# Algorithmes : biais, discrimination et équité[1]

Bertail Patrice[2], Bounie David[3], Clémençon Stephan[4], et Waelbroeck Patrick[5]

14 février 2019

Résumé

Les algorithmes s'immiscent de plus en plus dans notre quotidien à l'image des algorithmes d'aide à la décision (algorithme de recommandation ou de scoring), ou bien des algorithmes autonomes embarqués dans des machines intelligentes (véhicules autonomes). Déployés dans de nombreux secteurs et industries pour leur efficacité, leurs résultats sont de plus en plus discutés et contestés. En particulier, ils sont accusés d'être des boites noires et de conduire à des pratiques discriminatoires liées au genre ou à l'origine ethnique. L'objectif de cet article est de décrire les biais liés aux algorithmes et d'esquisser des pistes pour y remédier. Nous nous intéressons en particulier aux résultats des algorithmes en rapport avec des objectifs d'équité, et à leurs conséquences en termes de discrimination. Trois questions motivent cet article : Par quels mécanismes les biais des algorithmes peuvent-ils se produire ? Peut-on les éviter ? Et, enfin, peut-on les corriger ou bien les limiter ? Dans une première partie, nous décrivons comment fonctionne un algorithme d'apprentissage statistique. Dans une deuxième partie nous nous intéressons à l'origine de ces biais qui peuvent être de nature cognitive, statistique ou économique. Dans une troisième partie, nous présentons quelques approches statistiques ou algorithmiques prometteuses qui permettent de corriger les biais. Nous concluons l'article en discutant des principaux enjeux de société soulevés par les algorithmes d'apprentissage statistique tels que l'interprétabilité, l'explicabilité, la transparence, et la responsabilité.

[1] Ce travail a été réalisé avec le soutien de la Fondation Abeona et s'inscrit également dans le cadre du programme de recherche de la Chaire Finance Digitale et de son axe sur l'Intelligence artificielle dans la banque conduit en partenariat avec La Banque Postale.
[2] Université Paris Nanterre, MODAL X, UFR SEGMI, 200, ave de la République, 92 001 Nanterre.
[3] i3-SES, Télécom ParisTech, CNRS (UMR 9217), 46 rue Barrault, 75634 Paris Cedex 13.
[4] LTCI, Télécom ParisTech, 46 rue Barrault, 75634 Paris Cedex 13.
[5] i3-SES, Télécom ParisTech, CNRS (UMR 9217), 46 rue Barrault, 75634 Paris Cedex 13.

## 1. Les algorithmes en question

Les algorithmes régulent nos vies. Qu'il s'agisse d'outils d'aide à la décision pour le choix d'un parcours universitaire (Parcoursup), de recommandations de vidéos (Netflix), de recherches d'information en ligne (moteur de recherche) ou de logiciels intégrés aux véhicules, les algorithmes s'immiscent de plus en plus dans notre quotidien.

Un algorithme est un ensemble de règles et d'instructions écrites en vue d'obtenir un résultat. Une recette de cuisine ou bien une partition musicale sont des algorithmes. Ils décrivent un ensemble d'étapes et de règles en vue d'atteindre un résultat, la production d'un plat ou bien une mélodie. Un algorithme peut s'insérer dans un programme informatique et être interprété par des langages de programmation. Par exemple, en consultant des fiches de livre sur le site Amazon.fr, un algorithme peut proposer aux internautes des conseils d'achat liés à d'autres livres (algorithme de recommandation).

De nombreux types d'algorithme existent, au service de finalités diverses telles que le tri, le classement, et la prévision. Leur développement s'est considérablement accéléré ces dernières années au gré des opportunités offertes par l'apprentissage statistique[6], l'optimisation[7] et le calcul distribué[8], en bref ce qu'on appelle l'apprentissage machine (machine learning), la branche de l'intelligence artificielle[9] fondée sur l'analyse automatique des données massives[10]. Les algorithmes permettent avec une certaine efficacité d'automatiser des traitements répétitifs à grande échelle, de trier des centaines de milliers de résultats, ou bien de formuler des recommandations après avoir examiné des millions de décisions passées.

