés et la discrimination font actuellement l'objet d'une grande attention. Cependant, il est fréquent qu'ils ne soient pas considérés comme prioritaires dans les mises en œuvre technologiques et qu'ils soient traités comme des problèmes isolés, ne recevant de l'attention que lorsque cela est nécessaire. De nombreuses organisations privilégient plutôt des objectifs tels que les gains d'efficacité ou l'augmentation des profits, qui nécessitent souvent des ensembles de données plus riches, mais ne tiennent pas compte de l'impact éventuel de leurs méthodes de traitement des données sur les questions fondamentales de justice sociale.[1]
Les conséquences de la mise en œuvre de technologies sans une compréhension complète des problèmes de confidentialité, de partialité et de discrimination éventuelle qu'elles posent menacent à la fois les individus et les entreprises. Les préjugés intégrés peuvent nuire à la capacité d'un individu à recevoir un traitement équitable dans la société. Pour les organisations, le potentiel négatif comprend l'atteinte à la réputation, l'impact financier, des litiges, le contrecoup réglementaire, les problèmes de confidentialité et une moindre confiance des clients et des employés. [2] Les développeurs d'applications technologiques devraient s'efforcer de les rendre impartiales, non biaisées et neutres, et les organisations devraient tenir compte de ces questions fondamentales lors de la mise en œuvre des technologies émergentes afin de s'assurer que les préjugés et la discrimination ne sont pas des éléments fondamentaux de la conception d'un système.

Comportement éthique

L'éthique est généralement définie comme un ensemble de normes permettant de déterminer quel comportement est considéré comme bon ou mauvais dans un groupe, une culture ou une société particulière, sur la base de normes acceptées. [3,4] Bien que certains comportements fassent souvent l'objet d'un consensus (mentir et tricher sont généralement considérés comme contraires à l'éthique), les opinions sur ce qui constitue un comportement éthique peuvent parfois diverger considérablement d'une culture à l'autre. Parmi les dilemmes éthiques liés à la technologie, citons l'utilisation de l'intelligence artificielle (IA) pour remplacer les humains dans l'exercice de certains rôles et l'utilisation de ces systèmes pour prendre des décisions automatisées au sein des organisations avec peu ou pas de surveillance, ce qui peut avoir des conséquences négatives pour la société.

Questions systémiques

En ce qui concerne la mise en œuvre de l'IA et de l'apprentissage machine (AM), les questions d'éthique ont atteint un point d'inflexion critique, obligeant les organisations à trouver un équilibre entre les objectifs opérationnels et les droits individuels.[5] La confiance est une composante de la garantie de la confiance dans la technologie ; il est essentiel de savoir qu'un système a pris une décision au bon moment et pour la bonne raison. Il en découle la nécessité fondamentale qu'un système soit explicable, de sorte qu'il soit facile d'expliquer pourquoi le système a pris une décision donnée et de maintenir un niveau élevé de confiance dans la conception. [6, 7]

Un article a proposé 10 lignes directrices pratiques pour l'application de l'IA à un large groupe de parties prenantes. Bien que l'accent ait été mis sur l'utilisation de l'IA dans des cas médicaux, les lignes directrices se prêtent à une application universelle. Elles visent notamment à garantir que les opérations technologiques peuvent être facilement expliquées, que les conceptions sont transparentes, que les décisions sont reconnaissables et reproductibles, et que les humains s'approprient ces décisions.[8]

Deux des dix lignes directrices traitent de questions fondamentales pertinentes :

JOSHUA SCARPINO | CISM, CISSP

Il est le directeur mondial de la sécurité et de la conformité chez Harver, où il dirige l'équipe mondiale de sécurité et de conformité. Il a plus de 18 ans d'expérience dans le domaine de l'informatique et de la sécurité et 16 ans d'expérience dans l'US Air Force. Il a supervisé les opérations de sécurité pour des entreprises du classement Fortune 500 et a renforcé les contrôles critiques d'organisations financières en déployant, établissant, développant et auditant leurs programmes de sécurité pour atteindre la conformité au niveau international. Tout au long de sa carrière, M. Scarpino a fait le lien entre plusieurs domaines de sécurité afin de résoudre les problèmes opérationnels, de gouvernance, de risque et de conformité. Passionné par l'éducation, il enseigne actuellement à temps partiel en tant qu'instructeur adjoint et mène des recherches pour son programme de doctorat sur les considérations éthiques relatives à l'intelligence artificielle et à l'apprentissage automatique à l'université Marymount (Arlington, VA, États-Unis).

