



Est-il possible de construire un algorithme neutre ?

Rapport HTI01

Julien CALONNE - Kevin JAULIN - Matthieu SCHULER

Semestre P23



utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

Sommaire

Sommaire.....	1
Introduction.....	2
I - Le manque de neutralité dans la construction des algorithmes.....	3
II - Pourtant, on pourrait espérer une neutralité car les algorithmes proposent des résultats de manière automatique et mathématique.....	6
III - Vers une utilisation responsable des algorithmes et une amélioration de leur équité.....	8
Conclusion.....	12
Bibliographie.....	13

Introduction

Un algorithme est une suite d'instructions exécutés par une machine. Il est donc en théorie neutre puisqu'il ne s'agit pas d'une intelligence humaine, qui aurait été de par son caractère sociable mais aussi au travers de son éducation, influencée par la société. L'algorithme neutre permettrait alors d'émettre des avis impartiaux, équitables et transparents. Il pourrait appliquer les mêmes critères pour assurer une continuité des décisions et ainsi être plus fiable qu'un être humain.

Bien que ces algorithmes soient majoritairement utilisés comme aide à la décision en Europe, ils peuvent avoir un impact non négligeable sur des décisions humaines. Cependant, il est récurrent que des biais apparaissent au fur et à mesure de leur utilisation engendrant des conséquences préjudiciables. De plus, alors que les systèmes informatiques et l'intelligence artificielle occupent une place grandissante dans notre société, il est essentiel de s'interroger sur la neutralité des algorithmes qui les composent pour s'assurer qu'ils ne soient pas vecteurs de discrimination.

En effet, la question de la neutralité des algorithmes ne peut être abordée de manière isolée. Les algorithmes sont souvent basés sur des données historiques qui peuvent refléter des préjugés et des inégalités profondément enracinées dans la société. Même si les algorithmes eux-mêmes sont construits de manière neutre, ils peuvent reproduire et amplifier les inégalités lorsqu'ils sont appliqués dans des contextes défavorables.

A titre d'exemple, des études ont relevé que des algorithmes utilisés pour évaluer les demandes de prêt pouvaient introduire des biais basés sur des caractéristiques personnelles telles que le sexe ou l'origine géographique des demandeurs. Cela a conduit à des refus injustifiés de prêts pour certains groupes et à des taux d'intérêts plus élevés pour d'autres, perpétuant ainsi des inégalités.

Dans un autre domaine, les algorithmes de surveillance notamment utilisés par les forces de police aux Etats-Unis peuvent également être sujets à des biais en fonction de leur construction et des données utilisées. Cela peut entraîner une surveillance disproportionnée et injuste sur certains groupes, en particulier les minorités raciales.

Il est certain que tous les algorithmes ne sont pas neutres, tout comme les humains qui les utilisent, les données qui les composent ou encore le programme lui-même. Il est donc nécessaire de mener une réflexion approfondie afin de comprendre pourquoi les biais algorithmiques sont inévitables et comment ceux-ci apparaissent, entraînant ainsi des enjeux éthiques majeurs. Nous pouvons ainsi nous demander comment nous pouvons identifier ces biais affectant la neutralité des algorithmes. Puis, s'il est possible d'y faire face en les corrigeant ?

I - Le manque de neutralité dans la construction des algorithmes

Tout d'abord il est important de comprendre la différence entre impartialité et neutralité, en effet alors que tous deux impliquent de ne pas prendre parti pour l'un ou l'autre des acteurs concernés, le principe d'impartialité introduira de plus la notion de prise de décision.

“L'impartialité, contrairement à la neutralité, implique donc une décision. L'impartial choisit, le neutre s'abstient de le faire.”, La Chronique Mensuel d'Amnesty International, N°294, avril 2011

Or, comme explicité plus tôt, n'étant pour le moment en France encore qu'au stade d'aide à la décision, on ne peut pas vraiment considérer l'IA en tant que décisionnaire et donc évaluer son caractère impartial.

