Perspective éthique en intelligence artificielle : décoder les biais discriminatoires dans les décisions algorithmiques

Sandrine Charbonneau, *Université Laval*

RÉSUMÉ : Les développements d'algorithmes d'intelligence artificielle (IA) sont porteurs de nombreux bénéfices et probablement d'autant de risques pour la collectivité. Nous croyons qu'ils méritent d'être étudiés et encadrés par une perspective éthique, afin de ne pas reproduire ni renforcer les inégalités et les injustices sociales et économiques déjà présentes. Notre intention sera d'éclairer les risques d'effets discriminatoires dans les décisions algorithmiques en expliquant leur fonctionnement et les sources possibles de leurs biais. Nous verrons qu'un usage de ces technologies fait sans préoccupation pour les membres historiquement plus vulnérables et marginalisés de la société peut rapidement mener à des cas très graves de discrimination. Nous soulignerons qu'il est difficile, mais pas impossible de minimiser ces biais discriminatoires dans les décisions des algorithmes. Au final, nous défendrons que même si l'IA peut nous sembler complexe et opaque dans son fonctionnement, il n'en tient qu'à nous de placer les limites de son utilisation en fonction des valeurs que nous voulons promouvoir.

Technology is neither good nor bad; nor is it neutral. - Melvin Kranzberg

Introduction

Dans la dernière décennie, l'intelligence artificielle (IA) a connu des avancées extrêmement rapides, alors que des algorithmes toujours plus sophistiqués ont été conçus afin d'optimiser nos prises de décisions dans de multiples domaines[1]. On peut notamment observer des applications communes de l'IA dans notre quotidien par la présence de publicités en ligne, dont les contenus sont personnalisés en fonction de nos préférences. Dans bien des cas cependant, les choix effectués grâce à ces programmes sont moins connus, mais peuvent avoir des impacts majeurs dans nos existences. Pensons par exemple à des autorisations de prêts bancaires, des offres d'embauches, ou encore à la durée des peines d'emprisonnement[2]. S'il est complexe de formuler une définition précise et consensuelle de ce qu'est l'IA, cette technologie implique néanmoins les capacités suivantes : « correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation[3] ».

Divers biais[4] peuvent toutefois se glisser à tout moment dans le processus de création algorithmique, entre autres chez les programmeurs et programmeuses, ou encore dans les bases de données qui sont utilisées pour le codage. L'ampleur des impacts que pourrait avoir cette technologie sur nos sociétés et les incertitudes qui persistent quant à notre connaissance de celle-ci devraient nous inciter à adopter une analyse éthique de ses enjeux, afin de minimiser les risques et les dommages que l'IA peut causer aux individus. À cet égard, la programmeuse informatique et activiste Joy Buolamwini décrit notre empressement face aux développements de ces algorithmes en disant : « We have entered the age of automation overconfident, yet underprepared. If we fail to make ethical and inclusive artificial intelligence we risk losing gains made in civil rights and gender equity under the guise of machine neutrality[5] ».

Suivant les propos de Buolamwini, nous discuterons dans cet article des effets potentiellement discriminatoires de la logique prédictive des algorithmes décisionnels d'IA d'apprentissage artificiel. Notre objectif sera de démontrer comment divers biais peuvent se glisser dans les algorithmes d'IA lors de leur programmation

et d'expliquer les effets potentiellement discriminatoires qui peuvent résulter de leur utilisation. Nous soutiendrons que ces technologies peuvent apporter des bénéfices à la société, à condition que celle-ci établisse des balises éthiques qui garantissent que l'utilisation de ces outils « intelligents » ne contribue pas à renforcer les inégalités et les injustices déjà présentes, ce qui n'est pas si aisé à mettre en place. Pour l'expliquer, nous décrirons en premier lieu brièvement le fonctionnement de ce type d'algorithme et relèverons certains de ses apports pour la société, en mettant aussi en garde contre plusieurs dommages qu'il pourrait causer. En second lieu, nous présenterons les types de dommages qui peuvent être causés par des algorithmes biaisés et d'où ces biais discriminatoires peuvent provenir. En troisième lieu, nous mentionnerons deux obstacles majeurs que représente la lutte aux biais, d'abord en ce qui concerne l'aspect technique et ensuite, l'aspect politico-économique. En dernier lieu, nous proposerons une piste de solution afin de réduire la possibilité de biais dans les décisions algorithmiques, passant par des procédures d'encadrement éthique et d'audit.

1. Qu'est-ce que l'IA ?

1.1. Algorithmes d'apprentissage artificiel : prédire et corréler

À l'origine, les approches classiques « symboliques » de l'IA fonctionnaient avec des algorithmes postulant des règles logiques lors de situations précises à partir d'un jeu fini de données d'entraînement[6]. Pensons par exemple à un programme capable de battre à chaque fois les humains aux échecs par sa capacité supérieure de calculer les meilleurs coups possibles. Les progrès les plus récents en IA tiennent surtout du développement de techniques d'apprentissage artificiel avancé sans supervision, tel que l'apprentissage profond (*deep learning*). Ces types d'algorithmes peuvent alors effectuer des corrélations dans des données qui leur permettent d'effectuer des jugements probabilistes plus poussés. Résumons de manière très simplifiée le fonctionnement du modèle de *deep learning* : à partir d'une grande quantité de données, souvent qualifiées de données massives ou de big data (qui peuvent être notamment des chiffres ou des images), ce type d'algorithme d'IA est programmé pour pouvoir

