CHERCHEUR.E.S en responsabilité sociale et développement durable
ESG UQÀM

Séminaire théorique
L'éthique et l'IA

Présenté par Léa Goldman en collaboration avec Axelle Ferrant

Sous la supervision de

Mme Alice Friser

Mme Corinne Gendron

Mme Stéphanie Yates

Mme Marie-Luc Arpin

Les Cahiers du CRSDD • collection recherche

No 11-22

Table des matières

Business Ethics. https://doi.org/10.1007/s10551-021-05032-7 _____ *59*

Présentation du séminaire

Par Léa Goldman

Ce cahier de recherche présente un résumé critique d'articles consacrés aux thèmes de l'éthique et de l'intelligence artificielle.

Ces articles ont été sélectionnés pour ce séminaire dans l'objectif de mener une discussion autant sur les controverses que sur les fondements de l'éthique en IA.

« En novembre 2021, les 193 États membres de l'UNESCO ont adopté la Recommandation sur l'éthique de l'intelligence artificielle, le tout premier instrument normatif mondial sur le sujet. » (UNESCO, 2022) Cet accord vient démontrer l'intérêt grandissant de normaliser les pratiques afin d'assurer une gestion éthique de ce nouvel impératif.

Plus spécifiquement, ce séminaire aborde les enjeux de neutralité de ces algorithmes et la responsabilité des actions prises en fonction de ceux-ci. Le développement de ce type de technologie amène avec lui des risques qui sont à analyser et tenir en compte.

De plus, un aspect souvent déploré est le manque de réflexion en amont, et de questionnement sur le fondement même de cette technologie qui peut servir de moyen de légitimation. Ce qui mène, également, à se demander s'il est possible de mesurer la légitimé des décisions fondées sur ces dits algorithmes.

S.A. (2022) Éthique de l'intelligence artificielle. UNESCO.https://www.unesco.org/fr/artificial-intelligence/recommendation-ethics

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

Par Arnauld Chyngwa

Question

Au regard des principes et lignes directives publiés à travers le monde sur l'éthique de l'intelligence artificielle (IA), existe-t-il une convergence mondiale sur ces questions en pleine émergence ?

Réponse

Une analyse thématique faite par les auteurs, révèle qu'il existe des divergences de fond par rapport à quatre facteurs majeurs : (i) comment les principes éthiques sont interprétés, (ii) pourquoi ils sont jugés importants, (iii) à quel problème, domaine ou acteurs ils se rapportent, et (iv) comment ils doivent être mis en œuvre. Cependant, cette analyse montre l'émergence d'une apparente convergence entre les parties prenantes sur la promotion des principes éthiques de transparence, de justice, de non-malfaisance, de responsabilité et de confidentialité.

En effet, la communauté internationale semble converger sur l'importance de la transparence, de la non-malfaisance, de la responsabilité et de la confidentialité pour le développement et le déploiement d'une éthique de IA. Cependant, enrichir le discours éthique actuel sur l'IA grâce à une meilleure évaluation des principes éthiques critiques mais sous-représentés tels que la dignité humaine, la solidarité et la

durabilité est susceptible d'aboutir à un paysage éthique mieux articulé pour l'intelligence artificielle. De plus, déplacer l'accent de la formulation des principes vers la traduction dans la pratique doit être la prochaine étape. Un programme mondial pour une éthique de l'IA devrait équilibrer la nécessité d'une harmonisation transnationale et inter domaine par rapport au respect de la diversité culturelle et du pluralisme moral.

Argumentaire

Dans une analyse qui vise à cartographier le paysage mondial des principes et directives existants pour une éthique de l'IA, les auteurs ont identifié 84 documents en la matière, produits respectivement par des entreprises privées, les agences gouvernementales, les universités et institutions de recherche, les organisations intergouvernementales, les organisations à but non lucratif, des alliances, des syndicats et des partis politiques.

Les auteurs insistent sur le fait qu'au terme de la répartition géographique de ces produits documentaires, les données montrent une représentation significative des pays les plus développés économiquement, notamment les États-Unis (n=20 ; 23,8 %) et le Royaume-Uni (n=14 ; 16,7 %) qui représentent ensemble plus d'un tiers des produits documentaires des principes éthiques de l'IA. Ils sont suivis du Japon (n=4 ; 4,8 %), de l'Allemagne, de la France et de la Finlande (chacun n=3 ; 3,6 % chacun). Le Canada, l'Islande, la Norvège, les Emirats Arabes Unis, l'Inde, Singapour, la Corée du Sud, l'Australie sont représentés avec 1 document chacun.

A l’issu de l’analyse de ces données documentaires, les auteurs ont pu constater que onze valeurs et principes éthiques fondamentaux ont émergé. Ce sont, par fréquence du nombre de sources dans lesquelles elles figuraient : la transparence, la justice et l'équité, la non-malfaisance, la responsabilité, la vie privée, la bienfaisance, la liberté et l'autonomie, la confiance, la dignité, la durabilité et la solidarité.

La transparence : Présente dans 73 sources sur le 84 identifiées, ce principe de l’IA est celui le plus rependu dans la littérature actuelle. Elle apparait dans les sources comme un moyen pour minimiser les dommages et améliorer l’IA. Plusieurs sources documentaires suggèrent une divulgation accrue des informations par ceux qui développent et déploient l’IA, afin de parvenir à une plus grande transparence.

La justice et l'équité : l’analyse des auteurs a pu faire état de ce que pendant que certaines sources se concentrent sur la justice ne tant que respect de la diversité, l’inclusion et l’égalité, d’autres demandent la possibilité de faire appel ou de contester les décisions, ou le droit à la réparation. D’autres encore évoquent l’importance d’un accès équitable à l’IA. Par ailleurs, les sources émises par le secteur public, mettent un accent particulier sur l’impact de l’IA sur le marché du travail, et la nécessité d’aborder la question démocratique ou sociétale.

La non-malfaisance : Ce principe implique d'éviter des risques spécifiques ou des dommages potentiels, par exemple une utilisation abusive intentionnelle via la

cyberguerre et le piratage malveillant. Et il implique également de suggérer des stratégies de gestion des risques. A cet effet, les directives de prévention des préjudices se concentrent principalement sur les mesures techniques et les stratégies de gouvernance, allant des interventions au niveau de la recherche sur l'IA, aux approches latérales et continues. Ainsi, les stratégies de gouvernance proposées incluent une coopération active entre les disciplines et les parties prenantes, le respect de la législation existante ou nouvelle, et la nécessité d'établir des processus et des pratiques de surveillance.

La responsabilité : Ce principe malgré les nombreuses références dans les sources documentaires, il est rarement défini dans celles-ci. Néanmoins, tandis que certaines sources recommandent d'agir avec intégrité et clarifier l'attribution de la responsabilité et la responsabilité juridique, d'autres sources suggèrent de se concentrer sur les raisons sous-jacentes et les processus qui peuvent conduire à un préjudice potentiel. Par ailleurs, d'autres soulignent la responsabilité du dénonciateur en cas de préjudice potentiel.

La vie privée : L'éthique de l'IA considère la confidentialité à la fois comme une valeur à défendre et comme un droit à protéger. Bien que souvent indéfinie, la vie privée est parfois présentée en relation avec la protection et la sécurité des données. Quelques sources lient la vie privée à la liberté ou confiance.

La bienfaisance : Rarement défini aussi dans les différentes sources, quelques exceptions mentionnent l'augmentation des sens humains, la promotion du bien-être humain et de

l'épanouissement, la paix et le bonheur, la création d'opportunités socioéconomiques, et la prospérité économique. Ainsi selon les auteurs, les stratégies de la promotion de la bienfaisance incluent l'adéquation de l'IA avec les valeurs humaines, faisant progresser la compréhension scientifique du monde, et minimisant la concentration de puissance.

La liberté et l'autonomie : Ici, certaines sources la réfèrent spécifiquement à la liberté d'expression ou à l'autodétermination informationnelle. D'autres sources s'orientent plutôt vers la liberté, la responsabilisation ou l'autonomie. Par ailleurs l'autonomie est vue comme une liberté positive par certaines sources. Les auteurs pensent que la liberté et l'autonomie sont promues par la transparence et une IA prévisible, en ne réduisant pas les options et les connaissances des citoyens, et en augmentant activement les connaissances des gens sur l'IA.

La confiance : Parlant de la question de la confiance dans l'IA, les auteurs constatent qu'elle a été abordée par moins d'un tiers de toutes ces sources documentaires, ce qui soulèvent un dilemme éthique critique dans la gouvernance de l'IA. Alors que plusieurs sources, principalement celles produites par le secteur privé, soulignent l'importance de favoriser la confiance dans l'IA par le biais des activités éducative et de sensibilisation, un plus petit nombre de sources affirment que la confiance dans l'IA peut diminuer l'examen et saper certaines obligations sociétales des producteurs.

La durabilité : Sur le plan de la durabilité, elle est également sous représentée dans le discours éthique sur l'IA. Ce que les auteurs trouvent problématique du fait que le déploiement de l'IA nécessite des ressources informatiques massives, qui à leur tour nécessitent une consommation énergétique élevée. Cependant, ils précisent que l'impact de l'IA n'implique pas seulement les effets négatifs des infrastructures numériques sur l'environnement, mais aussi la possibilité d'exploiter l'IA au profit des écosystèmes.

La dignité : Bien que la dignité reste indéfinie dans les directives existantes, il s'agit bien d'une prérogative des humains mais pas des robots, il est fréquemment fait référence à ce que cela implique, notamment le fait d'être étroitement liée aux droits de l'homme ou autrement, signifier éviter le mal. Les auteurs pensent que la dignité n'est préservée que si elle est respectée par les développeurs d'IA en premier lieu.

La solidarité : La solidarité est principalement évoquée en relation avec les implications de l'IA pour le marché du travail. Des sources appellent à un filet de sécurité sociale solide. elles soulignent la nécessité de redistribuer les bénéfices de l'IA afin de ne pas menacer la cohésion sociale et en respectant les personnes et les groupes potentiellement vulnérables. Enfin, il y a une mise en garde contre la collecte de données et les pratiques centrées sur les individus qui peuvent porter atteinte à la solidarité en faveur d'un "individualisme radical".

Dans un tableau, les auteurs répartissent les onze principes éthiques en fonction de leur prévalence dans les ressources

documentaires analysées. Ils y indiquent pour chaque principe éthique, le nombre de documents dans lesquels il apparait sur les 84 documents. Aussi, ils indiquent les différents codes inclus pour chaque principe éthique.

Ce travail d'identification des principes éthiques a pu montrer qu'aucun principe ne semble être commun à l'ensemble du corpus de documents, bien qu'il y ait une convergence émergente autour des principes de transparence, de justice et équité, de non-malfaisance, de responsabilité et de vie privée; qui sont chacune référencées dans plus de la moitié de toutes les directives. Il y'a particulièrement une prévalence des appels à la transparence, la justice te à l'équité, ce qui indique une priorité morale émergente d'exiger une transparence du développement et de la conception des algorithmes, jusqu'aux pratiques d'utilisation transparentes. Aussi, la mise en garde des communautés mondiales contre le risque que l'IA puisse accroitre les inégalités.