Mais au-delà de leur efficacité, les algorithmes sont supposés disposer d'autres qualités telles que la neutralité ou encore l'équité. La neutralité d'un algorithme fait référence à sa capacité à restituer une représentation fidèle de la réalité. Mais de quelle restitution est-il question ? L'objectif d'un algorithme étant par nature de trier, classer ou ordonner les informations selon certains principes, ce concept de neutralité est difficile à appréhender, et on peut lui préférer celui de loyauté. La loyauté d'un algorithme implique que les résultats produits soient conformes aux attentes des utilisateurs/consommateurs. En d'autres termes, le classement, le tri ou la sélection des informations doivent être loyaux vis-à-vis des intérêts des personnes qu'ils sont censés servir. L'équité d'un algorithme va un cran plus loin. Les résultats ne doivent pas opérer de distinction entre les personnes en fonction d'attributs protégés par la loi telle que l'origine ethnique, le genre, ou encore la situation de famille[11]. Dit autrement, le classement, le tri ou la sélection des informations doivent être équitables entre les personnes sur la base d'attributs protégés par la loi.

[6] Domaine à l'interface des mathématiques et de l'informatique visant à développer des algorithmes pour l'analyse automatique des masses de données, ils permettent à une machine d'« apprendre » à réaliser des tâches de façon autonome avec une efficacité contrôlée, on parle d' « apprentissage machine » (machine learning). Ces algorithmes sont à la base du fonctionnement de solutions d'intelligence artificielle telles que la reconnaissance de parole, la vision par ordinateur ou les moteurs de recommandation de contenu.

[7] Branche des mathématiques dont l'objet est de développer des méthodes pour trouver ou approcher les valeurs pour lesquelles une fonction donnée est maximale ou minimale.

[8] Un calcul est distribué lorsqu'il n'est pas réalisé de façon centralisée sur une seule unité mais réparti sur plusieurs processeurs. La volumétrie des informations à traiter, sa dispersion physique (réseaux de capteurs) ou bien encore des contraintes de sécurité ou de rapidité peuvent par exemple motiver son usage.

[9] L'intelligence artificielle est un corpus de concepts et de techniques permettant à une machine de réaliser des tâches au moyen de programmes informatiques, simulant parfois ainsi, dans une certaine mesure, l'intelligence humaine.

[10] On parle de données massives (« Big Data ») ou megadonnées lorsque la volumétrie des données est telle qu'elle rend inopérant les outils standard de gestion des bases de données et de traitement de l'information. Les données du Web, les données décrivant les mesures réalisées par les technologies d'analyse moderne telles que la spectrométrie de masse ou la résonance magnétique nucléaire en sont des exemples.

[11] L'infraction de discrimination, i.e. toute distinction opérée entre les personnes physiques sur le fondement de leur origine, de leur sexe, de leur situation de famille, etc., est sanctionnée par le code pénal (voir les articles 225-1 à 225-4).

Mais l'équité des algorithmes existe-t-elle ? En première analyse, la mise en œuvre d'un algorithme correspond à l'application de règles et d'instructions qui devraient permettre de s'affranchir de la partialité qui caractérise parfois les décisions humaines. Cependant, les algorithmes n'échappent pas à cet écueil. Pour reprendre une expression de la mathématicienne Cathy O'Neil (2016), un algorithme n'est en réalité qu'une « opinion intégrée aux programmes », et nous savons que les opinions peuvent conduire à des pratiques discriminatoires à l'encontre des personnes.

L'actualité quotidienne atteste de nombreux cas de pratiques discriminatoires liées à des algorithmes, volontaires ou involontaires, à l'encontre de certaines populations. Aux Etats-Unis par exemple, des travaux ont mis en évidence que les populations afro-américaines étaient plus souvent pénalisées par les décisions de justice qui s'appuient sur le recours aux algorithmes (Angwin et al., 2016). Ces mêmes populations sont également plus discriminées sur les plateformes populaires de locations d'appartement en ligne (Edelman, Luca and Svirsky, 2017). Enfin, des publicités ciblées et automatisées en ligne relatives aux opportunités d'emploi dans les domaines des sciences, de la technologie, de l'ingénierie et des mathématiques aux Etats-Unis seraient plus fréquemment proposées aux hommes qu'aux femmes (Lambrecht and Tucker, 2017).

Les algorithmes ne seraient donc pas si équitables que cela et, à l'instar des décisions prises par des personnes humaines[12], pourraient conduire à des pratiques discriminatoires[13]. Les biais des algorithmes pourraient être alors définis comme une déviation par rapport à un résultat censé être neutre, loyal ou encore équitable.