s'entraîner à repérer lui-même des motifs (*patterns*) à travers ces mêmes données et peut, par la suite, lorsque de nouvelles données sont entrées, prédire des résultats. Par exemple, à partir d'une grande base de données d'entraînement comportant des images de chats et de chiens, on peut apprendre à l'algorithme à départager lui-même les différentes espèces. Lorsqu'on lui présenterait une nouvelle image d'un de ces animaux, il pourrait indiquer s'il s'agit de l'une ou de l'autre des espèces (ou aucune des deux), et ce, avec un certain pourcentage d'erreur[7]. Notons qu'il existe de multiples autres modes d'apprentissage artificiel, mais que nous discuterons dans notre article de ceux qui ont une logique probabiliste et corrélationniste, dont les implications en matière d'enjeux éthiques sont similaires.

Ce mode de fonctionnement largement prédictif et corrélationniste est donc ce qui rend l'IA aussi intéressante, mais tout aussi risquée. Comme l'expliquent les sociologues Dominique Cardon et Bilel Benbouzid, « les machines prédictives prétendent calculer les phénomènes sociaux sans s'appuyer sur les attributs catégoriels qui servent ordinairement à enregistrer les acteurs et leurs actions[8] » : puisqu'on laisse l'algorithme « apprendre » par lui-même, il peut effectuer des corrélations que nous n'aurions jamais pu effectuer nous-mêmes, faute de puissance de calcul, mais qui peuvent aussi se révéler erronées à notre sens.

Présentons d'abord les bénéfices que cette technologie représente, pour ensuite exprimer certains des risques qu'elle comporte.

1.2. Les bénéfices

Bien qu'une grande partie des bénéfices entourant les applications de l'IA reste à être prouvée - pour ne nommer qu'un exemple : l'automatisation complète des véhicules qui pourrait entraîner une diminution considérable des accidents sur la route[9]-, plusieurs avantages se font largement sentir. Les résultats les plus spectaculaires de l'utilisation de l'IA se font surtout remarquer dans le domaine de la santé, où certains algorithmes permettent de détecter des maladies avec plus de précision que les meilleurs médecins dans le domaine[10]. L'amélioration des soins de santé due à l'IA est constatée par cette plus grande rapidité et précision des diagnostics,

tout comme par la découverte de nouveaux traitements et par la possibilité d'effectuer un meilleur suivi chez les patients et patientes par l'utilisation de dossiers électroniques[11].

On peut également parler de la croissance économique des entreprises créant et utilisant des technologies d'IA. Depuis les dernières années, celles-ci permettent de générer des milliards de dollars en décuplant la productivité des compagnies, tout en promettant des hausses importantes de profits d'ici les dix prochaines années[12]. Toutes ces entreprises qui utilisent l'IA peuvent optimiser leur efficacité, que ce soit au sein de leur organisation interne, ou encore dans leur offre de produits et de services. L'IA permet aussi une meilleure analyse des opérations financières, aide à la détecter les fraudes[13], facilite la gestion des transports et du secteur agricole, tout comme la prédiction de la météo et de catastrophes naturelles[14].

L'IA promet donc de révolutionner une majorité de secteurs de la société en augmentant l'efficacité et la précision de multiples tâches.

1.3. Risques

Plusieurs individus et organisations ont toutefois commencé à s'intéresser aux principes et aux valeurs qui devraient accompagner notre utilisation et notre développement de l'IA, telle que la *Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*, publiée en 2018[15]. Parmi les principes qui ont été énoncés dans ce document, on retrouve notamment un principe d'équité des êtres humains par rapport aux décisions des algorithmes. Pour démontrer l'importance de telles considérations, prenons le cas d'un algorithme ayant entraîné des effets inattendus. Dans les dernières années, une équipe avait conçu un algorithme programmé pour différencier les chiens de race Husky des loups : les gens qui l'ont codé ont réalisé que l'algorithme se basait plutôt sur le paysage pour distinguer les images de chiens et de loups, ces derniers étant le plus souvent dans la neige. En enlevant les paysages, l'algorithme ne faisait plus la différence entre les deux[16]. Si ce cas peut sembler amusant, d'autres peuvent bien vite devenir profondément discriminatoires.

Joy Buolamwini, qui a la peau noire, a testé plusieurs logiciels de reconnaissance faciale commerciaux fonctionnant avec des algorithmes d'IA entraînés par apprentissage artificiel. Elle a remarqué que ceux-ci ne détectaient pas son visage. Ces logiciels détectaient cependant les visages de ses collègues au teint plus pâle et ont même détecté la présence de son visage lorsqu'elle a mis sur celui-ci un masque blanc[17]. Ces programmes, dans les meilleurs cas, parvenaient généralement à identifier les visages d'hommes blancs avec des taux d'erreurs de quelques pourcentages et d'une dizaine de pourcentages d'erreurs pour les visages de femmes blanches. Dans les pires cas, ils arrivaient à des taux d'erreurs de près de 50 % pour les visages de femmes Noires. En d'autres termes, plus le teint du visage étudié était foncé et plus les traits de ce visage tendaient vers des caractéristiques féminines, plus le taux d'erreurs était élevé. Notons que depuis la publication de cette étude de Buolamwini et de ses collègues, ces compagnies ont révisé leurs programmes et fait diminuer grandement leurs taux d'erreurs (sans pour autant atteindre un degré de précision aussi élevé pour chaque genre et teint de peau)[18].