Contribution & Utilité

Cet article contribue à faire le point sur les ressources documentaires existences et faisant une revue des lignes directrices et des principes éthiques de l'intelligence artificielle. Le travail de traitement des données et d'identification des principes éthiques convergents effectué par ces auteurs, peut servir de bonne base pour les futures recherches scientifiques en la matière. Aussi, ces travaux pourraient être un bon point de départ pour la communauté internationale dans la formulation et la formalisation au

niveau global, des lignes directrices et des principes éthiques de l'intelligence artificielle.

Critique

Je trouve que les auteurs font bien de montrer et de préciser dans cet article l'absence des ressources documentaires provenant de certaines régions du monde, notamment l'Afrique, l'Amérique du Sud et centrale, et l'Asie centrale. Au regard de ce fait qui pourrait surement s'expliquer par le niveau de pénétration numérique pour certains, il me semble que cela n'accorde pas à la convergence des principes indéfiées par les auteurs, un caractère exhaustif. Les résultats obtenus auraient pu être tout autre si ces régions du monde avaient des ressources documentaires produites. Ne pas avoir de documents formels signifie-il une absence éthique de l'intelligence artificielle? Je pense que la formulation et la formalisation des principes éthiques dans ces régions pourraient peut-être changer la classification faite par les auteurs, et par ricochet, les convergences identifiées.

Martin, K. 2022. Algorithmic Bias and Corporate Responsibility: How companies hide behind the false veil of the technological imperative. In The Ethics of Data and Analytics, Taylor & Francis.

Par ASSANI KIMWANGA BIN IBRAHIM AKIM

Questions

Les algorithmes sont-ils neutres ou dotés des biais chargés de valeur inscrits dans leur conception ? Les décisions de

conception des informaticiens et des scientifiques des données sont - elles neutres ? Par ailleurs, qui est responsable des bons et des mauvais résultats ou des mauvaises décisions lorsqu'une organisation ou un individu utilise un algorithme ?

Réponses

Juger de l'efficacité des algorithmes et prétendre qu'ils sont impénétrables produit un voile sur l'impératif technologique, qui protège les entreprises d'être tenues responsables des décisions chargées de valeur prises dans la conception, le développement et le déploiement des algorithmes.

Le développement des algorithmes devrait être examiné de manière critique pour élucider les biais chargés de valeurs encodés dans la conception et le développement

Reconnaître les biais chargés de valeurs de la technologie, y compris les algorithmes tout en préservant la capacité des humains à contrôler la conception, le développement et le déploiement de la technologie permet d'identifier et d'attribuer de manière appropriée la responsabilité de la performance des algorithmes et permet d'interroger mieux ceux qui conçoivent et développent des algorithmes pour augmenter nos décisions

Argumentaire

L'Auteure s'appuie sur un exemple récent d'entreprises utilisant l'IA : un pilote Amazon « a passé près de quatre ans à courir autour de Phoenix pour livrer des colis en tant que chauffeur contractuel pour Amazon.com Inc. Puis un jour, il

a reçu un e-mail automatisé. Les algorithmes qui le suivaient avaient décidé qu'il ne faisait pas son travail correctement. » Les conducteurs comprenaient que leurs performances étaient surveillées – comment ils conduisaient leur itinéraire, où ils mettaient des colis sur le porche, etc. – et recevaient parfois des courriels avec une note allant de fantastique à à risque. Cependant, « Amazon savait que déléguer du travail à des machines conduirait à des erreurs et à des titres dommageables », ont déclaré ces anciens gestionnaires, mais a décidé qu'il était moins coûteux de faire confiance aux algorithmes que de payer des gens pour enquêter sur des licenciements erronés tant que les pilotes pouvaient être remplacés facilement ».

Les arguments se répartissent traditionnellement en deux camps : ceux qui se concentrent sur l'algorithme en tant qu'acteur qui « fait » les choses et est en faute (déterministes technologiques) et ceux qui se concentrent sur les utilisateurs de cet algorithme comme déterminant le résultat (déterministes sociaux).

Contribution

L'article s'offre l'ambition de rendre plus explicites les implications pour la responsabilité des entreprises. L'auteure examine également les implications de la formulation d'arguments technologiques impératifs, en définissant les algorithmes comme évoluant sous leur propre inertie, en fournissant des décisions plus efficaces et plus précises et en dehors du domaine de l'interrogation. Elle soutient spécifiquement que juger l'IA sur l'efficacité et prétendre que les algorithmes sont impénétrables produit un voile sur

l'impératif technologique qui protège les entreprises d'être tenues responsables des décisions chargées de valeur prises dans la conception, le développement et le déploiement d'algorithmes. Elle fait remarquer qu'il est important de noter que les affirmations selon lesquelles les algorithmes sont impénétrables et efficaces fournissent un bouclier aux entreprises qui prennent des décisions chargées de valeur.

Enfin, Elle offre comment l'IA et les algorithmes devraient être interrogés compte tenu de ce que nous savons actuellement sur les décisions chargées de valeur des informaticiens et des scientifiques des données. Les biais sont des caractéristiques de conception chargées de valeurs dont l'utilisation a des implications morales.

Utilité

L'article présente une utilité indéniable pour les chercheurs du domaine des sciences techniques, ingénierie et mathématiques (STIM) intéressés aux questions de l'intelligence artificielle et pour les juristes intéressés aux questions de la responsabilité juridique. L'auteure explique l'IA et les biais chargés de valeurs selon les arguments déterministes (déterministes sociaux et déterminismes technologiques). Ces explications révèlent une fausse tension existante entre les déterministes sociaux et les déterministes technologiques, tension qui est à la fois un exercice académique et permet d'identifier et d'attribuer des responsabilités entre l'acteur de l'acte et celui qui doit en répondre de ses effets dommageables.

Critique

Dans son article, l'auteure propose un cadre pour examiner de manière critique le développement d'algorithmes afin de mettre en lumière le biais chargé de valeurs conçu dans le programme IA. Ce cadre prévoit des évaluations morales à 5 étapes à savoir, une évaluation morale des résultats, une évaluation morale des critères de fonctionnement des algorithmes, une évaluation morale des choix en matière de données ; une évaluation morale des choix en matière de données ; une évaluation morale des hypothèses dans l'élaboration du modèle et enfin, une évaluation des plans de résilience. A chaque étape d'évaluation, une série des questions est posée pour évaluer de manière critique un algorithme. Il en ressort de ces évaluations qu'un biais chargé de valeurs d'un algorithme peut répondre soit à une variable de résultat choisie avec des implications quant à ce que l'organisation utilisatrice juge important et dont les intérêts sont priorisés dans la conception de l'algorithme, soit que les critères de fonctionnement d'un algorithme sont choisis par l'entreprise qui le développe, soit en considération des données d'apprentissage et des données utilisées lors du déploiement d'un algorithme (l'informaticien et le scientifique des données décident quelles données sont appropriées à utiliser et fait le jugement de valeur non seulement quant à l'utilisation d'un ensemble de données, mais aussi quels facteurs dans l'ensemble de données à inclure et enfin soit que les développeurs sont obligés à s'attendre à ce que des erreurs se produisent et la conception d'algorithmes doit inclure la planification de la capacité à identifier, juger et corriger les erreurs.

Il s'agit d'un article intéressant qui met en lumière, les différents problèmes que suscite ou peut susciter l'IA, mieux les algorithmes dès sa conception, son déploiement jusqu'à son usage. Cependant, nous estimons que le voile sur l'impératif technologique qui pourrait protéger les entreprises de leur responsabilité dans les décisions chargées de valeur n'a pas été complètement levé. Le modèle d'évaluation en 5 phases proposé par l'auteure est un modèle complexe. Il émet beaucoup d'hypothèses d'analyse critique, laissant ainsi une véritable marge de manœuvre en faveur par exemple des développeurs. Ces derniers gardent la possibilité de jongler avec des explications techniques afin de renvoyer la faute ou les erreurs soit aux informaticiens, soit aux utilisateurs et ainsi se soustraire facilement de leur responsabilité au détriment des autres acteurs dans la chaine.

Par ailleurs, devant la panoplie des questions d'ordre technique soulevée par la conception, le développement et le déploiement d'algorithmes, Un juriste appelé à identifier et à attribuer la responsabilité en cas de contentieux entre acteurs sur les décisions chargées de valeur, se voit abandonner à la merci des experts pour l'éclairer sur une question de responsabilité dont le régime parait englober à la fois une responsabilité éthique et juridique.

Ainsi, cet article fait émerger des questions importantes de recherche en faveur des juristes intéressés à la problématique de la responsabilité du fait des choses inanimées.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics, 160,* 835–850. https://doi.org/10.1007/s10551-018-3921-3

Par Axelle Ferrant

Question

Les entreprises qui développent des algorithmes ont-elles une responsabilité envers l'utilisation qui est faite de leurs algorithmes, de quoi ces entreprises sont-elles responsables et quel est le fondement normatif de cette responsabilité ?

Réponse

Les entreprises qui conçoivent et développent des algorithmes sont responsables des implications éthiques de ces algorithmes qui sont utilisés dans la prise de décision (ex. embauche, promotion, octroi d'un prêt, etc.). En effet, une obligation est créée lorsque ces entreprises vendent volontairement ces algorithmes en connaissant le contexte de la prise de décision et les capacités uniques de l'algorithme. Par ailleurs, cette responsabilité se fonde sur le principe selon lequel une entreprise qui conçoit un algorithme influençant volontairement la prise de décision d'une autre entreprise devient responsable des décisions de cet algorithme.

Argumentaire

Les algorithmes déterminent qui sera embauché, qui recevra un prêt ou quels articles de presse seront proposés aux

consommateurs. En ce sens, l'autrice soutient que les algorithmes « structurent silencieusement nos vies ». Elle considère par ailleurs que les algorithmes sont chargés de valeurs et ne sont pas, contrairement à l'image que certains en ont, neutres. En effet, ils sont empreints des objectifs et des valeurs de ceux qui les créent.

Pour l'autrice, les entreprises qui conçoivent les algorithmes ont une responsabilité à plusieurs niveaux. D'une part, elles sont responsables de la délégation des rôles et responsabilités dans le cadre de la décision algorithmique. Les concepteurs choisissent les rôles des personnes et de l'algorithme dans la prise de décision. D'autre part, ces entreprises sont responsables de la décision prise par l'algorithme et des implications éthiques de l'algorithme utilisé, particulièrement lorsque l'algorithme est conçu pour empêcher les individus de prendre leurs responsabilités dans une décision. Ces entreprises sont ainsi responsables de produits qui fonctionnent conformément à la manière dont ils ont été conçus.

L'autrice base une partie de son argumentaire sur l'exemple d'un algorithme utilisé pour évaluer le risque présenté par un détenu pour permettre (ou non) sa sortie de prison. Elle montre comment le système informatique basé sur cet algorithme a produit des décisions non éthiques et est chargé de valeurs, notamment en catégorisant davantage les personnes racisées comme « à risque élevé » sans que ces décisions soient soutenues par les résultats empiriques ultérieurs. Elle met en lumière, notamment à travers l'apprentissage automatique (« machine learning »), la façon dont le recours à des données passées (biais historique) tend

à perpétuer des inégalités et des rapports de pouvoir existants.