Alors que les champs d'application de l'IA ne cessent d'augmenter, elle touche aussi aujourd'hui des domaines sensibles tels que la justice, la médecine ou encore la finance. Cependant l'accès par les algorithmes à des données personnelles aussi sensibles, se voudrait être rassurant pour l'utilisateur, tant au niveau de la sécurité appliquée sur celles-ci que sur le traitement qui sera effectué dessus. Seulement, comme il sera présenté plus bas, on peut observer que certains programmes s'éloignent plus ou moins consciemment de ce caractère neutre. Un algorithme étant construit suivant trois composantes principales, le développeur, les données utilisées et le code en lui-même, chacun de ces trois éléments peuvent être plus ou moins soumis à des acteurs extérieurs qui vont influencer leur comportement et leur évolution.

Dans la majorité des cas, ce seront, dans un premier temps, les données qui ont permis d'entraîner le modèle qui seront la source de biais. En effet, la qualité des données, en dépit de l'algorithme lui-même, est primordiale puisque le résultat de l'algorithme sera différent en fonction de la façon dont il est entraîné. Les développeurs doivent donc s'assurer que les données fournies, respectent de façon appropriée à trois facteurs, la véracité de l'information, une couverture quasiment totale, mais surtout une représentation fidèle et contextuellement correcte. En effet, le contexte et l'évolution de la donnée doivent être pris en compte dans le processus de développement de l'algorithme, à titre d'exemple les données boursières peuvent tout à fait illustrer ce point, effectivement les données récoltées sur une période d'un krach boursier par rapport à une période plus classique seront bien différentes et ne représenteront pas du tout la même situation, une recontextualisation adaptée de la donnée est donc nécessaire.

En tant que programme d'aide à la décision automatique, le code algorithmique est lui aussi un facteur influent, et peut amener à des changements de comportement au sein de la société. De façon intentionnelle ou non, il est possible d'influencer au travers des résultats calculés et présentés à l'utilisateur, sa façon d'appréhender des concepts ou encore de renforcer des préjugés, et par extension affecter son jugement, ses décisions et son comportement de manière plus ou moins subtile. En effet, bien que majoritairement corrigés aujourd'hui, l'on pourrait évoquer les problèmes concernant les suggestions de recherche Google, qui en 2016 lorsque l'utilisateur commençait ses recherches par “Les musulmans sont...”, proposait des suggestions d'autocomplétion telles que “Les musulmans sont mauvais” ou encore “L'Islam doit être détruit”. Ceci pouvant de fait renforcer une discrimination religieuse. Le problème majeur ici, serait que même de façon non-intentionnelle ou non prévue par l'entreprise créatrice de l'algorithme, les

utilisateurs pourront être en présence d'éléments discriminatoires qui suivant leurs cultures propres ou bien juste l'action répétée, les inciteront à développer un regard différent sur la société qui les entoure. Un autre effet de bord de ce phénomène sera l'image rendue par l'entreprise en cas de divulgation publique du problème, entraînant ainsi une perte de confiance et une baisse de légitimité du grand public envers les systèmes. Un second élément peut parfois intervenir en tant que créateur de biais au sein même du programme, l'interface présentée à l'utilisateur. En effet, alors que les utilisateurs accordent une importance de plus en plus grande à l'ergonomie des outils qu'ils utilisent, certains automatismes, retranscrits dans le code, commencent à apparaître, avec par exemple l'utilisation de la couleur verte pour valider ou représenter quelque chose de positif, et à contrario la couleur rouge à connotation souvent négative. Par extension, la disposition et colorimétrie utilisées pour représenter une information, peuvent donc permettre d'influencer l'utilisateur, en cherchant soit à attirer son regard ou tout au contraire rendre l'information cognitivement difficile à appréhender.

Un dernier point reste à aborder ici, puisque point d'algorithmes ne serait possible sans une équipe de développeurs. Pour le moment, ceux-ci sont humains et donc sujets à être influencés consciemment ou non. Point central, du développement du programme, ce seront cependant eux les maillons faibles de la chaîne de vie de l'algorithme. Effectivement la quantité de responsabilités qui lui incombe est bien plus importante que le reste des éléments constitutifs d'un algorithme, puisque c'est lui qui aura la responsabilité de contrôler la conformité des données utilisées. On entend par là que le développeur devra vérifier que les données sont à la fois représentatives mais surtout équilibrées. Un second élément jouant en la défaveur des développeurs sera leurs caractères humains, ils ne seront, en effet, pas capables de par leur nature, de certifier l'absence de biais, aussi bien des biais concernant les données utilisées que la présence de ceux-ci dans le code. S'illustrant par exemple, sur la période de 2014 à 2018 où une équipe d'Amazon s'essayait au développement d'une IA permettant un recrutement automatisé, seulement les données utilisées correspondaient alors au CV reçus entre 2004 et 2014, période pendant laquelle on pouvait observer une nette prédominance masculine dans le secteur des nouvelles technologies. La résultante de cette expérience fut qu'après entraînement, l'IA détermina que les candidats masculins pour ces postes étaient préférables et rejetait donc toutes candidatures portant une référence aux femmes..