1. « Une décision, une action ou une communication de l'IA ne doit pas violer une loi applicable et ne doit pas entraîner de préjudice pour l'homme. »
2. « Une décision, une action ou une communication de l'IA ne doit pas être discriminatoire. Cela s'applique notamment à l'entraînement des algorithmes. »[9]

Malgré les nombreux efforts déployés pour identifier ce qui est nécessaire pour se conformer, certains contributeurs fournissant même des cadres pour guider les déploiements, il n'existe toujours pas de méthode fiable pour identifier et aider à établir des priorités lorsqu'il existe un risque de préjudice, de discrimination ou d'autres problèmes éthiques. Et bien que les conséquences soient importantes pour l'IA et l'AM lorsqu'ils sont déployés à grande échelle, ce risque potentiel ne se limite pas à ces technologies.

> « Bien qu'il y ait une sensibilisation accrue, il n'y a pas encore d'approche unifiée pour identifier ces problèmes systémiques, et il n'y a pas de procédures normalisées pour les traiter de manière cohérente. »

ces dernières années, le droit à la vie privée des personnes a fait l'objet d'une attention accrue, et la sensibilisation à la justice sociale est devenue un point de discussion central dans de nombreuses régions. Dans le même temps, la nécessité d'adopter des normes universelles pour garantir une mise en œuvre éthique des technologies s'est accrue. À mesure que les capacités de l'IA et de l'AM évoluent, une approche normalisée est nécessaire pour déterminer si une organisation est soumise à un risque supplémentaire. Il est impératif que les dirigeants d'entreprise comprennent les conséquences de l'incapacité à atténuer le risque lié à l'absence de contrôles pour faire face aux problèmes de confidentialité, de partialité et de discrimination dans les technologies de l'IA et de l'AM. Bien qu'il y ait une prise de conscience accrue, il n'y a pas encore d'approche unifiée pour identifier ces problèmes systémiques, et il n'y a pas de procédures standardisées pour les traiter de manière cohérente. Il s'agit d'un problème fondamental pour de nombreuses organisations dans le monde. Les dirigeants doivent comprendre l'importance de ces questions et agir de manière appropriée pour éliminer les préjugés et la discrimination de toutes les mises en œuvre technologiques.[10]

Création de systèmes de confiance

Les exemples actuels de l'industrie et les avis d'experts peuvent être utilisés pour déterminer si les organisations ont intérêt à quantifier le risque potentiel de violation de la vie privée, de partialité et de discrimination présent dans leurs mises en œuvre technologiques. Un modèle peut être utilisé pour comprendre comment les organisations gèrent actuellement ce risque et mettre en évidence les avantages d'une approche unifiée qui met en œuvre des contrôles technologiques tout en sensibilisant aux préoccupations éthiques potentielles associées à ces questions fondamentales de justice sociale. Les avantages de l'élaboration d'un tel modèle et de l'aide apportée aux organisations pour mettre en évidence ces questions essentielles lors des phases de mise en œuvre et de conception des technologies sont évidents, comparés aux inconvénients de la mise en œuvre de technologies sans tenir compte de ces questions fondamentales. Il est essentiel d'examiner les exigences pertinentes pour un système donné et les approches actuelles qu'une organisation peut adopter pour comprendre ces domaines à risque avant la mise en œuvre, afin de s'assurer que les questions critiques concernant la partialité potentielle sont traitées de manière appropriée. Un examen préliminaire peut servir de point de départ à une conversation, aider à sensibiliser à ces problèmes et fournir une base pour les efforts de remédiation qui atténueront le risque identifié.[11]

Un examen de la littérature actuelle concernant la façon dont les organisations abordent actuellement la mise en œuvre en ce qui concerne les questions fondamentales révèle des préoccupations importantes.[12] Un exemple décrit une mise en œuvre de l'IA par Amazon qui a utilisé un algorithme de recrutement qui s'est avéré être biaisé. Le système privilégie des mots comme « saisi » ou « exécuté ». Ces mots se retrouvent plus souvent dans les CV masculins, ce qui conduit l'algorithme à favoriser les candidats masculins. Ce déploiement de l'IA a injustement limité la participation des femmes candidates à un emploi. Même si