Cet article décrit les biais liés aux algorithmes, et esquisse des pistes pour y remédier. Nous nous intéressons en particulier aux résultats des algorithmes en rapport avec des objectifs d'équité, et à leurs conséquences en termes de discrimination. Trois questions motivent cet article. Par quels mécanismes les biais des algorithmes peuvent-ils se produire ? Peut-on les éviter ? Et, enfin, peut-on les corriger ou bien les limiter ? Les réponses à ces questions ne sont pas triviales et renvoient plus généralement à des questions de société qui traversent la communauté académique mais également, en France, des autorités administratives indépendantes comme la Commission Nationale Informatique et Libertés, des commissions consultatives (Conseil national du numérique), ainsi que des associations civiles et des fondations. Ces questions sont multiples et portent successivement sur la transparence, l'auditabilité, l'explicabilité, l'interprétabilité, et la responsabilité des algorithmes. Ces enjeux seront également abordés dans cet article.

Cet article est structuré en quatre parties. Dans une deuxième partie, nous décrivons comment fonctionne un algorithme d'apprentissage statistique. Dans une troisième partie, nous nous intéressons à l'origine des biais des algorithmes qui peuvent être de nature cognitive, statistique ou économique. Dans une quatrième partie, nous présentons quelques approches statistiques ou algorithmiques prometteuses qui permettent de corriger ou limiter les biais. Nous concluons enfin l'article avec les principaux enjeux de société soulevés par les algorithmes d'apprentissage statistique.

[12] De nombreux travaux de recherche psychologie, à l'image des travaux pionniers de Khaneman et Tversky (1974), montrent que le raisonnement humain peut être biaisé en raison de raccourcis mentaux, de biais de stéréotype, ou encore d'erreurs de calcul ou d'appréciation.
[13] De nombreux cas de discrimination à l'embauche ont été par exemple mis en évidence en France ces dernières années, et Emmanuel Macron, dans sa présentation des mesures pour les banlieues en mai 2018, s'est engagé à mettre en place des tests anti-discriminations dans les grandes entreprises françaises.

## 2. Comment fonctionne un algorithme d'apprentissage statistique ?

Un algorithme est une suite d'opérations ou d'instructions permettant d'obtenir un résultat. Les algorithmes sont au cœur de nombreuses disciplines. Un domaine nous intéresse en particulier en raison de son développement significatif au cours de ces dernières années : l'apprentissage machine (*machine learning*).

Les algorithmes de machine learning permettent à une machine d'apprendre à partir d'exemples réels. Ils cherchent à découvrir la structure d'un ensemble de données à partir de valeurs observées. Ils sont utilisés dans de nombreux contextes comme la classification et la prévision et sont au cœur des algorithmes d'aide à la décision (moteur de recherche, etc.). Nous emploierons donc les algorithmes utilisés dans le contexte du machine learning pour illustrer les problèmes de biais des algorithmes.

Comment fonctionne un algorithme de machine learning ?

Les paradigmes du machine learning, initiés avec les travaux précurseurs de V. Vapnik, ont été élaborés en grande partie il y a plus d'un demi-siècle, bien avant les progrès réalisés dans le domaine de la collecte et du stockage de données, et la mise au point d'infrastructures de calcul distribué. Les principes du machine learning peuvent être décrits à travers les principes de la reconnaissance de formes.

Dans ce type de problème, l'algorithme doit accomplir la tâche suivante : à partir de données d'entrée X, il doit reconnaître automatiquement la catégorie Y (d'un type donné, spécifié à l'avance) associée à l'objet/individu décrit par X, avec un risque d'erreur minimal.

De très nombreuses applications correspondent à cette formulation, de la biométrie au diagnostic/pronostic médical assisté en passant par la gestion du risque de crédit. Dans le cas de la vision par ordinateur par exemple, X correspondra à une image pixélisée et la sortie Y à une 'étiquette' associée à l'image indiquant la présence éventuelle d'un objet spécifique dans celle-ci. Ce problème est de nature prédictive dans la mesure où la règle de décision déterminée par un algorithme d'apprentissage opérant sur une base de données étiquetées (i.e. les « données d'apprentissage ») ne doit pas seulement pouvoir prédire le passé, mais permettre de prédire efficacement, lorsqu'elle sera déployée, le label Y associé à de nouvelles données d'entrée X, non encore observées. On dira le cas échéant que la règle prédictive a alors de « bonnes capacités de généralisation ».