Pour soutenir son propos, l'autrice fait appel aux théories des études des sciences et des techniques de Latour et Alkrich, notamment pour déconstruire l'illusion de la neutralité des algorithmes et de la déresponsabilisation de l'acteur humain par la délégation de la décision à une technologie. Elle remet ainsi en cause l'idée qu'en retirant les individus de la prise de décision, les décisions prises par des algorithmes seraient moins biaisées. Les algorithmes deviennent des acteurs non humains de la prise de décision auxquels des acteurs humains délèguent des tâches et des responsabilités. Elle démontre que déléguer la prise de décision ne retire pas les responsabilités liées à cette prise de décision.

Si l'autrice confirme que les entreprises sont responsables des décisions prises par les algorithmes, elle établit différents degrés de responsabilité. Ainsi, au plus l'algorithme est conçu de façon opaque et autonome, au plus le concepteur et l'algorithme portent la responsabilité des décisions prises par l'algorithme.

Finalement, l'autrice confirme le fondement normatif de la responsabilité des développeurs envers les implications éthiques des algorithmes. En effet, si l'algorithme conçu par une entreprise influence la prise de décision d'autres institutions, cette entreprise peut être tenue responsable des actes, des préjugés et de l'influence qui résultent de cet algorithme.

Contribution

L'article contribue à faire avancer la réflexion autour de la responsabilité des entreprises conceptrices d'algorithmes envers les implications éthiques découlant de l'utilisation de ces algorithmes. Elle contribue à mettre en lumière la manière dont l'utilisation de données peut mener à des décisions algorithmiques qui perpétuent des systèmes de pouvoir injustes.

Utilité

L'article est utile pour quiconque s'intéresse aux enjeux éthiques de la conception et de l'utilisation d'algorithmes dans le cadre de prises de décisions. Il permet aussi d'interroger de façon pertinente l'argument de neutralité ou d'objectivité parfois attaché aux décisions prises par des technologies.

Critique

Cet article suggère des réflexions intéressantes autour des implications éthiques de l'utilisation d'algorithmes dans la prise de décision. Dès le début de l'article, l'autrice propose de définir l'algorithme comme chargé de valeurs. Cette proposition sert ensuite de fil d'Ariane pertinent au reste de l'article.

Si l'autrice conclut que les concepteurs d'algorithmes doivent être tenus responsables des implications éthiques de ces derniers, elle ne remet toutefois pas en question l'utilisation même de tels algorithmes. Or, selon moi, rendre les concepteurs responsables ne permet pas de répondre au réel problème sous-jacent à savoir la reproduction de biais

sociétaux par l'entremise de ces algorithmes. Par exemple, si le biais raciste ou sexiste d'un algorithme correspond à la reproduction d'un biais existant dans la société, ne faut-il pas plutôt remettre en question l'utilisation même d'algorithmes dans la prise de décisions impactant directement les citoyens ?

Bender et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? https://dl.acm.org/doi/10.1145/3442188.3445922.

Par Élizabeth Durand

Cet article pose une question qui doit se trouver au cœur du développement de l'intelligence artificielle, mais qui reste néanmoins la tache aveugle du progrès technologique. Intéressés aux modèles de langage (LM), les auteurs tentent de déterminer si la communauté de scientifiques et d'experts s'interroge sur « les risques potentiels associés à leur développement et sur les stratégies permettant de mitiger leurs impacts » (traduction libre, Bender et al. 2021).

Le texte, très bien structuré, répond à cette question en six temps en abordant les risques, environnementaux, économiques, sociaux et structurels associés au développement des LM. À travers leur argumentaire, les auteurs constatent que malgré l'omniprésence des risques ceux-ci prennent peu d'importance dans l'élaboration des technologies de traitement automatique des langues. Soit par naïve omission, l'enthousiasme du progrès prenant alors le pas sur les réflexions éthiques, soit par claustration entre les chercheurs, l'objet d'étude et leurs parties prenantes. En

effet, le syndrome du chercheur dans sa tour d'ivoire est d'autant plus manifeste quand l'objet d'étude peut se réduire à une suite de 1 et de 0. Il devient alors nécessaire de mettre en place une culture de recherche et une structure, promouvant le décloisonnement du chercheur l'obligeant à considérer la chaîne d'évènement, en aval et en amont, qu'enclenche son expérimentation.

Avant de s'enfoncer plus amplement dans l'argumentaire soutenu par Bender et al. (2021), il convient de faire un bref tour d'horizon des concepts clé qui y seront abordés.

D'abord, qu'est-ce que le traitement automatique des langues ? Généralement abrévié NLP pour *natural language processing*, il s'agit d'un domaine de recherche multidisciplinaire regroupant la linguistique, l'informatique et l'intelligence artificielle. L'objectif de ces recherches est de créer des programmes informatiques capables de reconnaître, d'analyser et de comprendre une quantité importante de données linguistiques pour ensuite produire une réponse adéquate tant au niveau linguistique que contextuel. Ces programmes sont utilisés dans la reconnaissance automatique de la parole (SIRI), les logiciels de traduction simultanée, les programmes de compréhension de langage naturel (compréhension de textes écrits), etc. La recherche sur les NLP débuter alors de la Guerre froide étend tranquillement sa chape d'angoisse sur l'Amérique. Dès 1950, le test de Turing, visant à évaluer l'intelligence d'une machine, comprend une section sur la capacité de cet appareil à comprendre le langage naturel et à générer des phrases intelligibles. Est-ce que la machine peut flouer l'homme quant à son identité ?

Pour créer des programmes capables de comprendre et d'interpréter des textes à haut niveau, certains développeurs utilisent les modèles de langage. D'abord grandement utilisé dans l'élaboration de logiciel de traduction, le LM s'appuie sur des méthodes de calcul statistique pour déterminer la probabilité qu'un mot x soit suivi d'un mot y. Pour que de tels modèles soient pertinents et fiables, ils doivent s'appuyer sur un nombre imposant de données, mais quelles données prendre et où les choisir ? Ces deux questions sont au cœur des réflexions avancées par Bender et al. (2021). Au profit de la simplicité, et étant donné l'objectif du présent texte, nous diviserons les méthodes de collectes de données en deux familles. Bien entendu, tout expert de la question pourrait contredire ce découpage et apporter de brillantes nuances. Néanmoins, pour l'usage que nous ferons de cette distinction, je me permets de croire en sa pertinence.

La première méthode, un peu plus contrôlée, demande au chercheur de cibler les sources où s'abreuvera le LM. En identifiant par lui-même les sources d'information, le chercher à une meilleure emprise sur le contenu qui constituera la base des associations linguistiques que tissera son programme. Il devient alors plus imputable des résultats. La deuxième méthode consiste à faire du *web crawling*. Selon cette technique, les LM explorent automatiquement l'internet en quête de texte, comme des forums de discussions, et se nourri à même les publications « quotidiennes » des internautes. Cette méthode est peut-être moins fastidieuse, car l'implication du chercheur est moindre et elle permet au LM d'intégrer une quantité faramineuse de données.

Cette incursion très sommaire et vulgarisée dans le monde de la linguistique et de l'intelligence artificielle est maintenant complétée. Néanmoins, ce détour s'avère nécessaire pour se figurer avec précision les enjeux entourant la création de LM.

Commençons par aborder la face cachée de toute technologie, celle que nous préférons souvent ignorer pour nous bercer de l'illusion d'une technologie indubitablement salvatrice. Toute technologie, à plus haut fait les technologies informatiques, ont un important coût environnemental. On peut croire qu'à long terme ces investissements permettront d'améliorer (sauver) la vie sur terre, n'empêche, que leur développement est une source importante de pollution. Pensons simplement à l'extraction des minerais nécessaires à la fabrication d'ordinateurs, de téléphones ou de tous organes dérivés de l'IA. L'exploration du sous-sol géologique ne se fait pas sans dommages environnementaux : émission de CO2, vibration et érosion des sol, pollution de la nappe phréatique, déforestation, etc.

De plus, une part non négligeable des ressources premières sont prélevées dans des pays en développement. Ainsi, l'avènement de nouvelles technologies se fait souvent au détriment des populations les plus vulnérables et les moins susceptibles d'en profiter. Dans le cas des LM, Bender et al. (2021) soulève à juste titre, le manque de diversité des langues intégrées aux modèles linguistiques. En effet, dans une forte proportion, l'anglais reste la langue dominante utilisée par les LM et les NPL. Dans ce contexte, les peuples ayant payé le plus lourd tribut environnemental et social à la création d'une technologie n'ont même pas la consolation

d'en retirer un objet en adéquation avec leurs besoins. Prenons l'exemple de Google translate, accessible et suffisamment efficace, cet outil de traduction est grandement utilisé pour faciliter les échanges entre nations. Que ce soit pour traduire sommairement un texte ou jouer l'interprète lors d'une discussion. Or, la plupart des langues nationales africaines sont absentes du traducteur simultané. Toutefois, les langues des pays colonisateurs sont toutes présentes, l'obligation d'y recourir pour utiliser ces programmes peut être vécue par certains comme renforçant d'une manière pernicieuse l'ascendant colonial.

Le manque d'équité et de diversité des langues intégré dans les LM peut aussi engendrer des erreurs de compréhension et de traduction ayant de lourdes conséquences. Ce qui est d'autant plus vrai lorsque ces programmes deviennent omniprésents et utiliser pour surveiller la population. À ce titre, Bender et al. (2021) relève l'exemple d'un Palestinien arrêter par la police israélienne à la suite d'un message écrit sur Facebook. Sa publication d'origine écrite en arabe souhaitait « bon matin », ces mots anodins ont d'abord été traduits en anglais par « hurt them » résultat plus que médiocre, puis traduit à nouveau en hébreu ce qui a donné « attaquez-les ». Cet exemple illustre bien le danger que peut entraîner une base de données incomplète, mais aussi les risques de se baser sur une langue intermédiaire de traduction. Le jeu du téléphone fini rarement bien.

Le développement des LM comporte aussi des risques sociaux. Notamment, celle s'appuyant sur du web crawling. Selon Bender et al. (2021) voici les sources majeures auprès desquels se nourrissent les LM explorant le web au petit

bonheur la chance : Reddit, Twitter et Wikipédia. Cette liste est problématique à plus d'un égard. D'abord, ces plateformes sont investies d'une population particulière qui ne peut pas être considérée comme représentative. Par exemple, selon Pew Internet Research's (2016), aux États-Unis, 67 % des gens utilisant Reddit sont des hommes et 64 % sont âgés entre 18 et 29 ans. Quant à Wikipédia seulement 15 % des contributeurs sont des femmes. Ces statistiques abordent seulement la différence de genre et d'âge, mais il est raisonnable de penser que de telles différences peuvent aussi exister en matière de diversité ethnique, religieuse et sexuelle. Les LM puisant leur base de données à même ces plateformes se retrouvent donc dans une espèce de chambre à écho numérique, assimilant les opinions des mêmes groupes d'individus, créant et renforçant des associations linguistiques qui peuvent s'avérer préjudiciables pour les populations marginalisées. De plus, l'internet n'est pas un lieu où règne la politesse et la retenue. Les commentaires haineux abondent et ils seront inévitablement traités par les LM en web crawling. Bien sûr il est possible de coder les programmes de manière à ce qu'ils omettent tout contenu présentant des mots interdits. Toutefois, cette liste doit être préalablement dressée et elle ne peut pas être exhaustive ou exempte de biais. Pour ajouter en complexité, il existe aussi plusieurs façons de dénigrer plus subtilement. Par exemple, les commentaires associés aux récits de femmes racontant des agressions sexuelles sont fréquemment invalidants. Ils vont ridiculiser les femmes ou minimiser les gestes posés par les agresseurs en comparant ces dénonciations à des colères d'enfants (tantrum) ou en invoquant l'instabilité hormonale des femmes. Aucune de

ces associations ne peut être prévenue par une liste de mots interdits. Un LM qui ingère ces conversations dans sa base de données devient forcément biaisé et aura plus de chance de proposer des mots ou des traductions consolidant ces associations.