Le problème avec ces programmes tient surtout du fait qu'ils sont en vente libre, prêts à être utilisés de multiples façons, comme à des fins de sécurité et de vérification de l'identité[19]. Les entreprises ou organismes qui les achètent peuvent alors commettre et reproduire des actions discriminatoires. Si les systèmes d'ouverture de portes d'un bâtiment étaient dotés de l'un de ces programmes, bien des gens pourraient être incapables d'y entrer. Nous verrons dans la prochaine section à quel point ce genre de technologie, avec ses résultats biaisés, peut avoir des effets bien plus préjudiciables sur les individus que ce genre d'agacement.

2. Biais discriminatoires : effets concrets et sources multiples

Nous présenterons dans cette section deux types de dommages, soit d'allocation de ressources et d'opportunités, puis de représentation. Nous poursuivrons avec deux sources de biais liés aux IA, venant d'une part des gens qui programment et de l'autre, des bases de données qui les composent. Le souci de coder

des algorithmes précis et efficaces peut réduire les caractéristiques sociales des humains à de simples données, au détriment du droit de chaque personne à un traitement juste et équitable.

2.1. Types et étendue des dommages

Les dommages pouvant découler des résultats biaisés des algorithmes peuvent être divisés en deux catégories selon Kate Crawford, chercheuse en informatique sur les impacts sociaux des IA : des dommages d'allocation et de représentation[20]. Les problèmes d'allocation sont plus directs, ils influencent par exemple l'octroi de ressources (comme les prêts bancaires et hypothèques), d'opportunités (comme les offres d'embauche et places dans une université) et de services (comme la livraison dans certains secteurs). Ces dommages transactionnels sont plus faciles à quantifier et sont le résultat d'une décision à un temps précis. Les dommages de représentation touchent quant à eux les attitudes et les croyances, reflétant les diverses représentations que des individus peuvent avoir de la société au niveau culturel. Ces derniers sont le plus souvent ignorés en IA, étant plus difficiles à formaliser. Dans cette catégorie, on a par exemple détecté des stéréotypes selon le genre dans les outils de traduction et des outils de reconnaissance faciale ayant des difficultés à identifier les visages de gens non caucasiens[21] : pensons à Google qui avait développé une application de photo ayant étiqueté deux personnes Noires comme étant des singes[22].

De plus, comme le fait remarquer Crawford, les modèles d'algorithmes d'IA fonctionnant par apprentissage artificiel peuvent commettre des erreurs se répandant très vite et à grande échelle. Une même erreur de biais pourrait par exemple toucher jusqu'à un ou deux milliards d'utilisateurs et utilisatrices par jours[23] (pensons simplement à la quantité de gens utilisant des services comme Facebook et Google qui peuvent être affectés si les algorithmes de ces derniers comportent des biais).

2.2. Sources des biais

2.2.1. Les humains et la programmation :

Nombreuses sont les occasions où des algorithmes d'IA ont pris des décisions injustes en discriminant des gens, notamment en fonction de leur genre, de leur race, ou de diverses conditions socio-économiques. Les biais menant à ces discriminations peuvent venir consciemment ou inconsciemment des gens qui programment par l'entretien de préjugés sur divers groupes sociaux, ou par un manque de connaissances et de souci envers des réalités marginalisées. Ces biais s'introduisent alors dans les algorithmes lors des choix de programmation, notamment par les paramètres qui sont sélectionnés pour les configurer.

Cathy O'Neil, une docteure en mathématiques qui travaille dans le domaine des algorithmes, a par exemple déjà interrogé un docteur en statistique qui codait des algorithmes calculant les risques de récidives pour des prisons d'État aux États-Unis. Elle lui a demandé s'il utilisait la « race » comme critère pour coder le programme, ce qu'il a nié. Il a cependant affirmé utiliser les codes postaux, puisqu'ils offrent beaucoup plus de « précision » dans les résultats. Le problème est toutefois que les codes postaux peuvent être de bons indicateurs par « proxy[24] » de la « race » et de la situation économique. L'algorithme, en calculant les risques de récidives, pouvait effectuer des corrélations entre les codes postaux de gens vivant dans des milieux plus défavorisés - dans ce cas, une majorité de gens Noirs - leur assignant des indices de récidives systématiquement plus élevés. O'Neil remarque que plusieurs scientifiques travaillant avec des données massives se voient plutôt comme des techniciens, qui doivent suivre leurs livres et leurs définitions d'optimisation, sans penser aux plus grandes conséquences de leur travail sur la vie et les droits des gens. Selon la mathématicienne, ce raisonnement est le paradigme de la situation actuelle, où certaines valeurs d'efficacité sont plus importantes chez bien des programmeurs et programmeuses que les concepts d'égalité et d'équité[25].