Le langage des probabilités et des statistiques, particulièrement adapté pour raisonner en univers incertain et décrire la variabilité des données, permet de formaliser ce problème. L'idée très simple est que le risque d'erreur que l'on cherche à minimiser au moyen du procédé d'apprentissage est la probabilité de prédire de façon erronée l'étiquette Y relative à une observation X présentée aléatoirement à la machine, et issue de la même population statistique que les exemples ayant servi à l'apprentissage.

Deux problèmes se posent. Premièrement, la probabilité de prédire une étiquette est la plupart du temps inconnue, la complexité de la distribution de la paire (X,Y) « entrée-sortie » échappant à la modélisation, et la base de données d'apprentissage étant souvent loin de couvrir l'univers de tous les possibles (i.e. contenir toutes les paires (X,Y) pour lesquelles la machine aura à effectuer une prédiction dans le futur). Le deuxième problème est lié à ce que l'on entend exactement par « apprentissage » par la machine d'une règle de décision. L'apprentissage consiste simplement en la mise en œuvre d'un programme d'optimisation visant à minimiser une version statistique (empirique) du risque d'erreur calculée à partir de la base de données d'entraînement (la fréquence des erreurs commises sur les exemples d'apprentissage dans le cas le plus simple) et opérant sur une classe de règles donnée.

Au-delà des avancées significatives réalisées récemment dans la mise en œuvre des techniques d'optimisation, permettant en particulier d'accroître leur rapidité ou leur capacité de « passage à l'échelle », les travaux théoriques de V. Vapnik offrent un cadre de validité aux techniques de minimisation du risque empirique. En substance, ils garantissent de bonnes capacités de généralisation à la règle prédictive issue de la procédure d'apprentissage. Toutefois, deux conditions sont nécessaires : d'une part, la classe de règles sur laquelle est opérée l'optimisation doit être d'une complexité contrôlée, tout en étant suffisamment riche pour contenir des règles s'ajustant bien aux données (X,Y), et d'autre part, le nombre d'exemples d'apprentissage présentés à la machine doit être suffisamment grand pour que le risque d'erreur théorique puisse être approché par sa version statistique.

Dès la fin des années 70, ces concepts fondamentaux ainsi que certaines approches algorithmiques telles que les réseaux de neurones sont documentés dans la littérature scientifique sous une forme quasi-achevée. Ce n'est qu'avec le Big Data cependant que le machine learning a commencé à rencontrer un succès grandissant. Cette longue phase de lancement s'explique en partie par la rareté de l'information numérisée alors disponible. En effet, l'apprentissage statistique s'effectuait généralement à partir de données issues de questionnaires très coûteux de tailles insuffisantes, engendrant une erreur statistique parfois considérable. Par ailleurs, les capacités de mémoire et de calcul étant alors très limitées, les programmes d'optimisation pouvant être mis en œuvre à l'époque opéraient sur des classes de règles trop frustes pour réaliser un apprentissage efficace.

Les briques technologiques, dont certaines sont à l'origine du développement du web (Hadoop, MapReduce, etc.), ont en effet engendré des progrès considérables dans le domaine de la collecte, du stockage et du traitement des données. Les avancées réalisées dans la gestion de la mémoire et dans le domaine du calcul parallélisé permettent la mise en œuvre de programmes d'apprentissage opérant sur des classes très flexibles, telles que les réseaux de neurones profonds (deep learning), susceptibles, pour de nombreux problèmes, de rendre compte très efficacement de la façon dont l'information X en entrée permet de prédire la sortie Y.

Les mégadonnées du web, les immenses bibliothèques d'images, de sons ou de textes « étiquetés » auxquelles il permet d'accéder, entraînent ainsi les moteurs de reconnaissance de contenu avec d'innombrables exemples. Le Big Data permet alors en quelque sorte aux approches statistiques d'accéder au « paradis asymptotique » promis par la loi des grands nombres. Mais ces avancées ne doivent pas faire oublier le fait qu'une procédure d'apprentissage statistique, aussi automatisable qu'elle soit, n'est valide que dans un cadre spécifique. Sa validité repose fortement sur les hypothèses faites sur les mécanismes aléatoires inhérents à l'observation des données (mécanismes parfois difficiles à contrôler) et le catalogue des règles de décision jugées potentiellement performantes utilisé : sa simplicité ne doit donc pas inciter certains à jouer aux apprentis sorciers.