La question de responsabilité se pose alors, en effet qui blâmer en cas de difficulté : l'algorithme pour ne pas avoir appris correctement, les développeurs pour ne pas avoir détecté les problèmes, ou encore faut-il imputer les propriétaires pour avoir autorisé la mise sur le marché d'une application défaillante ?

Tout d'abord, il faut savoir que dans l'Union Européenne et depuis le règlement d'Avril 2021, quoi qu'il arrive l'algorithme ne peut être légalement et pénalement reconnu responsable. Malheureusement, il n'existe pas de réponse parfaite pour cette question, cependant certaines actions sont possibles afin de se prémunir ou de tout du moins se déresponsabiliser. Avant tout début de développement, il est primordial que l'équipe de développeurs mène une réflexion éthique notamment sur les risques de biais, ainsi que leurs implications possibles afin de prévoir comment limiter les risques en cas de dérive. Ensuite, la nature neutre de l'algorithme peut être plus ou moins défendue par les propriétaires, en justifiant par exemple de critères essentiels de fonctionnement comme pour Tinder, utilisant un critère de désirabilité. Une autre solution possible

serait de volontairement responsabiliser l'utilisateur, l'exemple aujourd'hui le plus significatif étant celui de ChatGPT. Effectivement, lorsque l'on pose à l'IA, la question de sa propre neutralité voici sa réponse :

- *Es-tu un algorithme neutre ?*
- *En tant qu'algorithme de langage développé par OpenAI, je suis conçu pour être neutre et impartial. Mon objectif principal est de fournir des réponses et des informations basées sur les données disponibles et les modèles de langage appris lors de mon entraînement. Cependant, il est important de noter que je suis une création humaine et que je suis influencé par les données qui m'ont été fournies lors de mon apprentissage. Par conséquent, il est possible que des biais involontaires ou des erreurs se glissent dans mes réponses. Il est toujours recommandé de consulter plusieurs sources d'information et de faire preuve de discernement critique lors de l'évaluation des réponses fournies par tout système automatisé."*

ChatGPT, à la date du 27 Mai 2023

On peut noter ici que bien qu'il soit conscient de son objectif, il est tout aussi conscient de ses propres défaillances, et invite même l'utilisateur à compléter sa réponse par des recherches plus approfondies.

Ainsi, on observe aujourd'hui que la neutralité d'un algorithme est un objectif à atteindre plutôt qu'une garantie absolue. Les créateurs d'algorithmes doivent être conscients des biais potentiels et prendre des mesures pour minimiser ces biais afin de maximiser la neutralité et puis dans le futur l'impartialité de leur algorithme. Cela signifierait qu'à l'heure actuelle aucun algorithme ne serait neutre, il n'est pas possible d'être aussi catégorique sur le sujet.

II - Pourtant, on pourrait espérer une neutralité car les algorithmes proposent des résultats de manière automatique et mathématique

Nous venons de montrer qu'un algorithme ne pouvait pas être neutre. Cependant, les algorithmes sont basés sur des instructions logiques et mathématiques. Cette construction leur permet alors de traiter les données de manière impartiale et cohérente. Contrairement aux humains, les algorithmes ne sont pas influencés par des préjugés, des émotions ou des opinions personnelles. Ils sont conçus pour appliquer les mêmes critères et processus à toutes les données d'entrée, ce qui peut contribuer à la neutralité des résultats.