S'il est difficile de se faire un portrait exact de la communauté travaillant à concevoir des IA quant à leur niveau de sensibilisation

par rapport aux impacts sociaux qu'ont leurs algorithmes, plusieurs chercheurs et chercheuses comme O'Neil constatent qu'il reste d'importants efforts de conscientisation sociale à effectuer.

2.2.2. Les humains et les données

Le manque de sensibilisation des humains influençant leurs choix de programmation n'est pas le seul élément en cause dans la présence de biais dans les IA. Celles-ci sont conçues pour apprendre et fonctionner avec les données qui sont à leur disposition : « Ce sont les données, la variété et la qualité de celles-ci qui rendent l'algorithme capable d'un meilleur discernement. Des données peu nombreuses ou relatant des pratiques discriminatoires peuvent reproduire des biais ou en créer, par exemple, en faisant des corrélations entre des éléments qui ne devraient pas être liés[26] ». Un exemple flagrant de discrimination en lien avec l'exemple précédent d'O'Neil a été rapporté par des journalistes de ProPublica, soit celui de l'outil COMPAS, utilisé aux États-Unis pour prédire le risque de récidive des gens accusés de crimes. Dans ce dossier, les journalistes donnent l'exemple de deux crimes similaires, l'un commis par une personne blanche et l'autre par une personne Noire. L'algorithme prédisant leur risque de commettre à nouveau un crime dans le futur donne pourtant un risque plus élevé à la personne Noire, alors qu'elle donne un risque plus faible à la personne blanche. Pour des raisons de choix de critères de calcul dans la programmation, mais aussi et surtout de quantité de données, l'algorithme a « appris » de l'historique des crimes et des sentences passés, reproduisant alors dans ses prédictions les biais racistes des arrestations et des jugements criminels aux États-Unis envers la population Noire[27]. Les scores rapportés par ces outils prédictifs ne sont d'ailleurs pas censés permettre aux juges de donner des sentences plus sévères, mais seulement d'orienter leur jugement. Dans les faits, plusieurs ont cité les résultats de l'algorithme pour justifier leurs décisions[28].

Crawford critique l'approche qu'elle nomme *data fundamentalism*, où corrélation est associée à causalité et où les données massives sont toujours perçues comme offrant des vérités objectives. Selon ses travaux, les données sont souvent prises comme un reflet précis du

monde social, alors qu'il y a toujours des distorsions dans la collecte de données. Celles-ci ne sont jamais « neutres » ou « impartiales », mais peuvent exclure et diviser. Crawford explique que les sciences travaillant avec des Big Data doivent se demander : d'où proviennent leurs données, quelles ont été les méthodes employées pour les analyser et quels sont les biais possibles lors de leur interprétation[29] ? La vigilance est de mise par rapport à la création et l'utilisation de bases de données non représentatives et discriminatoires. Certaines peuvent receler des historiques de pratiques injustes, comme un projet d'algorithme de recrutement d'Amazon, qui désavantageait les candidatures des femmes à l'embauche. Heureusement, il a été décidé que cet algorithme ne serait pas utilisé[30]. Il est important de critiquer les données qui seront utilisées pour entraîner l'IA et de comprendre d'où elles viennent, afin de prévoir quelles discriminations elles pourraient engendrer.

3. Obstacles pour minimiser les biais

Nous montrerons en quoi les algorithmes d'IA posent de sérieux enjeux éthiques d'explicabilité lors de leur utilisation, par un obstacle technique et un obstacle politico-économique. D'une part, la complexité des IA fonctionnant par apprentissage artificiel (et plus particulièrement par apprentissage profond) rend leurs résultats très opaques. De l'autre, les algorithmes développés par les entreprises sont actuellement protégés par les droits commerciaux de propriétés intellectuelles et de propriété privée, ce qui rend souvent impossible la possibilité d'examiner et de critiquer leurs décisions.

3.1. Boîte noire : manque de transparence et d'explicabilité des décisions

Le problème de la boîte noire (*black box*) est celui où on peut observer les entrées (*inputs*) et les sorties (*outputs*) dans un programme informatique, mais sans bien comprendre comment les uns ont pu mener aux autres[31]. Comme l'expliquent les auteurs du rapport Villani de 2018, des algorithmes d'IA fonctionnant par apprentissage profond déploient tant de calculs et de paramètres qu'il est « presque impossible de suivre le cheminement de l'algorithme

de classement[32] », ce qui peut rendre leurs résultats peu ou pas explicables. Ci-dessous, voici une schématisation de ce à quoi peut ressembler un « arbre décisionnel » informatique très simple, représentant quelques couches de calculs des « réseaux neuronaux artificiels » de l'algorithme[33].

Avec des milliers, voire des millions de variables et d'entrées, il peut être pratiquement impossible de retrouver les origines des biais discriminatoires dans ces technologies. Même en testant l'algorithme, on ne pourrait pas savoir, par exemple, si les discriminations viennent plutôt de la programmation, des données d'entraînement, ou même des deux.