En outre, il est difficile d'inverser une tendance déjà assimilée. Comme les programmes de LM se nourrissent à même le web, pour renverser une tendance associative, ces derniers doivent consommer un nombre important d'articles et de textes qui défait cette liaison. Ce processus est problématique, pour deux raisons. D'abord, plusieurs enjeux n'ont pas la couverture médiatique nécessaire pour engendrer une multitude de publications. Ainsi les changements incrémentaux ont peu de visibilité et sont conservés dans les modèles de LM. Ce qui vient en contrepartie nuire à ce changement. Les bouleversements radicaux, comme BLM ou #metoo, sont alors surreprésenté, mais la couverture médiatique sensationnaliste et les commentaires haineux prolifère aussi. Deuxièmement, il peut arriver qu'aller à l'encontre des associations linguistiques demande une certaine révolte politique. Le soulèvement d'une minorité aurait tôt fait de renforcer l'association déjà en place simplement par le traitement médiatique négatif de la nouvelle. Bien entendu, l'assimilation de biais n'est pas uniquement réservée aux LM développés par web crawling. Les LM dont la base de données est structurée directement par les chercheurs sont aussi susceptibles de présenter des biais préjudiciables. Toutefois, ceux-ci sont attribuables aux chercheurs en question et demandent donc à ces derniers une certaine

introspection pour choisir minutieusement les bases d'informations à exposer aux LM.

Un autre risque que les auteurs associent aux LM, qui est selon moi fort intéressant est directement imputable aux utilisateurs. Le langage est une faculté complexe et les conclusions que l'on tire des discussions dépendent de plusieurs facteurs, entre autres le contexte de la discussion et la personne avec laquelle nous échangeons. La façon dont elle se présente et les caractéristiques que nous lui attribuons. Ainsi, l'énoncé « vive la liberté » a le potentiel de renvoyer à un ensemble de représentations différentes, le sens que nous lui donnerons et la portée de sa signification variera selon son contexte. Sommes-nous en train de fêter la patrie ou sommes-nous en train de la fustiger, l'allégresse ou la révolte ? Cette façon de se référer au contexte et d'attribuer à l'interlocuteur diverses intentions est presque un réflexe. Ce comportement est si ancré dans notre compréhension du langage que face aux LM et au NLP, l'humain a tendance à l'adopter inconsciemment. Or, les programmes de LM et NLP n'ont aucune intention et les phrases produites par SIRI ou les suggestions de réponses offertes par les claviers intelligents des téléphones ne sont rien d'autre qu'un résultat statistique, qu'un apprentissage théorique. Ces interactions ne reposent en aucun cas sur une discussion intentionnelle et consciente. Néanmoins, cet aspect est souvent oublié par les utilisateurs. Ces derniers projettent alors un contexte et une personnalité aux programmes qui vient conforter leurs points vus et amplifier le cas échant certains biais.

L'article de Bender et al. (2021) est très intéressant. En étalant intelligiblement leur argumentaire, les auteurs

plaident pour l'éveil d'une conscience technologique. Les technologies ne sont pas créées en vase clos et les coûts et bénéfices varient en fonction des lieux étudiés. Ces coûts ne sont pas anodins, qu'ils soient environnementaux, sociaux, politiques et même économiques. Jusqu'à présent, la question financière n'a pas été abordée dans ce compte rendu. Non pas, car elle n'est pas importante, mais plutôt, car elle est évidente. Ces modèles sont imposants et complexes et plus ils sont complexes et immenses, plus leur développement coûte cher en ressources humaines et matérielles. Il est donc essentiel de bien réfléchir et planifier le développement de ces modèles. Présentement, la mode est au « *the bigger, the better* ». Cependant les auteurs remettent en question cette position tant pour son coût financier que pour les risques mentionnés précédemment. Le mieux ne serait-il pas d'avoir des modèles adéquats et efficaces, adaptés à la tâche qui leur est demandée ? Pour ce faire, miser sur base de données monumentale n'est pas toujours la solution. Certaines tâches peuvent être accomplies avec un nombre inférieur de donnée bien choisie. En fait, dans tous les scénarios, le choix des données est un enjeu primordial. Une fois déployé, il devient difficile et coûteux d'apporter des modifications aux programmes. Il faut donc réfléchir et mapper l'architecture de données avec soin, impliquer les parties prenantes dans le développement et faire des prétests. Pour Bender et al. (2021), la préparation en amont et l'ouverture de ce champ disciplinaire à ses parties prenantes et le moyen sur lequel miser pour favoriser le développement éthique de ces technologies.

Ces technologies qui imitent les humains sans en être devraient apporter une plus-value à la société et non amplifier ses défauts. « Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy », (Birhane et Prabhu, 2021 dans Bender et al, 2021)

Hao, K. (2020). We read the paper that forced Timnit Gebru out of Google. Here's what it says, https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Par Charles Duprez

Question d'éthique en Intelligence artificielle : quand Timnit Gebru met le feu aux ailes d'Icare

L'humain pourrait créer des machines capables de l'égaler voir de le surpasser sur le plan de l'intelligence. Apogée d'un rêve prométhéen aux conceptions technicistes, l'IA serait dès lors capable de pallier tous les problèmes sociaux et environnementaux tout en nous libérant de nos tâches « *d'animal laborans* » comme le définit Hannah Arendt[1]. C'est donc, sans surprise, que l'IA est devenue l'objet d'un véritable engouement. Rien qu'à Montréal, le gouvernement

[1] Arendt, H. (1988). *Condition de l'homme moderne* (Agora (Paris, France) ; 24). Paris: Presses pocket.

a investi près de 100 millions de dollars pour pousser les recherches dans ce domaine. Depuis quelques années, les acteurs cet écosystèmes ont enrichi leurs rangs de chercheurs et chercheuses en sciences sociales, notamment en éthique, pour les aider à mieux prendre en compte les enjeux éthiques dans le développement de ces technologies. Soulignons aussi qu'en 2017 a été lancée la Déclaration de Montréal pour un développement responsable de l'intelligence artificielle qui vise à élaborer un cadre éthique de l'IA et à l'orienter pour qu'elle profite à toutes et à tous.

Derrière les discours officiels, des voix dissonantes s'élèvent tout de même pour souligner les enjeux éthiques qui accompagnent dès à présent le développement de ces technologies. En 2020, Timnit Gebru, chercheuse internationalement reconnue en éthique de l'IA au sein de Google a même été licencié pour, comme elle l'explique, avoir publié un article scientifique sur les IA du langage, au sein de l'entreprise américaine (« On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' »)[2]. Gebru est notamment l'autrice du fameux papier sur la reproduction du biais et la difficulté de reconnaissance faciale des femmes et des personnes de couleurs (les

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

machines n'ayant été qu'entraîné sur des bases de données composées majoritairement d'hommes blancs).

Les enjeux éthiques de l'IA dans les technologies de modèle du langage

Dans un article paru sur le journal MIT Technology Review[3], Karen Hao revient sur l'article de Gebru qui a lui a valu son licenciement et en extrapole les principaux enjeux.

Avant toute chose, les IA étudiées dans le cadre de l'article de Gebru sont des modèles de langage. Elles sont formées sur la base de grandes bases de données textuelles et apprennent à reconnaître le « sens » des mots pour pouvoir ensuite construire leurs propres phrases. Ces technologies sont devenues de plus en plus populaires ces dernières années et ont gagné en performance – on peut par exemple citer l'outil de traduction de Google. Mais, selon les auteurs, suffisamment de réflexions n'auraient pas été consacrées à leurs risques. Ils identifient trois principaux risques :

1. *Les coûts environnementaux et financiers*

Sur ce premier aspect, les données apportées dans l'article sont formelles, entraîner des IA demande une grande capacité de calcul qui consomme beaucoup d'électricité. Dans un article de 2019, des chercheurs montaient déjà que

[3] MIT. (Dec. 2020). *We read the paper that forced Timnit Gebru out of Google. Here's what it says.* Récupéré de https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

la consommation d'énergie et l'empreinte carbone des modèles d'IA du langage avaient considérablement augmentés depuis 2017[4], en lien avec l'utilisation de bases de données de plus en plus conséquentes.

Pour donner des ordres de grandeur, la formation d'une version du modèle linguistique de Google, BERT – qui sous-tend le moteur de recherche de l'entreprise – émet près de 650 kg de CO2 pour une consommation de 1507 kWh. C'est l'équivalent d'un aller-retour San Francisco - New-York. Un chiffre qui grimpe à près de 280 tonnes de CO2 dans le cas d'une formation via la méthode de « recherche d'architecture neuronale » (NAS) (656 347 KWh). Dans la pratique, les modèles doivent de plus être entraînés plusieurs fois avant d'être opérationnels. Rappelons aussi qu'à l'horizon 2050 chaque individu devrait émettre dans les 2 tonnes d'équivalents CO2/an pour répondre aux objectifs des accords de Paris. En contexte de dérèglement climatique, cette utilisation des ressources souligne encore les inégalités environnementales entre les pays riches (et plus spécifiquement, quelques entreprises riches) qui peuvent utiliser d'importantes ressources pour faire tourner leurs modèles, lorsque les pays pauvres sont les principales victimes des enjeux environnementaux. Comme l'expliquent Gebru et ses collègues: « *It is past time for*

[4] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

researchers to prioritize energy efficiency and cost to reduce negative environmental impact and inequitable access to resources ».

Ces arguments méritent d'être pensés dans une perspective critique plus large des nouvelles technologies numériques. En effet, l'IA comme toutes les technologies du numérique repose sur des infrastructures lourdes, fortement consommatrices de ressources et dont les pollutions, bien que souvent invisibles, sont en perpétuelle augmentation. La part des émissions de GES imputable au numérique représenterait ainsi en 2020 près de 4% des émissions mondiales[5]. Certes, comme le rappel Yann LeCun (codétenteur du prix Turing et directeur scientifique IA chez Meta), ce chiffre - bien qu'augmentant - reste aujourd'hui marginal à l'échelle globale[6]. Mais ce type de raisonnement traduit une incapacité à prendre en compte le caractère exponentiel de ces évolutions. Car avec une augmentation de la consommation énergétique numérique de 8,5% par année, l'énergie nécessaire pour alimenter les besoins en calculs devrait dépasser la production énergétique mondiale d'ici

---

[5] The Shift Project. (2018, octobre). *Pour une sobriété numérique* [étude]. Récupéré de https://theshiftproject.org/wp-content/uploads/2018/11/Rapport-final-v8-WEB.pdf

[6] France Culture. (2019, 25 octobre). *L'intelligence artificielle nous veut-elle du bien ?* [Vidéo en ligne]. Récupéré de https://youtu.be/GOExyRWQ9w8?t=2253

2040[7]. La pollution liée aux technologies numériques est donc bien réelle. S'ajoute à cela une autre pollution liée à l'extraction des métaux rares et l'épuisement de ces ressources. Ces matériaux sont des composantes essentielles des nouvelles technologies numériques sans lesquelles l'IA ne pourrait pas voir le jour. Comme l'explique Pitron : ces métaux rares sont en grande partie extraits en Chine dans des conditions désastreuses pour l'environnement[8].