Dans une interview accordée à France Inter en février 2023, le sujet autour de l'impact de l'intelligence artificielle (IA) dans le domaine médical a été évoqué par Jean-Emmanuel Bibault, un oncologue spécialisé dans l'IA en médecine. L'avancée fulgurante de cette technologie dans le domaine a suscité de nombreuses interrogations sur la place du médecin dans le système de santé et les questions éthiques qui en découlent. Néanmoins, une neutralité est rendue possible grâce aux algorithmes utilisés. En effet, ils offrent des résultats de manière automatique et mathématique. Au cours de l'interview, on comprend avec les explications de Jean-Emmanuel Bibault que l'IA dans ce domaine repose sur des algorithmes conçus pour traiter les données de manière cohérente et impartiale. Contrairement à nous, ces algorithmes ne sont pas influencés par des préjugés, des émotions ou des opinions personnelles. Ils appliquent les mêmes paramètres et processus à toutes les données qu'on lui donne en entrée, ce qui permet une approche plus neutre dans les résultats obtenus. Le principal rôle de l'IA en médecine est l'analyse d'images médicales telles que les scanners ou les IRM. Certains algorithmes sont capables de détecter automatiquement des anomalies ou des signes de maladies, comme les hémorragies cérébrales ou les cancers du sein. En utilisant des critères objectifs et en évitant les biais humains, ces algorithmes fournissent des résultats plus fiables et peuvent contribuer à une approche neutre dans le diagnostic médical.

Une autre situation qui témoigne de l'impact de la nature automatique et mathématique des algorithmes se trouve dans le domaine de la traduction automatique. Lorsqu'il s'agit de traduire un texte d'une langue à une autre, les algorithmes de traduction automatique utilisés par DeepL ou encore Google Traduction permettent d'obtenir des résultats bluffants. Ces algorithmes se basent sur des modèles statistiques et des règles linguistiques pour analyser et traiter les phrases afin de produire des traductions.

Nous avons étudié un article du Journal Of Data Mining and Digital Humanities¹ "Vers une robotique du traduire", qui porte un avis objectif sur le devenir du métier de traducteur, traductrice au vue de l'émergence des algorithmes de traductions automatiques toujours plus performants. Il est montré que de prime abord, la machine permet de gagner en productivité par rapport à un humain à hauteur de 150 à 200%. De plus, avec l'arrivée des algorithmes par réseaux neuronaux (bien plus performant que ceux utilisant une simple base de données), la machine est capable de

¹ JDMDH est un projet conjoint du CNRS, de INRAE et d'INRIA se focalisant sur la rencontre de l'informatique et des disciplines des sciences humaines, profitant des outils numériques comme la visualisation de données, la recherche d'informations, les statistiques, la fouille de textes et publie des travaux scientifiques dans le prolongement des sciences humaines traditionnelles.

retranscrire de la façon la plus neutre possible une phrase écrite dans une autre langue. Évidemment, on voit ici aussi la limite qui se situe dans la difficulté d'obtenir certaine tournure de phrase spécifique à des régions de certains pays. Cependant, d'un point de vue algorithmique, sa nature mathématique et automatique permet dans ce contexte d'obtenir des résultats plus neutres et objectifs.

Ensuite les résultats des algorithmes peuvent s'approcher de la neutralité et jouer un rôle important dans la réduction des inégalités. Intrinsèquement et même si nous avons vu que cela pouvait être difficile à atteindre, les algorithmes sont conçus pour traiter les données de manière impartiale, en appliquant des critères objectifs et en évitant les biais humains. Cette approche peut contribuer à une prise de décision plus équitable et permettre une égalité des chances plus juste.

Nous pouvons ici reprendre l'exemple concernant la santé. Nous avons vu que leurs fonctions principales dans le domaine de la santé se concentrent sur l'analyse. En effet, ces algorithmes peuvent analyser de grandes quantités de données médicales pour fournir des diagnostics et des recommandations de traitement. L'avantage de ces algorithmes réside alors dans leur capacité à traiter objectivement les informations, en se basant sur des critères médicaux et des preuves scientifiques. Cela permet de réduire les discriminations dans les diagnostics et les traitements, donnant ainsi une approche plus équitable et plus précise de la prise en charge des patients, indépendamment de leur origine sociale ou de leur statut économique.

Enfin, les algorithmes offrent la possibilité d'approcher la neutralité dans leurs résultats en appliquant des règles similaires à toutes les données d'entrée garantissant une cohérence et une fiabilité pouvant contribuer à réduire les biais subjectifs.

Nous pouvons prendre l'exemple de l'évaluation des candidatures pour des postes d'emploi. De nombreuses entreprises utilisent des systèmes d'évaluation automatisés pour trier et classer les candidatures en fonction de critères prédéfinis. Ces algorithmes analysent des données telles que les compétences, l'expérience et les qualifications des candidats pour générer un score ou une recommandation.