Le rapport Villani donne l'exemple d'algorithmes de Google de ciblage publicitaire, qui ont eu tendance à proposer aux femmes des offres d'emplois à plus faible rémunération que celles qu'ils proposent aux hommes, ou encore d'algorithmes qui ont tenté de prédire les secteurs les plus susceptibles de criminalité aux États-Unis, ayant pour résultat l'augmentation de la surveillance dans des quartiers pauvres à prédominance afro-américaine[34]. Ces cas ayant été constatés, on pourrait dans le premier établir que ces biais viennent, par exemple, uniquement de l'historique des données utilisées, avec des statistiques montrant des salaires moins élevés pour les femmes. Toutefois, on pourrait aussi penser que ces traitements discriminatoires proviennent des préjugés entretenus par les gens ayant programmé les algorithmes, qui auraient pu inconsciemment proposer certains types d'emplois moins bien rémunérés aux femmes et non aux hommes. Le problème reste donc la difficulté d'établir

d'où proviennent les biais, en raison de l'opacité des algorithmes. En examinant les codes qui les composent, on ne saurait départager avec précision quelles données et quels choix ont pu donner tels résultats.

Bien que la plupart des pays où ces algorithmes sont utilisés possèdent des lois pour protéger les populations des discriminations, on ne peut établir avec certitude le poids qu'ont les prédictions des IA dans la prise de décisions aux effets discriminatoires. Comme dans le cas de l'outil COMPAS mentionné plus haut, les juges ne devraient pas se baser uniquement sur les algorithmes pour déterminer les sentences, mais ils et elles le font parfois. À plus long terme, si l'utilisation d'algorithmes d'aide à la décision continue à se répandre, sans être plus transparents dans leurs résultats, nous pourrions décider de ne plus leur confier certaines tâches pour éviter le maintien et la reproduction d'injustices[35].

3.2. Entreprises privées : maximisation des profits et de l'efficacité

Un obstacle supplémentaire se dresse à la possibilité de comprendre les décisions des algorithmes d'IA et de critiquer leurs biais le cas échéant : le fonctionnement de ces algorithmes est considéré comme un secret d'affaires, les formules qui les composent sont donc protégées par les droits de propriété intellectuelle et privée, empêchant qui que ce soit d'analyser en détail les étapes de leurs calculs[36]. Dans bien des cas, le souci de générer plus de profits passe outre le respect des humains pouvant être affectés par les indications des algorithmes. Plusieurs compagnies cachent délibérément les résultats de leurs modèles, avec des hordes d'avocats et de lobbys pour les défendre, comme Google, Amazon et Facebook, dont les algorithmes seuls valent des centaines de millions de dollars[37].

Les domaines de la finance et des assurances sont particulièrement reconnus pour avoir massivement recours à des algorithmes afin d'évaluer leur clientèle actuelle ou future. Une énorme quantité de données est récoltée sur ces gens, avec leur consentement ou à leur insu[38]. Comme dans le cas des programmes d'IA prédisant les risques de récidives chez les criminels et criminelles, les algorithmes

de finance et d'assurance affectent les scores de fiabilité des individus en fonction d'informations directement pertinentes (par exemple, le respect du Code de la route) et d'autres par « proxy », dont les liens de causalité sont très discutables (comme le code postal résidentiel du conducteur ou de la conductrice, ou encore les régions visitées en voiture, suivies par géolocalisation)[39]. De nombreuses personnes ont critiqué ces scores opaques, qualifiés par exemple d'« arbitrary, and discriminatory[40] », mais sans obtenir de garanties de ces compagnies ou des gouvernements que la situation soit étudiée.

Cette réalité concrète illustre bien que malgré la conscientisation montante dans le domaine de l'IA sur les impacts sociaux des algorithmes d'apprentissage artificiel, il semble manquer de mesures concrètes de contrôle et d'évaluation de ces logiciels quant aux impacts de leur utilisation.

4. Pistes de développement et d'encadrement

Faute de pouvoir développer longuement sur les multiples façons dont pourraient être encadrés les algorithmes d'IA, nous ne ferons qu'esquisser une piste de solution aux problèmes précédemment énoncés. Nous avancerons que les biais indésirables se retrouvant dans des programmes informatiques peuvent être limités par une réglementation très serrée des IA, impliquant des procédures d'audit, où sont testés les algorithmes avant et après leur mise en marche. L'objectif serait de s'assurer qu'à chaque étape, nous puissions avoir une compréhension suffisante de leurs résultats et que nous nous assurions que ceux-ci ne soient pas discriminatoires. Notre conclusion sera que si les algorithmes démontrent des biais dans leurs décisions, ils devraient être révisés ou sinon, interdits.

4.1. Audits et réglementation

Face aux défis techniques et juridiques que posent l'opacité et le secret d'affaires des algorithmes d'IA, nous proposons d'adopter des réglementations exigeant plus de transparence et un suivi serré des compagnies programmant et utilisant ces algorithmes. Nous demandons pour ce faire que ces derniers puissent être sujets à des examens et à des audits avant et après leur mise en application. Il

va sans dire que nous contestons les lois empêchant l'accès aux codes des algorithmes pour des raisons de propriété privée ou intellectuelle. Nous proposons la mise en place de comités d'analyse et de révisions indépendants, soumis à des clauses de confidentialité. Par exemple, un groupe pourrait être composé d'experts et d'expertes en informatique et de spécialistes en éthique qui n'ont pas de conflits d'intérêts avec les compagnies développant et utilisant l'algorithme étudié.