Amazon a corrigé le problème, mais les personnes concernées n'ont probablement pas eu droit à une réparation du préjudice occasionné.[13] Cet exemple illustre l'une des nombreuses façons dont les technologies déployées peuvent contribuer à des problèmes fondamentaux. Permettre une prise de décision basée sur des préjugés intégrés ne cause pas seulement du tort aux personnes directement touchées, mais peut également accroître la méfiance envers ces systèmes. Un autre exemple est l'opinion répandue selon laquelle la discrimination est intégrée dans le système d'évaluation du modèle classique de la Fair Isaac Corporation (FICO).[14] Ses détracteurs affirment qu'il favorise les américains blancs par rapport aux personnes de couleur parce qu'il valorise davantage le crédit traditionnel que les antécédents de paiement positifs. Aracely Panameno, directrice des affaires latinos pour le Center for Responsible Lending, a fait remarquer que « si les données que vous fournissez sont basées sur une discrimination historique, alors vous cimentez fondamentalement la discrimination à l'autre bout. »[15]

« De nombreux algorithmes d'AM sont difficiles à expliquer et il est également difficile de déduire comment la réponse a été obtenue par le système ».

Bien que les organisations prétendent que les données sont impartiales, elles sont souvent incapables ou réticentes à fournir des preuves de leurs affirmations.[16] De nombreux algorithmes de l'AM sont difficiles à expliquer et il est difficile de déduire comment la réponse a été obtenue par le système ; ces systèmes sont connus sous le nom de « boîtes noires ». Les résultats sont basés sur des hypothèses concernant la manière dont les systèmes prennent leurs décisions.[17,18] Ces systèmes ont un impact considérable. Une recherche étudiant une grande partie des publications montre que de nombreux sujets qu'elles abordent ne sont que partiellement liés aux « systèmes explicables, responsables et intelligibles. »[19] Les seules catégories identifiées comme ayant un lien quelconque avec l'éthique et la vie privée sont « la confidentialité des big data, la confiance, l'équité algorithmique, et l'explication et le raisonnement. » Ces domaines d'intérêt ne représentent qu'une faible proportion de l'ensemble des recherches effectuées dans ce domaine.[20] La recherche sur l'éthique et la discrimination dans le domaine de l'IA fait l'objet d'un manque important d'attention. Lorsque l'IA et les systèmes connexes sont conçus et mis en œuvre, il est essentiel de comprendre comment ils peuvent contribuer à la prise de décisions susceptibles d'avoir des implications éthiques au sein d'une organisation.

La voie à suivre : Une approche cohérente

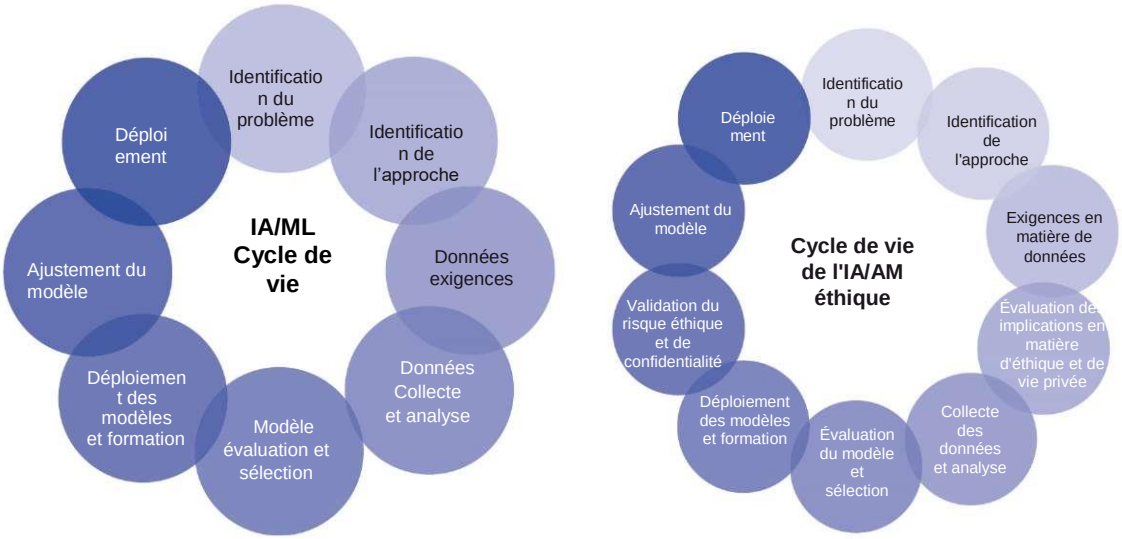
À mesure que la technologie évolue, il est primordial de prévenir les préjugés, de promouvoir la confidentialité des données et de protéger les droits humains inhérents. Dans les discussions sur ces questions fondamentales, il est essentiel de reconnaître que les êtres humains ont des préjugés intrinsèques et que tout ce qui est créé par l'homme risque de l'être. [21] Malgré les efforts déployés pour concevoir des systèmes d'IA et de l'AM exempts de biais ou les réduisant au minimum, ceux-ci peuvent néanmoins exister, que ce soit en raison des convictions personnelles des créateurs, de biais culturels ou d'une discrimination intégrée dans la manière dont les ensembles de données sont utilisés pour entraîner les systèmes. Les organisations doivent faire l'effort de comprendre fondamentalement comment les systèmes qu'elles conçoivent pourraient avoir un impact sur les utilisateurs.[22]