D'après un article, "Les algorithmes dans le recrutement et le risque discriminatoire", rédigé par le ministère de la transition écologique, un algorithme est sujet à un risque discriminatoire. Cependant, ils ont l'ambition d'être plus objectifs que ne le seraient les personnes. Ils permettent une meilleure neutralité dans le résultat obtenu de par les règles qui lui sont appliquées. En effet, l'article met en avant le fait que si les développeurs réalisant ces algorithmes respectent un certains nombres de compétences, ces derniers ne peuvent être que plus fiables et objectifs dans les résultats qu'ils fournissent. Voici une liste non-exhaustive des conditions et compétences à respecter pour rédiger un algorithme neutre dans le domaine du recrutement :

- former les concepteurs d'algorithmes à l'éthique, aux risques de biais et de discrimination
- tester les algorithmes et évaluer / identifier les biais et discriminations potentielles
- développer des chartes d'engagement éthique de l'utilisation des algorithmes.

De plus, l'utilisation de règles et de critères préétablis dans les algorithmes d'évaluation des candidatures permet de minimiser les influences subjectives et les biais inhérents aux évaluations humaines. Les algorithmes sont programmés pour se concentrer sur les compétences, l'expérience et les qualifications pertinentes pour le poste, sans tenir compte de caractéristiques personnelles telles que le genre, l'origine ethnique ou l'apparence physique. Cette approche automatique et mathématique assure une cohérence dans l'évaluation des candidatures et contribue à réduire les biais discriminatoires.

Toutefois, en dépit des avantages potentiels des algorithmes en termes d'efficacité et de neutralité, il est important de souligner que leur utilisation doit être guidée par une responsabilité accrue et une volonté de promouvoir l'équité. Les défis associés à la partialité et à la discrimination algorithmique ne peuvent être ignorés, et il est crucial d'adopter des mesures pour atténuer ces risques. En examinant de plus près les limitations des algorithmes et en prenant des mesures pour améliorer leur équité, il devient possible de construire un avenir où ces outils technologiques sont utilisés de manière plus responsable et éthique.

III - Vers une utilisation responsable des algorithmes et une amélioration de leur équité

Nous avons vu qu'espérer d'un algorithme qu'il soit neutre, fiable et impartial n'est pas une réalité qu'un idéal difficilement atteignable. Il est cependant important de tendre vers cet idéal et de tenter de s'en rapprocher au maximum. Différents acteurs travaillent dans cet objectif : entreprises, communautés, Europe.

Certaines entreprises s'engagent sur les thématiques de développement responsable et proposent aux développeurs des outils pour identifier et corriger des biais. L'entreprise IBM, par exemple, a lancé en 2017 "**AI Fairness 360**" : une véritable boîte à outils disponible pour les développeurs. Elle leur permet d'évaluer et d'atténuer les biais présents dans leurs modèles d'apprentissage. Cette boîte à outils est accessible dans un package Python et téléchargeable par n'importe qui. Elle utilise un ensemble de métriques qui compare les données des modèles afin d'en ressortir les parties contenant des biais. Ainsi, les développeurs peuvent adapter leurs données en fonction des réponses d'AI Fairness 360 et anticiper par avance de possibles inégalités créées par leurs algorithmes.

IBM propose une démonstration sur son site internet qui nous permet en quelques clics de nous rendre compte des biais créés par les données. La démonstration donne pour exemple une IA qui prédit si le revenu d'un individu dépasse 50.000 \$ par an en fonction des données d'un recensement. Le graphe visible sur la figure 1 nous montre la différence du taux de résultats favorables reçus par le groupe non privilégié au groupe privilégié. Dans notre cas, le groupe privilégié correspond à des individus blancs et le groupe non privilégié à des individus noirs. La différence idéale correspondrait à une valeur de 0 et d'après le site d'IBM l'équité est respectée si la valeur est comprise entre -0.1 et 0.1. L'outil AI Fairness 360 nous montre alors que la valeur -0.18 correspond à un biais vecteur de discrimination entre les deux groupes étudiés.