Nos suggestions, à l'instar des recommandations du rapport AI Now 2017, se concentrent principalement sur les tâches des IA reliées aux organismes publics dont les décisions peuvent avoir des impacts majeurs sur la vie des gens, par exemple dans les domaines de la justice criminelle, de la santé, de l'aide sociale et de l'éducation[41]. Comme nous l'avons mentionné, la complexité des algorithmes d'IA d'apprentissage automatique rend très difficile, voire impossible la compréhension claire de leurs résultats. Nous ne nous attendons donc pas à une explicabilité complète du programme dans les résultats fournis, mais nous tenons à la possibilité d'interroger ses bases de données, pour mieux comprendre le fonctionnement de l'algorithme.

De plus, avant d'utiliser des données pour entraîner un modèle, il faut s'assurer de pouvoir comprendre d'où viennent ces données et ce qu'elles représentent, en s'interrogeant notamment sur les méthodes employées pour les collecter. Comme le soutient Crawford, il faut se tourner vers les sciences sociales pour cette partie du travail et aller plus loin que de demander « combien » aux gens, mais aussi « pourquoi » et « comment »[42]. Il est important que cette étape d'analyse des données soit effectuée avant d'entraîner des programmes pouvant rendre des décisions lourdes de conséquences. Une fois que les algorithmes auront été testés exhaustivement et que leur utilisation ne semblera pas mener à des conclusions injustement discriminatoires[43], il faudrait s'assurer que leurs résultats soient les plus transparents que possible pour ceux et celles qui les interprètent.

Revenons à l'exemple d'Amazon, où l'entreprise a décidé d'elle-même de ne pas utiliser son algorithme pour les embauches, puisqu'il défavorisait les candidatures des femmes[44]. Dans une autre situation,

il a été révélé par Bloomberg que cette même compagnie n'offrait pas son service de livraison en 24 heures dans certains quartiers à majorité afro-américaine aux États-Unis, les algorithmes utilisant les codes postaux pour optimiser les possibilités de profits. À la suite de la publication de l'article sur cette réalité discriminatoire, les maires de New York, Boston et Chicago ont critiqué Amazon, qui a rapidement décidé d'offrir le service à ces endroits précédemment exclus. Cependant, aucune réglementation ne les obligeait à le faire et aucune loi ne les empêchait concrètement de concevoir et d'utiliser ces algorithmes discriminatoires[45].

Selon nous, dans des cas plus délicats (comme pour juger des sentences de prison), les machines ne devraient pas être autorisées à rendre une décision automatique ou une suggestion sans autorisation ou examen humain. Nous privilégierons toujours des approches plus complètes avec un jugement humain, même si celles-ci s'avèrent plus coûteuses en temps, puisqu'un gain en efficacité pourrait se solder par de graves préjudices. Si l'algorithme, après avoir été analysé et testé sous les conseils venant de disciplines aux points de vue variés, ne nous permet pas d'affirmer que ses résultats sont exempts de biais causant de la discrimination, il ne devrait pas avoir le droit d'être utilisé.

5. Conclusion

Dans cet article, nous avons tenté d'expliquer le fonctionnement de certains types d'IA, ainsi que les bénéfices potentiels de son utilisation, tout comme certains risques qu'elle comporte pour la société. Nous avons voulu démontrer de quelle manière des biais aux effets discriminatoires peuvent être introduits dans les algorithmes lors de la programmation, par les préjugés ou le manque de sensibilité de ceux et celles qui programment envers certains groupes de personnes, ou encore dans leur choix des données utilisées. Nous avons décrit quelques-uns des nombreux obstacles techniques et politiques se dressant contre les tentatives d'amenuiser ou d'annuler les effets discriminatoires des biais de l'IA. Finalement, nous avons lancé une piste de réflexion sur les possibilités d'encadrer

la conception et les usages des algorithmes d'IA, par des procédures d'audit et de réglementation.

Nous sommes encore aux débuts du développement de ces nouvelles technologies, au moment où les lois n'ont pas encore balisé ni encadré l'ensemble de leur fonctionnement. Comme l'avancent la chercheuse en « gouvernementalité algorithmique » Antoinette Rouvroy et d'autres , les algorithmes d'IA commencent à s'étendre dans toutes les sphères de notre quotidien et sont appelés à jouer des rôles de plus en plus grands au sein de la gouvernance même de la société[46]. Les impacts de ces outils de décision basés sur des prédictions devraient être réfléchis sous des points de vue inclusifs et non discriminatoires. Il nous semble le respect les droits de chaque être humain à être traité justement et équitablement devraient guider de façon prioritaire nos choix futurs. Nous pensons que malgré le fait que les jugements humains puissent être défaillants, il faut trancher sur les cas où ils restent préférables aux décisions des machines, puisqu'il est plus facile de comprendre et de critiquer des raisonnements humains dans des cas de litiges.