Dans le cadre de la recherche sur les cycles de développement de l'IA/AM, la plupart des cycles de développement partagés publiquement partagent ou ressemblent aux étapes décrites à gauche de la **figure 1.** Un élément distinct pour combattre les préjugés dans les systèmes fédéraux est la réalisation d'une évaluation d'impact des préjugés avant le développement et la mise en œuvre, la détermination de la pertinence de la solution et la validation itérative tout au long du processus de déploiement.[23] Des changements fondamentaux dans la façon dont ces systèmes sont créés, conçus et mis en œuvre sont nécessaires. L'ajout d'étapes distinctes pour traiter les préjugés au cours du cycle de vie du développement de l'IA/AM profiterait à toutes les organisations et favoriserait une approche cohérente pour identifier ces problèmes fondamentaux. L'ajout de deux étapes distinctes (l'une pour évaluer le potentiel de partialité et les implications éthiques qui en découlent, l'autre pour valider le risque après le déploiement) pourrait contribuer grandement à garantir la prise en compte des défis en matière d'éthique lors de ces mises en œuvre technologiques. Les ajouts proposés sont représentés à droite dans la **figure 1.**

L'évaluation des implications en matière d'éthique et de respect de la vie privée exige que les organisations prennent en compte et documentent ces préoccupations dans le cadre du processus de développement. Bien qu'il existe de nombreux cadres d'IA éthique pour guider le développement et l'évaluation de l'IA éthique, ces étapes distinctes n'ont pas été universellement adoptées dans le cycle de vie général du développement de l'IA, et elles constituent généralement un ajout au processus. Il est essentiel d'en faire un élément central du cycle de vie de toutes les mises en œuvre de l'IA et de l'AM pour une adoption et un succès universels. La validation du risque éthique et du risque lié au respect de la vie privée est nécessaire pour garantir que le

FIGURE 1

Cycle de vie de l'IA/AM typique/éthique



le risque considéré et évalué n'a pas changé une fois un système développé et déployé.

Le problème n'est pas que les outils et les cadres ne sont pas disponibles ou que les individus ne se préoccupent pas de ces questions. Une recherche en ligne sur les cadres de l'« IA éthique » révèle un nombre important de ressources et d'experts qui promeuvent des approches éthiques. Le défi actuel est que de nombreuses organisations restent concentrées sur le respect des objectifs commerciaux ou opérationnels et veillent à ce que les délais et les budgets des projets soient respectés. Il est probable que les préoccupations relatives à la partialité ne sont pas intentionnellement écartées, celles-ci ne sont simplement pas suffisamment prioritaires en raison des objectifs opérationnels.

Conclusion

Pour que cette conversation progresse, il convient de s'attaquer aux modèles de comportement actuels du secteur. Privilégier les objectifs opérationnels et les gains d'efficacité sans donner la priorité à l'éthique n'est plus une approche acceptable. Le climat est peut-être en train de changer, comme le révèlent les difficultés rencontrées par certaines grandes entreprises face aux préjugés conçus dans leurs systèmes et déployés à grande échelle. Certains ont réévalué et modifié leurs mises en œuvre technologiques pour lutter contre les préjugés intégrés et ont affiné leurs approches. Les concepteurs de systèmes doivent reconnaître la tendance à transmettre leurs propres préjugés aux systèmes. De plus, les entreprises doivent réaliser que lorsque les implémentations technologiques ont des préjugés intégrés, leur utilisation pour exploiter des ensembles de données peut perpétuer la discrimination à grande échelle. Ce n'est qu'en comprenant où ces problèmes fondamentaux se posent qu'il sera possible de développer des systèmes non discriminatoires et d'atténuer les risques.