Dans le même temps, le développement de l'IA s'accompagne souvent d'un discours « techno solutionniste » selon lequel cette technologie pourra éclairer nos prises de décisions grâce à un traitement sans précédent de la masse d'informations disponibles. Selon le « rapport Meadows », le retard dans les signaux est l'une des causes de notre manque de réactivité face à des dangers imminents[9].

---

[7] Villani, C. et al. (2018). *Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne* (p. 123). Conseil national du numérique. Récupéré de https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

[8] Pitron, G. (2018). *La guerre des métaux rares : La face cachée de la transition énergétique et numérique*. Éditions les Liens qui libèrent.

[9] Meadows, D., Meadows, D. et Randers, J. (2013). *Les limites à la croissance (dans un monde fini) : Le rapport Meadows, 30 ans après* (Collection Retrouvailles). Montréal: Les Éditions Écosociété.

L'IA serait en ce sens porteuse d'espoir par sa capacité à traiter en temps réel un grand nombre de données.

Il est aussi couramment avancé que l'IA pourrait permettre de réduire la consommation énergétique mondiale grâce à sa capacité d'optimisation. Marc Paquet, président de WikiNet estime par exemple qu'elle permettra de créer des outils performants pour trouver des solutions innovantes aux problèmes environnementaux[10]. Nous pourrions ainsi décontaminer des sites pollués ou encore améliorer l'efficience énergétique de ce qui nous entoure[11]. Dans un tel schéma, il reste cependant à prendre en compte l'effet rebond, car une hausse de l'efficacité énergétique n'entraîne pas une réduction proportionnelle de la consommation d'énergie[12]. Le rapport Villani stipule ainsi qu'« *une vision réellement ambitieuse de l'IA devra donc aller au-delà d'un simple discours sur l'optimisation de l'utilisation de nos ressources : elle devra intégrer un paradigme de croissance plus économe et collectif, inspiré de la compréhension de la dynamique des écosystèmes, dont elle sera un outil clé.* »[13]. Il faut de plus souligner qu'il s'agit de répondre avec des solutions techniques à des problèmes qui sont justement

---

[10] Paquet, M. (2018). Réhabilitation des terrains contaminés De nouvelles solutions grâce à l'intelligence artificielle. Vecteur Environnement, 51(1), 12.

[11] Voir par exemple la promesse de la domotique.

[12] Owen, D. (2013). *Vert paradoxe : Le piège des solutions écoénergétiques* (p. 89-109). Montréal: Éditions Écosociété.

[13] Villani, C. et al., *op. cit.*, p. 124.

posés par la technique. Cela contribue à renforcer l'imaginaire dans lequel l'amélioration de la technique est la seule marge de manœuvre dont dispose l'espèce humaine, écartant toute remise en cause de la croissance économique[14].

2. Des modèles restreints à une vision appauvrie du monde

Second enjeu soulevé dans l'article, celui du type de données qui abreuvent ces IA. Pour les rendre plus performantes, elles travaillent avec une immense quantité de données issues d'Internet, mais une telle quantité rien impossible toute vérification humaine du contenu qui sert à développer ces IA. Le risque est alors de reproduire des langages racistes, sexistes ou plus largement discriminants tels qu'ils apparaissent sur la toile. Il peut alors être à craindre qu'elles renforcent les discriminations de certaines minorités si les données qui les nourrissent vont dans ce sens. À ce titre, il faut rappeler l'expérience malheureuse de Tay, l'IA développée sur Twitter par Microsoft qui dû être retirée en quelques heures pour faute d'être devenue antisémite et négationniste[15].

---

[14] Jackson, T., *op. cit.*

[15] Tual, M. (2016, 24 mars). A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter. *Le Monde*. Récupéré de

La difficulté réside aussi dans le fait que les IA ont du mal à comprendre les nuances du langage. En fait, elles sont très fortes pour reproduire, mais non pour véritablement comprendre le sens des mots, ce qui les rend incapables de saisir certaines utilisations du vocabulaire, comme ce fut le cas avec la création d'un vocabulaire non sexiste et antiraciste dans les mouvements MeToo et Black Lives Matter. De même, en projetant une forme de compréhension du langage, les IA imposent un cadre homogène (issu de l'élite du monde de la Tech américaine) au reste du monde.

La recherche de Gebru fait écho à d'autres études qui témoignent d'un réel manque de diversité des personnes impliquées dans le développement des IA. Il apparaît par exemple que seuls 12% des chercheurs en *machine learning* seraient des femmes[16]. Des biais dans les algorithmes arrivent alors pour plusieurs raisons comme une vision du monde propre aux ingénieurs qui travaillent sur les IA, mais également des données utilisées pour les entraîner qui sont

https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html

[16] Chin, C. (2018, 17 août). AI Is the Future—But Where Are the Women? *Wired*. Récupéré de https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/

elles-mêmes issues d'un monde profondément inégalitaire[17]. Les algorithmes risquent donc de reproduire ces biais, aggravant ainsi les problèmes de justice sociale. Plus généralement, le problème est que les IA se développent sur la base d'une fausse représentation de la réalité, tirée de la vision du monde d'une minorité privilégiée. Comme le souligne Jean-Claude Ravet, l'IA risque de s'aligner aux intérêts des *oligarchies technofinancières de la Silicon Valley* et des marchés boursiers[18].

Pour mieux comprendre cet enjeu, on peut le mettre en lien avec les critiques autour du « cyber-colonialisme » qui représente une menace d'asservissement via l'IA des peuples au profit de quelques grandes puissances qui en aurait le contrôle. Cette idée développée entre autres par Nicolas Miailhe est également pointée du doigt par le Conseil national du numérique français qui voit dans le capitalisme de plateforme une captation de toute la valeur ajoutée[19]. La valeur; c'est celle des cerveaux disponibles qui sont recrutés par les grands groupes du numérique, mais également celle des données recueillies sur les utilisateurs.

---

[17] Chabal, A. (2019, 12 août). Quand l'intelligence artificielle reproduit les biais sexistes. *World Economic Forum*. Récupéré de https://fr.weforum.org/agenda/2019/08/quand-lintelligence-artificielle-reproduits-les-biais-sexistes/

[18] Ravet, J. (2018). *L'éthique et l'intelligence artificielle*. Relations, (795), 5.

[19] Miailhe, N., *op. cit.*

Pour Cédric Villani, nous assistons véritablement à une démarche de type colonial : « *vous exploitez une ressource locale en mettant en place un système qui attire la valeur ajoutée vers votre économie. Cela s'appelle une "cybercolonisation"*[20]. De plus, le pouvoir numérique est centralisé entre une poignée d'acteurs. Il s'agit des GAFAMI[21] côté américain et des BHATX[22] côté chinois dont le modèle d'affaires repose en grande partie sur un accaparement du « temps de cerveau disponible »[23]. Dans ces conditions, il est difficile d'imaginer comment l'IA pourrait aider à réduire les inégalités puisqu'elle se fonde sur un principe de récupération de la richesse et d'asservissement par la consommation d'une partie de la population, notamment des pays du tiers-monde. Un

[20] Belot, L. (2018, 17 juin). Intelligence artificielle en Afrique : « Le risque de captation de valeur existe », décrypte Cédric Villani. *Le Monde*. Récupéré de https://www.lemonde.fr/afrique/article/2018/06/17/intelligence-artificielle-en-afrique-le-risque-de-captation-de-valeur-existe-decrypte-cedric-villani_5316644_3212.html

[21] Google, Amazon, Facebook, Apple, Microsoft, IBM.

[22] Baidu, Huawei, Alibaba, Tencent, Xiaomi.

[23] Ministère des Armées. (2016, août). Lettre n° 53. Dans *Direction générale des relations internationales et de la stratégie, Observatoire du monde cybernétique* [document pdf]. Récupéré de http://www.defense.gouv.fr/content/download/483341/7740229/version/1/file/OBS_Monde+cybern%C3%A9tique_201608.pdf

mécanisme déjà bien connu qui a fait la force du capitalisme[24].

De plus, comme le démontre Thomas Piketty dans ses ouvrages[25], l'histoire de la répartition des richesses est toujours politique. L'auteur rappelle que l'absence d'investissements dans la formation est l'un des principaux facteurs qui creusent les inégalités. En appliquant son analyse à l'IA, il apparaît que cette technologie sera inégalitaire s'il n'y a pas de volonté politique de former les gens à la comprendre et à en profiter. Or, ce sont bien quelques entreprises dans les pays les plus développés qui ont les moyens d'investir massivement dans l'IA. Cela aura certainement pour conséquence de renforcer encore davantage leur capital, aggravant d'autant plus les inégalités avec les pays plus pauvres.

3. Des coûts d'opportunités

Troisième grande critique de Gebru et de ses collègues, celui des « efforts de recherche mal orientés », c'est-à-dire que l'argent investi dans ces technologies pourrait être mieux

[24] Partant, F. (2007) *Ce tiers monde si nécessaire*. Dans Essai sur l'après développement (p. 63-78). Paris : La Découverte.

[25] Voir par exemple : Piketty, T. (2013). *Le capital au XXIe siècle*. Paris: Éditions du Seuil.

employé. En effet, la plupart des chercheurs estiment que l'IA ne peut pas réellement comprendre le langage, elle se contente de le manipuler. Or, puisque les perspectives économiques sont immenses, les Big Tech continuent de financer abondamment cette voie, au détriment d'autres approches.

De plus, un raffinement de ces techniques laisse craindre que les modèles d'IA puissent être utilisés pour générer massivement de la désinformation. Ce qui pose encore la question de la concentration des pouvoirs et de la non-neutralité de la technique dans nos sociétés.