Nous croyons tout de même que l'essor du développement en IA peut se faire de manière responsable. Il comporte de nombreuses opportunités d'améliorer notre qualité de vie, malgré les risques de reproduction et de maintien de discriminations. Sur le plan des bénéfices possibles, nous pouvons aussi mentionner la possibilité d'adopter des algorithmes moins biaisés que les humains sur certains plans. Des juges peuvent avoir tendance à accorder des peines d'emprisonnement inéquitables en raison de biais racistes, mais aussi pour des raisons aussi banales que l'heure à laquelle ils et elles rendent leur sentence - une étude ayant démontré que les peines étaient plus sévères avant la pause dîner en raison de leur faim[47]. Si la création d'algorithmes purement « neutres » ou sans biais est impossible, on peut toutefois tenter d'en programmer qui soient moins partiaux et plus justes dans leurs décisions.

---

1. Nous nous concentrerons sur les enjeux touchant l'intelligence artificielle « simple » ou « faible », soit les programmes informatiques d'algorithmes

de calcul actuels et non l'intelligence artificielle « complexe » qui pourrait être autant, sinon plus intelligente que l'humain, comportant selon plusieurs des risques existentiels. Pour plus de détails sur cette dernière, voir par exemple : Nick Bostrom et Eliezer Yudkowsky, « The ethics of artificial intelligence », dans Keith Frankish *et al.* [dir.], *The Cambridge Handbook of Artificial Intelligence*, Cambridge, Cambridge University Press, 2014, p. 316-334.

2. Notons que l'IA n'est pas systématiquement utilisée dans les situations mentionnées en exemple. Il peut être difficile, voire impossible pour le public de savoir si les décisions rendues par des systèmes informatiques résultent de l'utilisation de l'IA ou non. C'est toutefois une pratique en croissance dans la plupart des entreprises : Louis Columbus, « 10 Charts That Will Change Your Perspective On Artificial Intelligence's Growth », dans *Forbes,* [En ligne], https://www.forbes.com/sites/louiscolumbus/2018/01/12/10-charts-that-will-change-your-perspective-on-artificial-intelligences-growth/ (Page consultée le 6 juillet 2019).
3. Andreas Kaplan et Michael Haenlein, « Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence », dans *Business Horizons*, vol. 62, no^{o} 1, janvier 2019, p. 15.
4. Nous emploierons le terme « biais » dans le sens large de « biais cognitif », incluant toute erreur de raisonnement (que ce soit dans la théorie ou pratique, de façon consciente ou inconsciente). Nous mettrons l'accent sur les biais qui poussent vers la partialité et les préjugés : Marie van Loon, « Biais cognitifs (version académique) » dans *M. Kristanek* [dir.], L'Encyclopédie Philosophique, 2018.
5. Maureen McElaney, « Cognitive Bias in Machine Learning », The Data Lab [En ligne], https://medium.com/codait/cognitive-bias-in-machine-learning-d287838eeb4b (Page consultée le 9 juillet 2019).
6. Jocelyn Maclure et Marie-Noëlle Saint-Pierre, « Le nouvel âge de l'intelligence artificielle : une synthèse des enjeux éthiques », dans *Les cahiers de propriété intellectuelle*, vol. 30, no^{o} 3, octobre 2018, p. 748.
7. SAS Institute, « Deep Learning: What it is and why it matters » [En ligne], https://www.sas.com/en_us/insights/analytics/machine-learning.html (Page consultée le 6 juillet 2019).
8. Bilel Benbouzid et Dominique Cardon, « Machines à prédire », dans *Réseaux*, no^{o} 5, 2018, p. 28.

9. Waymo, « Waymo Safety Report. On The Road To Fully Self-Driving » [En ligne], https://waymo.com/safety/ (Page consultée le 11 juillet 2019).
10. Martin Stumpe et Craig Mermel, « Improved Grading of Prostate Cancer Using Deep Learning », Google AI Blog [En ligne], https://ai.googleblog.com/2018/11/improved-grading-of-prostate-cancer.html (Page consultée le 6 juillet 2019).
11. Lisa Morgan, « Artificial Intelligence in Healthcare: How AI Shapes Medecine », *Datamation* [En ligne], https://www.datamation.com/artificial-intelligence/artificial-intelligence-in-healthcare.html (Page consultée le 11 juillet 2019).
12. Louis Columbus, *op.cit*.
13. Santana Wilson, « 3 Ways AI is Used in Business Process Optimization », Oracle Data Science [En ligne], https://www.datascience.com/blog/ai-for-business-process-optimization, (Page consultée le 13 juillet 2019).
14. Magnimind Academy, « Invaluables Societal Benefits of AI », Medium [En ligne], https://becominghuman.ai/invaluable-societal-benefits-of-ai-2ed62f7a653f (Page consultée le 12 juillet 2019).
15. Université de Montréal, « Déclaration de Montréal pour un développement responsable de l'intelligence artificielle » [En ligne], https://www.declarationmontreal-iaresponsable.com/ (Page consultée le 6 juillet 2019).
16. Marco Tulio Ribeiro *et al*., « "Why Should I Trust You?" : Explaining the Predictions of Any Classifier », dans *arXiv:1602.04938* [cs, stat], février 2016, p. 9.
17. Joy Buolamwini et Timnit Gebru, « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification », dans J*ournal of Machine Learning Research,* vol. 81, février 2018.
18. Deborah Raji Inioluwa et Joy Buolamwini, « Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products », dans *Conference on Artificial* Intelligence, Ethics, and Society, 2019.
19. Face++, « Face Detection » [En ligne], https://www.faceplusplus.com/face-detection/ (Page consultée le 13 juillet 2019).
20. Il existe bien sûr d'autres manières de les traiter, celle de Crawford nous apparaissant la plus complète au niveau des biais ayant des effets discriminatoires dans les décisions prises par des algorithmes.
21. Kate Crawford, « The Trouble with Bias », Conférence prononcée au NIPS 2017, Longbeach, CA, le 5 décembre 2017, 17:00 [En ligne],