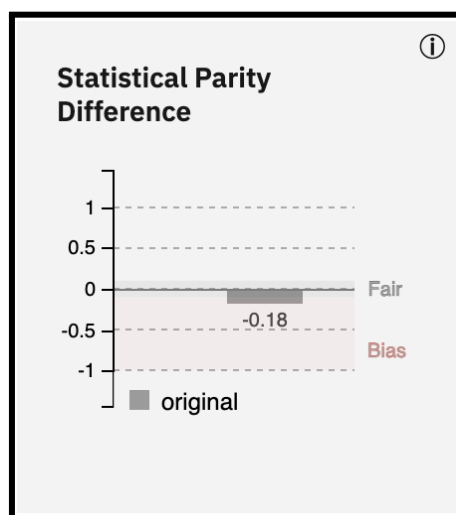


Figure 1 : Graphique de différence statistique entre deux groupes

Google s'est aussi lancé, il y a 5 ans, dans la détection des biais dans les jeux de données avec un outil Open-Source : **What-If Tool** (WIT). Il s'agit d'une interface visuelle qui permet de mieux comprendre les résultats de modèles de machine learning. Cet outil examine la diversité des données en vérifiant si les résultats des algorithmes sont similaires en changeant des données précises : origine ethnique, localisation... L'outil fonctionne sur la plupart des plateformes, peut être utilisé avant et après la création des modèles et fait ses preuves même sur les algorithmes de Google.

Malgré tous les outils développés pour détecter les biais, l'humain est souvent le meilleur pour faire preuve de discernement entre certains caractères distinctifs (la détection d'émotions par exemple). Il existe une méthode permettant d'inclure les humains dans l'apprentissage des IA, elle s'appelle l'Humain dans la boucle. D'après le site [l'IA de A à Z](#), il s'agit d'utiliser un humain pour éduquer, tester ou paramétrer un système d'IA pour l'aider à produire des résultats plus fiables. Cette méthode hybride permettrait de combiner les deux intelligences simultanément afin de s'assurer que l'apprentissage du modèle va dans la direction souhaitée. Dans le cas de la détection d'image, l'humain indique au système quel signe distinctif chercher lorsque la nuance est faible. Par exemple, il est simple d'identifier un léopard d'un zèbre en fonction de son pelage, mais moins évident de le discerner d'un guépard. Il faut donc se baser sur d'autres critères comme l'allure de l'animal (plus élancée pour le léopard). Cette méthode appliquée, sur des données plus sensibles, permet de débiaiser les données en amont en sélectionnant manuellement des critères.

L'humain dans la boucle se rapproche du crowdsourcing (production participative en français). Il s'agit d'utiliser l'intelligence, la créativité et le savoir-faire d'un grand nombre de personnes pour réaliser certaines tâches habituellement réalisées à l'intérieur d'une entreprise. En impliquant un large éventail de contributeurs, le crowdsourcing peut aider à réduire les biais individuels en donnant la possibilité à un groupe plus représentatif de participer à la collecte des données. Cela contribue à élargir la portée des ensembles de données et à réduire les biais inhérents à une seule personne ou à un petit groupe restreint. De plus, le crowdsourcing peut également être utilisé pour vérifier et valider les données collectées. En faisant appel à une communauté de contributeurs pour examiner et annoter les données, on peut identifier et corriger les biais potentiels qui pourraient être présents. Ces mécanismes de vérification par les pairs peuvent permettre de détecter les préjugés et de garantir une plus grande impartialité des données utilisées pour l'entraînement des modèles d'IA.

D'autre part, les avancées juridiques autour de l'intelligence artificielle témoignent d'une prise de conscience croissante des enjeux éthiques et réglementaires entourant cette technologie. En avril 2021, la Commission européenne a présenté sa proposition de règlement sur l'IA, qui vise à établir un cadre juridique clair et cohérent pour garantir l'utilisation responsable et éthique des IA dans l'Union européenne (UE).

Le règlement proposé comprend des dispositions régissant les systèmes d'IA à haut risque, tels que ceux utilisés dans les secteurs de la santé, de la sécurité, des transports et de l'administration publique. Ces systèmes seraient soumis à des obligations spécifiques, telles que la documentation, l'évaluation de la conformité, la gouvernance des risques et la transparence. Des interdictions absolues sont proposées pour certains types de systèmes d'IA considérés comme nuisibles pour les droits fondamentaux, comme la notation sociale automatisée. Le règlement prévoit également des exigences relatives à la collecte et à l'utilisation des données dans les systèmes d'IA. Il met l'accent sur la nécessité d'une gestion responsable des données, en veillant à ce qu'elles soient collectées et utilisées de manière éthique, respectueuse de la vie privée et en conformité avec les lois sur la protection des données de l'UE (RGPD).