À ces enjeux nous pourrions rajouter la critique du système technicien. En effet, avec l'IA, la thèse de Jacques Ellul est plus que jamais d'actualité, car la technique - qui obéit à sa propre détermination - serait concrètement capable de se réaliser d'elle-même. L'IA serait alors le paroxysme du *système technicien*[26]. De plus, faute de pouvoir parvenir à rendre les machines humaines, les ingénieurs tentent de moduler l'humain en machine. Les *emoji* et autres architectures d'interactions développées par exemple sur les réseaux sociaux peuvent être vus comme une façon d'isoler, de décortiquer et de quantifier jusqu'à nos émotions. Miguel Benasayag met notamment en garde contre l'aspect de colonisation de l'IA qui s'infiltre cette fois-ci dans tous les pans de notre vie. Il estime que l'IA tend à réduire

[26] Janicaud, D. (1985). Y a-t-il un système technicien ? [Chapitre de livre]. Dans *La puissance du rationnel* (p. 143-149). Paris : Gallimard.

l'intelligence humaine à une machine à calculer au service de la technologie[27]. Pour qu'elle serve la liberté, l'IA devrait donc être au service du vivant. Cependant, une bonne façon de le faire serait déjà d'arrêter de la comparer à l'intelligence humaine, car cela la restreint à une logique purement computationnelle qui réduit l'humanité et la vie à la technique[28]. L'IA risque aussi de nous déresponsabiliser, car il y aurait un décalage encore plus prononcé entre nos perceptions, notre imagination et nos émotions. Comme le souligne Anders, avec la technique, nous perdons la capacité de nous représenter les effets de notre agir comme étant les nôtres[29]. Pour en comprendre toute la portée, imaginons qu'avec l'IA le soldat du futur « *pourrait être le berger d'un troupeau de robots* »[30], de quoi rendre le « monstrueux » tout à fait possible.

Ces critiques sont d'autant plus nécessaires que la réaction de Google de se séparer de l'autrice du fameux article en dit

---

[27] Godefridi, T. (2019, 15 octobre). La Tyrannie des algorithmes, de Miguel Benasayag. *Contrepoints*. Récupéré de https://www.contrepoints.org/2019/10/15/355749-la-tyrannie-des-algorithmes-de-miguel-benasayag

[28] Ravet, J., *op. cit.*

[29] Anders, G. (2003). Nous fils d'Eichmann [Extraits]. Dans *Nous fils d'Eichmann* (p. 51-59 et p. 89-104). Rivages poche.

[30] Noël, J., *op. cit.*

long sur l'importance d'une liberté académique et d'expression autour de ces débats de société.

En guise de conclusion : de l'importance d'élargir le débat sur l'IA

L'IA pourrait finir par s'imposer dans nos vies sous la forme d'un monopole radical nous supprimant la possibilité de vivre sans[31]. En cas de « crise », l'IA nous fait alors perdre en grande partie notre capacité de résilience. Et cela d'autant que les humains vont reléguer le pouvoir de décision à la machine qui sera supposément capable de faire de meilleurs choix. Le philosophe Éric Sadin ajoute qu'avec l'IA l'humain érige une « *nouvelle instance à nous dire la vérité en toute chose* ». L'exactitude d'une situation est perdue au profit d'une vérité décrétée par la machine qui, par sa puissance d'interprétation, va remplacer notre capacité de jugement[32]. Mais l'IA est un système d'expertise et en aucun cas un oracle de la vérité universelle. Or, être libre, c'est être autonome ce qui, comme le rappel Castoriadis, revient à être

[31] Illich, I. (2003). Les deux dimensions de la contre-productivité institutionnelle [Chapitre de livre]. Dans *Œuvres complètes (p. 659-676)*. Volume 1. Paris : Fayard.

[32] Thinkerview. (2018, 8 novembre). Éric Sadin : l'asservissement par l'Intelligence Artificielle ? [Vidéo en ligne]. Récupéré de https://www.youtube.com/watch?v=VzeOnBRzDik

capable de se donner ses propres lois[33]. Or que reste-t-il de notre autonomie si c'est l'IA qui par son analyse et ses choix façonne notre propre réalité ?

L'IA, très certainement, sera une technologie qui transformera nos réalités, et ses nombreuses promesses rendent d'autant plus difficile sa remise en question. Mais, comme le conçoit Miguel Benasayag, ce monde est de toute façon déjà là et la clé est peut-être de sortir d'une posture manichéenne vis-à-vis de l'IA et de chercher comment « *favoriser la colonisation de la technologie dans l'intérêt de la vie et de la culture* »[34] et ainsi remettre la technique au service de l'humain. Enjeu au cœur des réflexions des groupes de travail sur l'éthique de l'IA, dont fait partie Gebru.

[33] Castoriadis, C. (1998). L'individu privatisé. *Le Monde Diplomatique*. Récupéré de https://www.monde-diplomatique.fr/1998/02/CASTORIADIS/3528

[34] Goulet, M.-C. (2019, 26 septembre). L'intelligence artificielle : entre promesses et périls. *Nouveaux Cahiers du socialisme*. Récupéré de https://www.cahiersdusocialisme.org/lintelligence-artificielle-entre-promesses-et-perils/

Martin, É. (2021). L'éthique de l'intelligence artificielle, ou la misère de la philosophie 2.0 à l'ère de la quatrième révolution industrielle. Cahiers Société, (3), 189–218. https://doi.org/10.7202/1090182ar

Par Geneviève Dugré

Introduction

L'article de Martin est en quelque sorte une charge contre la philosophie kantienne, rawlsienne et utilitariste qui serait prégnante en éthique de l'intelligence artificielle. Il déplore également le fait qu'il s'agit d'une éthique des a posteriori, c'est-à-dire qu'elle ne réfléchit que lorsque la technologie est là et pour en limiter superficiellement les aspects les plus négatifs (ex. voiture autonome, robots tueurs) sans se questionner fondamentalement sur les technologies elles-mêmes, leur finalité et leur ancrage macrosociologique.

Ces éthiques, loin de vouloir transformer les régimes d'accumulation, de reproduction capitaliste et l'aliénation, servent en quelque sorte de forces légitimatrices, comme des manières d'acceptabilité sociale « pour légitimer des transformations socioéconomiques et politiques » (p.189). Elles contribuent de fait à évacuer et à rejeter l'apport des différentes théories critiques.

Les approches

Déterminisme et substantialisme

L'auteur est également critique de certaines de ces théories critiques. Il rappelle, en citant Feenberg que la technique est un objet d'intérêt moderne de la philosophie et que longtemps, la position dominante était celle de l'instrumentalisme et de la conception de la technique « comme moyen neutre » (p.91) aspect remis en compte par les déterministes et les substantialistes. Les premiers ont comme limite de subordonner la politique à la technique ou bien de se limiter à reconnaître que la technique elle-même est politique et insiste, de façon « romantique » (il réfère ici à Heidegger et Ellul) sur une position extrêmement pessimiste, mais, pourrait-on dire, sans réels débouchés. Les substantialistes insistent davantage sur la caractérisation de la technique dans la modernité, c'est-à-dire les travers d'un surplus de « rationalité, d'efficacité, de contrôle et de calcul » (p.191) et sur la domination que cela engendre.

Théorie critique

Les théories critiques, celles de l'École de Francfort et de Foucault, prôneraient une approche moins essentialiste et s'intéresseraient aux origines sociales de la domination, c'est-à-dire à l'importance de prendre l'organisation sociale, les institutions et les systèmes sociaux pour comprendre la technique, considérée ici comme « idéologie matérialisée » (p.193).

Constructivisme

S'inspirant toujours de la critique de Feenberg, il soulève les limites du constructivisme qui adhère aux niveaux dispositifs et se concentre sur des aspects très particularistes,

sans s'intéresser aux résistances sociales et aux groupes macro (classe, culture).

Les accusations

L'éthique des universitaires et des experts

Martin (2021; p.191) y va d'une charge contre « l'éthique produite par une recherche universitaire empirique » ou produite dans des « consultations publiques lorsque ces dernières sont dépouillées d'une prise en compte du « contexte politique général », de la résistance, de la macrosociologie, des luttes et des crises. La critique s'arrête selon lui à la prise en compte des thèses postmodernes et culturalistes. L'éthique ainsi développée servirait essentiellement à rendre acceptable l'intelligence artificielle et plus fortement, la pensée calculante qui la caractérise. Il appelle ainsi à une justice sociale plus englobante.

L'auteur a une tendance à personnaliser les attaques. Il consacre quelques pages à déconstruire les positions de MacClure, Bengio et Dilhac, qu'il associe aux éthiques précédemment mentionnées. En plus de promouvoir l'éthique a posteriori, ces auteurs se limiteraient à ne réfléchir qu'à l'atténuation des risques; à l'opacité des systèmes; aux cyberattaques; à la vie privée, aux biais discriminatoires et autres questions de ce type. Ils agiraient dans la promotion d'une IA responsable et ne serait essentiellement qu'une éthique « des concepteurs » visant à présenter Montréal comme une ville d'avant-garde dont la *Déclaration de Montréa*l en est l'aboutissement.

Ainsi, si la société civile était invitée à participer à son élaboration, le processus découlerait surtout d'un « travail étroit avec le milieu des affaires ».

Il décrit aussi la boussole écologique, issue du Pacte de la transition de Dominic Champagne, essentiellement comme un outil marketing pour assurer une légitimation auprès d'un plus large public, notamment sensible à des enjeux qui ne vont pas forcément de pair avec le développement technologique.

Il écorche aussi la charte de l'UNESCO, essentiellement élaborée par des experts; produite en très peu de temps pour un outil qui se veut ‘’un instrument normatif mondial’’ et qui se limite essentiellement aux biais discriminatoires (p.201)

Consultations et acceptabilité, éthique libérale, marché et marketing

L'auteur affirme que les problématiques éthiques sont le fruit de recensions effectuées lors des consultations publiques et que le rôle des éthiciens se limite souvent à produire un inventaire autour de questions limitées (vie privée, discrimination, développement soutenable) (p.202).

Il questionne la pertinence de ces questions dans un monde où la protection de la vie privée est déjà passablement mise à mal. Il questionne quelle est la vision du travail, de la société juste et de l'équitabilité qui est véhiculée lorsque l'on transfère à l'IA (ex. camion autonome) le travail; des questionnements qui seraient, à son avis, peu abordés par ces « experts universitaires » notamment ceux qui jouissent d'une reconnaissance médiatique (p.203).

Cela aurait pour conséquence de ne pas questionner le « mode de production économique », l'évolution des rapports de production et servirait surtout à neutraliser « rhétoriquement d'éventuelles résistances ou contestations » (p.203).

Il critique le rôle des universités, et notamment des départements de philosophie, qui se restreindraient trop souvent à « produire le cadre normatif » basé sur « les droits de l'homme et le respect de la vie privée » permettant aux concepteurs d'IA de prospérer.

Inspirations

Afin de réfléchir à des alternatives à ces modèles, il invite à considérer les travaux de Castoriadis et de Freitag.

Castoriadis

Castoriadis déplore en le fait que l'éthique vient compenser notre incapacité à répondre de la crise du politique, de l'abandon d'une vision globale, de la privatisation de l'existence, de l'enfermement dans la sécurité et les jouissances privées et se fait ainsi essentiellement individuelle, générale et abstraite (l'éthique).

Depuis le nucléaire et la crise de la technoscience, il y aurait eu une démultiplication de chaires universitaires en éthique se positionnant comme réponse pragmatique à la crise des valeurs.

Cependant, l'éthique ne s'avère pas capable de remettre en question les « régimes, constitutions, lois et institutions » (p.

204), à parfaire les institutions, à orienter globalement la société tout en tenant compte des particularités. À cet égard, l'éthique aurait quelque chose de fataliste et ne permettrait pas de sortir du cadre libéral, dépolitisé et post-politique. Elle serait de courte vue. Martin (2021; p. 205) donne ici l'exemple du développement de robots non sexistes ou non racistes, ce qui ne pose pas de questions d'ordre plus macrosociologiques.