https://www.youtube.com/watch?v=fMym_BKWQzk, (Page consultée le 30 novembre 2018).
22. Jana Kasperkevic, « Google says sorry for racist auto-tag in photo app », *The Guardian* [En ligne], https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app, (Page consultée le 12 juillet 2019).
23. Kate Crawford, *op. cit.*, 5:30.
24. La discrimination par « proxy » est une forme de discrimination par facteurs indirects : « We formalize a notion of proxy discrimination in data-driven systems, a class of properties indicative of bias, as the presence of protected class correlates that have causal influence on the system's output » : Anupam Datta *et al.*, « Proxy Discrimination in Data-Driven Systems Theory and Experiments with Machine Learnt Programs », dans *arXiv:1707.08120v1*, juillet 2017, p. 1.
25. Tom Upchurch, « To work for society, data scientists need a hippocratic oath with teeth », *Wired UK* [En ligne], https://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o-neil-weapons-of-math-destruction (Page consultée le 18 décembre 2018).
26. Jocelyn Maclure et Marie-Noëlle Saint-Pierre, *op. cit.*, p. 756.
27. Aliya Saperstein, *et al.* « The Criminal Justice System and the Racialization of Perceptions », dans T*he ANNALS of the American Academy of Political and Social Science*, vol. 651, no^{o} 1, janvier 2014, p. 104-121.
28. Julia Angwin *et al.*, « Machine Bias », ProPublica [En ligne], https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Page consultée le 23 novembre 2018).
29. Kate Crawford, « The Hidden Biases in Big Data », Harvard Business Review [En ligne], https://hbr.org/2013/04/the-hidden-biases-in-big-data (Page consultée le 23 novembre 2018).
30. Cathy O'Neil, « Amazon's Gender-Biased Algorithm Is Not Alone », Bloomberg [En ligne], https://www.bloomberg.com/opinion/articles/2018-10-16/amazon-s-gender-biased-algorithm-is-not-alone (Page consultée le 18 décembre 2018).
31. Frank Pasquale, *The Black Box Society*: *The Secret Algorithms That Control Money and Information*, Cambridge, Harvard University Press, 2015, p. 3.
32. Cédric Villani, *et al*., « Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne », France, mars 2018, p. 142 [En ligne], https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf (Page consultée le 18 décembre 2018).

33. Jonathan Shaw, « Artificial Intelligence and Ethics », Harvard Magazine [En ligne], https://harvardmagazine.com/2019/01/artificial-intelligence-limitations (Page consultée le 12 juillet 2019).
34. Cédric Villani, *op.cit.*
35. *Ibid.*
36. Kate Crawford, « Artificial Intelligence's White Guy Problem », *The New York Times* [En ligne], https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (Page consultée le 24 septembre 2019).
37. Cathy O'Neil, Weapons of math destruction : *how big data increases inequality and threatens democracy*, New York, Crown, 2016, p. 29.
38. Autre enjeu éthique majeur à propos duquel nous ne pourrons développer.
39. Cathy O'Neil, *op. cit.* p. 164-171.
40. Frank Pasquale, *op. cit.*, p. 23.
41. Alex Campolo, *et al.*, The AI Now Report : T*he Social and Economic Implications of Artificial Intelligence,* AI Now Institute, 2017, p. 1.
42. Kate Crawford, « The Hidden Biases in Big Data », *op. cit*.
43. Nous n'avons pas l'espace nécessaire pour expliquer de quelles manières nous pensons que les tests devraient être conduits. Nous croyons, à l'instar de Cathy O'Neil, qu'une suggestion intéressante pour conduire des audits sur les boîtes noires de programmes d'IA serait de les tester avec des cas variés, afin de s'assurer qu'ils ne mènent pas systématiquement à des résultats discriminatoires. (Tom Upchurch, *op. cit.*).
44. Cathy O'Neil, « Amazon's Gender-Biased Algorithm Is Not Alone », *op. cit*.
45. David Ingold et Spencer Soper, « Amazon Doesn't Consider the Race of Its Customers. Should It? », *Bloomberg* [En ligne] http://www.bloomberg.com/graphics/2016-amazon-same-day/, (Page consultée le 23 novembre 2018).
46. Marc-Olivier Bherer, « En 2018, résistez aux algorithmes avec la philosophe Antoinette Rouvroy », *Le Monde* [En ligne], https://www.lemonde.fr/idees/article/2017/12/29/en-2018-resistez-aux-algorithmes-avec-la-philosophe-antoinette-rouvroy_5235555_3232.html (Page consultée le 20 décembre 2018).
47. Shai Danziger *et al.*, « Extraneous factors in judicial decisions », *Proceedings of the National Academy of Sciences,* vol. 108, no° 17, avril 2011, p. 6889-6892.