Conclusion

La problématique de la neutralité des algorithmes suscite des débats et des préoccupations quant à leur capacité à être impartiaux et équitables. Après avoir examiné les différentes dimensions de cette question, il devient clair que l'idée d'un algorithme neutre absolu est difficile à réaliser.

Les algorithmes, bien qu'ils soient conçus pour traiter les données de manière objective, sont toujours influencés par des biais inhérents à la collecte et à l'utilisation des données. Les données historiques utilisées pour former les algorithmes peuvent refléter les inégalités et les préjugés existants dans la société. De plus, les choix faits par les développeurs lors de la conception de l'algorithme, qu'il s'agisse de sélectionner les données ou de définir les critères de traitement, peuvent introduire des biais involontaires.

Il est donc essentiel d'être conscient de cette situation et de garder un regard critique sur les résultats suggérés. Plutôt que de chercher une neutralité absolue, il est plus réaliste et constructif de se concentrer sur l'amélioration de la transparence, de la responsabilité et de la réduction des biais dans la conception et l'utilisation des algorithmes. Cela nécessite une collaboration étroite entre les experts techniques, les chercheurs en éthique et la société dans son ensemble afin de définir des normes et des réglementations qui guident l'utilisation des algorithmes de manière éthique et responsable.

Bibliographie

I – Le manque de neutralité dans la construction des algorithmes

- *La Chronique*, Mensuel d'Amnesty International, n°294, avril 2011
<http://www.jbjv.com/Comment-distinguer-neutralite-et.html#:~:text=C'est%2C%20lorsque%20l',s'abstient%20de%20le%20faire.>
- *The Guardian* : (Auto)complete fail : how search suggestions keep catching Google out
<https://www.theguardian.com/technology/shortcuts/2018/may/22/autocomplete-fail-how-search-suggestions-keep-catching-google-out>
- *Ouest France* : Amazon a automatisé des recrutements, mais le logiciel excluait les femmes de certains postes:
<https://www.ouest-france.fr/economie/entreprises/amazon/amazon-automatise-des-recrutements-mais-le-logiciel-excluait-les-femmes-de-certains-postes-6011655>
- *Intelligence Artificielle* : Qui est responsable ?
<https://www.murielle-cahen.com/publications/intelligence-artificielle.asp>

II – Pourtant, on pourrait espérer une neutralité car les algorithmes proposent des résultats de manière automatique et mathématique

- *Les algorithmes dans le recrutement et le risque discriminatoire*, Ministère de la transition écologique
https://www.cedip.developpement-durable.gouv.fr/IMG/pdf/les_algorithmes_dans_le_recrutement_et_le_risque_discriminatoire.pdf
- *Vers une robotique du traduire*, *Journal of data Mining and Digital Humanities*, CNRS, INRAE, INRIA
<https://jdmdh.episciences.org/>
- *Intelligence Artificielle* : Qui est responsable ?
<https://www.murielle-cahen.com/publications/intelligence-artificielle.asp>
- *Comment l'intelligence artificielle révolutionne le secteur de la santé ?* (2023, 13 février). *France Inter*.
<https://www.radiofrance.fr/franceinter/podcasts/grand-bien-vous-fasse/grand-bien-vous-fasse-du-lundi-13-fevrier-2023-4203390>

III – Vers une utilisation responsable des algorithmes et une amélioration de leur équité

- *L'IA de A à Z : L'humain dans la boucle*. (s. d.). L'IA de A à Z.
<https://atozofai.withgoogle.com/intl/fr-BE/human-in-the-loop/>
- *What-If Tool*. (s. d.). <https://pair-code.github.io/what-if-tool/>
- Moll, G. (2019, 8 novembre). *Apprentissage automatique et biais : Impacts et solutions* - IBM-France. IBM-France.
<https://www.ibm.com/blogs/ibm-france/2019/09/26/apprentissage-automatique-et-biais/>
- *Une approche européenne de l'intelligence artificielle*. (s. d.). Bâtir l'avenir numérique de l'Europe. <https://digital-strategy.ec.europa.eu/fr/policies/european-approach-artificial-intelligence>