Les chartes deviennent donc ici un « catalogue de règles mécaniques objectivées, de vertus pieuses, de commandements ou principes abstraits ne pouvant pas remplacer le véritable jugement et agir politiques, lesquels ne se situent pas dans quelque moment universel abstrait, mais se déroulent toujours, comme le dirait Aristote, dans un moment particulier et concret qui implique la phronesis, la prudence ou sagacité » (Martine, 2021; p.2016)

Freitag

Pour Freitag, selon Martin, le mode de régulation de la pratique sociale est celui du mode de reproduction décisionnel-opérationnel, systématique, cybernétique et post-moderne, qui se substitue aux ordres précédents : culturel-symbolique et politico-institutionnel. Le marché et la technique s'autonomisent par rapport à la société; la technique serait alors autofinalisée et ne répondrait plus à une « finalité normative substantielle » (p.206). Ce qui aurait permis cela est l'action réduite à un mouvement mécanique et formel et une liberté abstraite; à des effets mesurables en termes de valeur et d'utilité; à des besoins réduits aux plaisirs et à la jouissance; à un « monde objectif,

à une pure disponibilité, à des ressources matérielles pouvant être modifiées à volonté » (p.206). Plus concrètement, c'est la difficulté à s'inspirer de « normes relevant d'une culture et d'une structure symbolique commune » qui caractérisent ce mode qui est davantage axé sur la société de marché et la « liberté des puissances organisationnelles ». Il y aurait éclatement de « l'unité entre le sujet, la technique et l'objet » (p.207) et échec du « projet démocratique » qui s'incarne dans le « débat public, de normes, lois et institutions capables, à travers la verticalité du pouvoir, d'encadrer les différents moments de l'action »(p.207). La capacité de régulation s'abandonne ainsi dans la technique (p. 107-108). Les systèmes, l'informatique etc. se substituent donc à la culture et à la société concrète; l'efficacité devient presque la finalité, que cela se fasse ou non dans la destruction de la nature et de la société.

Sociologie

La sociologie elle-même ne viserait pas une pensée visant la participation culturelle et politique des individus, mais devient davantage empirique, localisée, prévisionnelle et axée sur la gestion du comportement.

IA et monde commun

L'IA nous dégagerait du devoir de penser, de juger, d'envisager un monde commun. Il y a ainsi destruction des médiations culturelles et politiques au profit d'une régulation basée sur des « mécanismes de gestion du social » (p. 208). Tout est ainsi décomposé en « processus techniques ou algorithmiques », ce qui évacue les dimensions « subjective, significative ou expressive » et qui modifie

négativement la conception de l'humain et du lien social. La technologie impulse sa « nouvelle puissance d'organisation des rapports sociaux désencastrés des anciennes finalités normatives et de toute synthèse sociale-historique/ (p.209). L'éthique permet de maximiser cette adhésion à la logique du « capitalisme technoscientifique » (p.209).

Une telle vision basée sur les dispositifs communicationnels et managériaux ; sur la gestion technocratique; sur ces nouvelles médiations; sur le rôle des experts, etc. modifierait le rapport à l'espace public. Pour Freitag, l'éthique est un discours qui sert de propagande en vue de construire l'acceptabilité sociale des nouvelles technologies à l'intérieur de l'espace publicitaire spectaculaire qui a remplacé l'espace public politique : l'éthique devient un argument de vente pour donner au développement technologique un visage positif et « responsable. » Pour illustrer cela, Martin (2021) évoque à nouveau la Déclaration de Montréal, dont le principal but semble être d'attirer les investisseurs et de positionner favorablement la ville dans le domaine de l'IA.

Rapport à la démocratie

S'inspirant de Freitag, Martin considère que les questions sociales ont été confisquées à la société au profit des experts adhérant à la « rationalité technico-sociale autorégulée » (p.211). Les balises éthiques, jugées comme étant basées sur des valeurs abstraites, serviraient surtout à apaiser « les craintes des opposants » et assurer des « carrières universitaires et médiatiques » (p.211). Ces balises seraient basées sur le plus « petit dénominateur commun » et

viseraient principalement à « satisfaire l'intérêt du maximum de stakeholders, ou plutôt la perception ou l'opinion de ces parties prenantes ». Les sondages sur les dilemmes éthiques serviraient surtout à « alimenter le débat public » plus qu'à offrir de véritables pistes de réflexion.

Les chartes auraient ainsi tendance à produire du « bricolage », des juxtapositions de valeurs qui ne sont ni hiérarchisées ni structurantes socialement.

Solutions

Face à cela, l'auteur invite à une conception plus holistique et dialectique qui ne repose–pas sur les « médiatiques techniques, machiniques et algorithmiques » et invite à réfléchir le politique, la société et la civilisation.

Cela implique, à la manière de Castoriadis, de sortir des dichotomies éthique/politique, individu privé et citoyen public et d'entrevoir des modes de régulation moins aliénants et liberticides et basés sur une vision de l'autonomie comme vecteur de régime démocratique et non pas pilotés par les algorithmes et l'IA.

S'inspirant de Freitag, l'auteur oppose au kantisme, le recours à une ontologie qui prend en considération les conditions du maintien de l'existence plutôt que des catalogues de principes se qui inclut une vision plus large de la participation et du sens basée sur une hiérarchisation et une synthèse de valeurs et les possibilités d'envisager de « nouvelles instances politiques de responsabilités ».

En ce qui a trait à la dialectique, elle est mise à mal par les éthiques issues de la logique formelle et analytique qui a

tendance à « supprimer toute pensée négative » et demeure ainsi « prisonnière de la réalité et de l'ordre établi » (p. 216), et qui s'enferment aussi dans la science positiviste, l'unidmentionnalité et l'aliénation; qui reproduit les dominations; qui est au service de « l'accélération technico-économique destructrice de l'être » et incapable de considérer des médiations véritablement sociales. Il en appelle aussi à une philosophie reconsidérant l'existence, les normes et le politique.

Critique

Même en souscrivant à la critique contre les limites et la récupération des éthiques Rawlsienne, Kantienne et de l'éthique a posteriori, on ne peut constater que l'auteur se limite surtout à soulever des anecdotes et à critiquer les travaux de certains des partisans de ces approches ainsi qu'à soulever que ces éthiques participent à l'appropriation capitaliste plutôt que d'en critiquer ou même d'en décrire les fondements. Les évoquer devient un leitmotiv plutôt qu'une critique véritablement radicale de ce qu'elles sont. On s'arrête surtout à leurs impacts, ce qui est un départ, mais ne saurait suffire à offrir une réflexion approfondie sur le « comment » les « transcender » sinon que, justement, recourir à certaines abstractions. L'idée de les considérer comme forme de légitimation des IA est également intéressante, mais l'analyse s'arrête à un peu répéter les mêmes arguments.

L'auteur présente des perspectives déterministe, substantialiste, des théories critiques et du constructivisme. On s'aperçoit des difficultés des théories critiques – malgré

leurs très grandes importances à considérer ces évolutions technologiques dans la complexité. Cela pose donc la question du « comment » relever le défi de parfaire ces théories et on reste un peu sans réponse.

On peut aussi adhérer à la perspective selon laquelle ces éthiques sont développées par des experts qui sont, bien souvent, des promoteurs de l'IA. Donc, pour faire face à cela, en attendant des nouveaux modes de régulation globaux, on peut se questionner sur quels types de mécanismes mettre en place à court terme pour favoriser de véritables délibérations rationnelles et signifiantes, pour au moins baliser, en partie, ces cadreurs, et ainsi, arriver à de nouvelles formes de normativité. Le processus de transition entre les modes est plus ou moins bien explicité. Ce n'était pas le but de l'article, certes, mais cela invite à la réflexion.

Le fait de personnaliser les attaques peut certes permettre des illustrations concrètes, mais cela est un peu irritant, car cela nous fait sortir un peu de la réflexion rationnelle pour ne créer au final qu'un effet rhétorique suscitant soit l'adhésion ou le décrochage.

Les critiques à l'égard des chartes comme outils de légitimation, d'acceptabilité sociale voire comme instrument marketing sont intéressantes, cependant, il y a certes d'autres problèmes avec ce type de documents et cela est très peu abordé de façon concrète dans l'article.

La critique à l'égard des départements de philosophie qui seraient en quelque sorte instrumentalisés par les promoteurs de l'IA, bien qu'elle ne soit pas sans fondement, relève peut-être un peu de l'exagération.

Les pensées de Castoriadis et de Freitag sont également pertinentes. Cependant, cela suppose un cadre d'analyse très large qu'il est parfois difficile d'appliquer au cas particulier de l'IA. En fait, cela a un peu l'effet de considérer l'IA comme quelque chose de très structurant socialement et politiquement, ce qui n'est peut-être pas encore tantle cas; en fait, cela serait à vérifier. On comprend que le recours à ces outils vise surtout à soulever la nécessité de mises en place de diverses formes de régulations appartenant à divers registres. Cela n'est pas sans intérêt. Ce détour est cependant un peu laborieux, pour arriver à des conclusions qui sont somme toute relativement évidentes, ou peut-être pas non plus…

Il est difficile de voir par quels moyens concrets on pourrait opposer à la logique techniciste et capitaliste les solutions proposées (holisme, dialectique, politique, démocratisation, ontologie existentielle, sortie de l'aliénation), cette omission est probablement en grande partie volontaire de la part de l'auteur et invite donc à la réflexion sur les capacités de mettre en place de nouveaux espaces de médiations.

Apports

- Réflexion sur comment empêcher une instrumentalisation de l'éthique à des fins quasi publicitaires
- Réflexion sur l'articulation entre les limites de l'éthique de l'IA et les nouvelles médiations qui peuvent – ou non – passer par des instances intermédiaires. À cet égard, comment tout cela peut fonctionner?

- Questionnement sur comment passer d'une vision éthique à une vision un peu plus politique et macrosociale de l'éthique et, plus globalement, sur les formes démocratiques les plus aptes à contrer la domination de la vision technocentrée.

Martin, K., Waldman, A. (2022). Are Algorithmic Decisions Legitimate? The Effect of Process and Outcomes on Perceptions of Legitimacy of AI Decisions. Journal of Business Ethics. https://doi.org/10.1007/s10551-021-05032-7

Par Zeynep Torun

Question

Est-ce qu'il est possible de mesurer empiriquement les perceptions de la légitimité des décisions des entreprises prises par le biais d'algorithmes ?

Réponse

L'objectif de cet article est de construire une étude empirique à partir des travaux d'Elsbach (1994), Finch et al. (2012) et John et al. (2020) dans le but de mesurer empiriquement les perceptions de la légitimité des acteurs commerciaux utilisant un système de prise de décision algorithmique. Les auteurs étudient les différents facteurs influençant la perception de la légitimité, ainsi que leurs importances relatives sur les attitudes, les jugements ou les opinions des personnes. Ils utilisent les termes de dividende de légitimité et pénalité de légitimité

pour décrire les effets positifs et négatifs de la présence ou de l'absence d'une condition a sur la perception de la légitimité d'une décision algorithmique.