Une voie prometteuse consiste à adopter et à ajouter des étapes distinctes au cycle de vie de l'IA et de l'AM, en veillant à ce que les préoccupations en matière d'éthique et de vie privée soient des éléments fondamentaux de tous les processus de conception et de développement de l'IA et de l'AM. En outre, il est essentiel de disposer d'un processus qui valide le risque résiduel potentiel après le développement mais avant la mise en œuvre pour garantir les résultats attendus. Pour une confiance durable dans la technologie et les organisations, il est essentiel de s'assurer que les systèmes émergents de l'IA et de l'AM respectent les droits individuels à participer de manière juste et équitable à la société.

Bibliographie

1 Lo Piano, S.; "Ethical Principles in Machine Learning and Artificial Intelligence: Cases From the Field and Possible Ways Forward", *Humanities and Social Sciences Communications,* vol. 7, iss. 1, 17 juin 2020, *https://www.nature.com/articles/ s41599-020-0501-9*

2 Cheatham, B.; K. Javanmardian; H. Samandari; "Confronting the Risks of Artificial Intelligence," *McKinsey Quarterly,* 26 avril 2019, *https://www.mckinsey.com/business-functions/ mckinsey-analytics/our-insights/confronting-the- risks-of-artificial-intelligence*

3 Singer, P.; "Ethics Philosophy," Britannica, 15 décembre 2021, *https://www.britannica.com/ topic/ethics-philosophy*

4 IGI Global, "What Is Ethics," *https://www.igi-global.com/dictionary/ethics-in-higher-education/10276*
5 *Op cit Lo Piano*
6 *Ibid.*
7 Muller, H.; M. Mayrhofer; E. Van Veen; A. Holzinger; "The Ten Commandments of Ethical Medical AI," *Computer,* juillet 2021, *https://ieeexplore.ieee.org/document/9473208*
8 *Ibid.*
9 *Ibid.*
10 Liu, X.; D. Murphy; "A Multi-Faceted Approach for Trustworthy AI in Cybersecurity," *Journal of Strategic Innovation and Sustainability,* vol. 15, iss. 6, 16 décembre 2020
11 Munoko, I.; H. L. Brown-Liburd; M. Vasarhelyi; "The Ethical Implications of Using Artificial Intelligence in Auditing," *Journal of Business Ethics,* vol. 167, iss. 2, 8 janvier 2020, *https://link.springer.com/article/10.1007/s10551-019-04407-1*
12 Trunk, A.; H. Birkel; E. Hartmann; "On the Current State of Combining Human and Artificial Intelligence for Strategic Organizational Decision Making," *Business Research,* 20 novembre 2020, *https://link.springer. com/article/10.1007/ s40685-020-00133-x*
13 Manyika, J.; J. Silberg; B. Presten; "What Do We Do About the Biases in AI?" *Harvard Business Review,* 25 October 2019, *https://hbr.org/20T9/T0/ what-do-we-do-about-the-biases-in-ai*
14 Consumer Financial Protection Bureau, "What Is a FICO Score?" 4 septembre 2020, *https://www.consumerfinance.gov/ask-cfpb/ what-is-a-fico-score-en-1883/*
15 Martinez, E.; L. Kirchner; "The Secret Bias Hidden in Mortgage-Approval Algorithms," The Markup, 25 août 2021, *https://themarkup.org/ denied/2021/08/25/the-secret-bias-hidden-in- mortgage-approval-algorithms*
16 *Op cit Manyika et al.*
17 *Op cit Liu*
18 *Op cit Lo Piano*
19 Abdul, A. ; J. Vermeulen ; D. Wang ; B. Lim ; "Trends and Trajectories for Explainable, Accountable and Intelligible Systems : An HCI Research Agenda," Proceedings of the 2018 Conference on Human Factors in Computing Systems, 21 avril 2018, *https://dl.acm.org/doi/10.1145/3173574.3174156*
20 *Ibid.*
21 Livingston, M.; "Preventing Racial Bias in Federal AI," *Journal of Science Policy and Governance,* vol. 16, iss. 2 mai 2020
22 Wallach, W.; "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making," *Ethics and Information Technology,* vol. 12, iss. 3 septembre 2010
23 Op cit Livingston