Argumentaire

La revue de littérature et les définitions

Prise de décision algorithmique des entreprises :

Les auteurs définissent la prise de décision algorithmique (PDA) comme des processus impliquant des algorithmes, ou des séquences d'opérations logiques et mathématiques, pour mettre en place des politiques par le biais de logiciels. Cet article se base sur des algorithmes développés à partir de données d'apprentissage et se concentre sur les perceptions de la légitimité des entreprises commerciales et privées. De ce fait, les auteurs partent d'une définition qui reconnaît le rôle des entrées de données, des ordinateurs et de l'automatisation des décisions. Toutefois, il existe des limites et des risques liés à ces systèmes de prise de décisions algorithmiques :

- La généralisation des situations par des prédictions probabilistes sur le futur avec le risque de commettre une erreur sur des situations ambiguës qui sont difficiles à prédire.
- L'invasion de la vie privée des personnes, car les systèmes algorithmiques fonctionnent grâce à une large collecte de données pour pouvoir apprendre et analyser les résultats. - Les données sur lesquelles

ces systèmes se basent sont biaisées en matière de la race, du genre, du sexe et de la situation socio-économique. Cela peut conduire à des résultats discriminatoires.

- Les données sur lesquelles ces systèmes se basent sont biaisées en fonctionne de la race, du genre, du sexe et de la situation socio-économique. Cela peut conduire à des résultats discriminatoires.

Légitimité :

Au niveau organisationnel, les entités sont perçues comme légitimes lorsque leurs actions sont « désirables, propres ou appropriées dans le cadre d'un système de normes, de valeurs, de croyances et de définitions socialement construit » (Suchman, 1995, p.574). Les entreprises légitimes poursuivent « les objectifs socialement acceptables d'une manière socialement acceptables » (Ashforth et Gibbs, 1990, p. 177). La légitimité est une « attitude » (Jahn et al. 2020, p.546) ou une « perception » qui est liée à la crédibilité institutionnelle.

Hypothèses :

Les auteurs développent 5 hypothèses comme les moteurs de la légitimité perçue des décisions algorithmiques à partir des travaux effectués sur la légitimité et la décision algorithmique.

Tableau 1 : Hypothèses

Type de décision La perception des individus de la légitimité de l'utilisation commerciale des algorithmes dans les prises de décision varie selon leur évaluation de l'importance de la décision dans leur vie.	H1 : Lorsque l'importance de décision augmente, la perception des individus de la légitimité de l'utilisation d'un algorithme pour prendre une décision diminue.
Résultats La recherche sociojuridique sur la légitimité suggère également que les résultats des décisions sont importants pour les perceptions populaires de	H2 : Tous les autres facteurs étant constants, un bon résultat positif est associé à un dividende de légitimité, ou à une augmentation de la légitimité perçue de la décision.

la légitimité des institutions et des processus qui ont conduit à ces résultats	
Raisons arbitraires Comme la littérature sur la responsabilité algorithmique met en évidence, les algorithmes prédictifs basant sur des données et des modélisations discriminatoires produisent des résultats injustes.	H3 : L'utilisation de facteurs arbitraires ou fondés sur la race a un impact négatif sur la légitimité perçue.

Gouvernance La littérature concernant la légitimité et la gouvernance défend que les perspectives de la légitimité doivent augmen ter autant que la solidité des mécanismes de gouvernance procédurale. Selon les auteurs, l'utilisation des facteurs raciaux ou arbitraires peut être si délégitimant qu'aucune procédure ne puisse le remédier.	H4 : Tout régime de gouvernance apporte un dividende de légitimité à la prise de décision algorithmique par rapport à l'absence de gouvernance, mais le dividende de légitimité est d'autant plus important que la gouvernance procédurale est solide. H4b : Les mécanismes de gouvernance procédurale robustes modèrent les pénalités de légitimité associées à de mauvais résultats. H4c : Quelle que soit la forme de la gouvernance procédurale utilisée, il existe une pénalité de légitimité associée aux algorithmes qui utilisent des facteurs arbitraires ou racia ux.

Méthodologie

Les auteurs utilisent la méthodologie de vignettes factorielles dans l'objectif d'explorer l'importance relative de la gouvernance, de la procédure et des résultats sur la légitimité perçue des décisions prises par des entreprises utilisant un système de PDA. Grâce à cette méthodologie, les auteurs proposent des scénarios à des participants où plusieurs facteurs sont systématiquement proposés dans la vignette. Par la suite, les participants jugent le scénario et effectuent une notation selon une échelle allant de -100 (« Fortement en désaccord ») à +100 (« Fortement en accord ») pour répondre à la question de légitimité (« Cette décision est légitime »).

Les auteurs effectuent 9 enquêtes différentes en quatre mois. Les enquêtes ont été menées sur Amazon Mechanical Turk, une plateforme de crowdsourcing pour les chercheurs aspirant à recruter des participants pour leurs études (« HIT »). Tous les participants ont noté 20 à 30 vignettes (dépendamment de condition), environ en 10 minutes, les états-uniens ont été payés 1.60-1.80, avec une note de 95% de HIT.

Facteurs

Les facteurs se trouvant dans des vignettes proposées aux participants constituent les concepts théoriquement importants qui peuvent influencer la tâche de notation.

Tableau 2 : Facteurs et leurs niveaux

Facteurs	Niveaux
Type de décision (H1) Deux conditions sont assignées par hasard aux participants : - L'importance au niveau individuel - L'importance au niveau sociétal	- Quelles publicités les gens voient-ils en ligne? - Quelles chansons sont suggérées par les plateformes de musique? - Quels candidats sont embauchés pour un emploi? - Quelles demandes d'assurance sont remplies? - Quel contenu vidéo est retiré par une plateforme en ligne?
Résultats (H2)	- Positif pour l'individu - Négatif pour l'individu
Arbitraire (H3)	- Arbitraire - Discriminatoire

	- Nulle
Gouvernance (H4, H4b, H4c)	- Transparence : l'organisation informe les individus qu'un logiciel a fait cette décision (Diakopoulos, 2020) - Évaluation d'impact : l'organisation a réalisé une évaluation de l'impact du processus algorithmique sur l'équité et la vie privée (Yam&Skorburg, 2021). - Gouvernance audit : une tierce organisation a réalisé un audit indépendant annuel pour garantir que les décisions algorithmiques ne sont pas biaisées (Mittelstadt, 2016). - Gouvernance humaine : « human in the loop » du

	processus algorithmique (Elish, 2019) - Recours : dans lesquels la décision peut être contestée par l'individu auprès d'une commission de révision interne (Mulligan et al. 2020).

Résultats

Tableau 3 : Hypothèses et résultats

Hypothèses	Résultats
H1 : Lorsque l'importance de décision augmente, la perception des individus de la légitimité de l'utilisation d'un	Les participants considèrent que les décisions moins essentielles prises par l'intelligence artificielle (IA) sont perçues plus légitimes que les décisions plus essentielles.

algorithme pour prendre une décision diminue.	
H2 : Tous les autres facteurs étant constants, un bon résultat positif est associé à un dividende de légitimité, ou à une augmentation de la légitimité perçue de la décision.	Les auteurs trouvent qu'un mauvais résultat est une pénalité de légitimité comparée à un bon résultat. Un bon résultat a un impact positif sur la légitimité perçue.
H3 : L'utilisation de facteurs arbitraires ou fondés sur la race a un impact négatif sur la légitimité perçue.	Toutes choses égales, l'inclusion des facteurs arbitraires dans l'Enquête 4 diminue le taux de légitimité, de 25,53 (Enquête 2) à -14,05 (Enquête 4). En général, les enquêtes où il se trouve des facteurs arbitraires sont les enquêtes ayant les taux de légitimité les plus faibles.

H4 : Tout régime de gouvernance apporte un dividende de légitimité à la prise de décision algorithmique par rapport à l'absence de gouvernance, mais le dividende de légitimité est d'autant plus important que la gouvernance procédurale est solide.	Le mécanisme de gouvernance plus solide, qui permet un recours, a un impact positif sur la légitimité perçue. Cependant, les mécanismes de gouvernance alternatifs, tels que l'évaluation d'impact, « human in the loop » et l'audit du programme, réduisent la légitimité de la décision comparée au simple avis. Ce résultat va à l'encontre de l'hypothèse et surprend les auteurs.
H4b : Les mécanismes de gouvernance procédurale robustes modèrent les pénalités de légitimité associées à des mauvais résultats.	Pour les mauvais résultats, l'appel à une procédure de recours apporte un dividende de légitimité. Mais si la décision prise ramène des bons résultats, des procédures de recours n'ont aucune importance. Le taux de légitimité moyen d'un mauvais résultat, même avec une

	procédure de recours, reste inférieur à celui d'un bon résultat.
H4c : Quelle que soit la forme de la gouvernance procédurale utilisée, il existe une pénalité de légitimité associée aux algorithmes qui utilisent des facteurs arbitraires ou raciaux.	Le dividende de légitimé d'inclure une procédure de recours améliore la légitimité des décisions incluant des facteur arbitraire come la journée de la semaine, mais détériore lors qu'on inclue des facteurs injustes comme la race.

Conclusion

L'article qui a été résumé dans ce travail nous propose une étude empirique sur la légitimité des décisions algorithmiques prises par des entreprises. Ils étudient de différents facteurs – le type de décision, les résultats, les facteurs arbitraires et la gouvernance - et leurs effets sur la perception de la légitimité décisionnelle. De plus, cette étude a démontré que la légitimité perçue varie inversement avec l'importance de la décision. Les auteurs sont partis de l'hypothèse qu'avoir des procédures de gouvernance solide sur les décisions algorithmiques prises par des entreprises a un effet positif sur la légitimité. Cependant, les données collectées par les

auteures présentent un résultat plus nuancé. La seule forme de procédure de gouvernance ayant un dividende de légitimité est celle qui est la plus solide : un recours à un décideur humain. Quant aux résultats, lorsque la décision prise a un bon résultat ou a un bon effet sur la vie d'un individu, la légitimité de cette décision remporte sur la procédure de gouvernance. Les auteurs ont également observé que les décisions ayant peu d'importance avec de bons résultats ont peu de dividende de légitimité.

Contribution

Cet article contribue à plusieurs niveaux dans la recherche en sciences sociales comme la légitimité, l'éthique de l'entreprise, l'éthique de l'intelligence artificielle et la politique publique. Cette étude confirme que n'importe quelle décision prise par un algorithme ayant des facteurs arbitraires discriminatoires a des effets sur la légitimité de l'entreprise. La contribution concernant la recherche sur l'éthique de l'IA, cet article met l'accent sur les implications morales qui devraient préoccuper les entreprises notamment, concernant la construction du système de PDA. Les entreprises devraient non seulement de se préoccuper de la transparence et de la responsabilité, mais aussi de la discrimination. Les recommandations des auteurs pour les personnes ayant un intérêt pour la politique publique est d'envisager des limites plus importantes sur les entrées et les utilisations algorithmiques plutôt que des garanties procédurales, car presque toutes les décisions algorithmiques sont considérées comme illégitimes lorsqu'ils utilisent des facteurs fondés sur la race ou arbitraires.