Ainsi, pour reprendre l'exemple précédent, la capacité de généralisation d'une règle produite par un algorithme standard de machine learning pour la reconnaissance de formes n'est assurée que dans les situations où les données sur lesquelles elle est appliquée suivent la même loi de probabilité que les données d'entraînement utilisées lors de l'étape d'apprentissage. Si par exemple, les images de la base d'entraînement sur lesquelles apparaissent des chats, pour reprendre l'un des exemples jouet les plus populaires en vision par ordinateur, présentent toutes un fond vert, on s'attendra naturellement à ce que la machine apprenne à reconnaître la présence d'un fond vert plutôt que celle d'un chat.

Mais au-delà de cet exemple aussi saisissant que simpliste, il convient de souligner que les Big Data, et les données du web en particulier, ne sont généralement pas obtenues à partir d'un plan d'expérience ou de sondage défini à l'avance. Certaines applications se contentent d'exploiter

a posteriori l'information qui a pu être collectée sans toute la rigueur scientifique nécessaire. Cette absence de contrôle sur le processus d'acquisition des données peut naturellement compromettre la découverte de régularités statistiques enfouies dans les masses de données. On se souviendra des sérieuses déconvenues essuyées par certaines applications relatives à la prédiction des épidémies ou du trafic routier.

## 3. Petit tour d'horizon des biais

Les algorithmes d'apprentissage statistique ont donc besoin de données pour produire des résultats, et la qualité des données détermine la qualité des résultats. Se posent alors de nombreuses questions : que se passe-t-il si les données sont erronées ou faussées ? Que se passe-t-il si les algorithmes entraînés sur des données de citoyens américains sont utilisées pour prédire des comportements de citoyens européens ? Que se passe-t-il si le contexte auquel est appliqué les données évolue dans le temps ? Pour répondre à ces questions, nous proposons un tour d'horizon de quelques biais que nous classons en trois catégories : les biais cognitifs, statistiques et économiques.

### 3.1 Les biais cognitifs

Les résultats des algorithmes dépendent de la manière dont les programmeurs les ont écrits. Or ces derniers restent des êtres humains, et de nombreuses recherches en psychologie et sciences cognitives montrent l'existence de biais cognitifs dans la prise de décision (Khaneman et Tversky, 1974). Ces biais cognitifs peuvent conduire à des biais dans les algorithmes.

Les biais cognitifs sont une distorsion de la manière dont l'information est traitée par rapport à un comportement rationnel ou à la réalité. Par exemple, le biais de « bandwagon » ou « du mouton de Parnurge » peut conduire le programmeur à suivre des modélisations qui sont populaires sans s'assurer de leur exactitude. Les biais d'anticipation et de confirmation peuvent conduire le programmeur à favoriser sa vision du monde même si des données disponibles peuvent remettre en question cette vision. Le biais de « corrélations illusoires » peut également conduire une personne à déceler des corrélations entre deux événements indépendants. Tous ces biais peuvent induire des choix de variables qui traduisent plus une certaine perception des phénomènes et qui vont guider l'algorithme vers des décisions biaisées. Les travaux très polémiques de Kosinski et Wang (2018) qui ont proposé de détecter l'orientation sexuelle d'un individu au faciès (soulevant des problèmes éthiques évidents) sont sans doute plus révélateurs de leurs propres perceptions que d'une quelconque réalité.

Un autre type de biais connu est le biais de stéréotype. Ce dernier peut survenir lorsqu'un individu agit en référence au groupe social auquel il s'identifie plutôt que sur ses capacités individuelles. De nombreuses recherches ont alors montré que la performance individuelle peut décroître lorsqu'un individu pense être jugé ou sélectionné sur la base de stéréotypes négatifs (Block et al. 2011). Ces biais peuvent apparaître par exemple dans les offres d'emploi en ligne où les femmes s'auto-sélectionnent, et répondent à des offres dont elles pensent qu'elles auront une probabilité plus élevée d'être acceptée. En retour, les algorithmes qui se nourrissent des données de *clicks* renforcent ces menaces de stéréotypes.

De tels biais de stéréotype ont été mis en évidence également dans les travaux sur l'incorporation automatique de mots en machine learning (word embedding). Les moteurs de recherche proposent par exemple des associations automatiques de mots sur la base de nombreuses occurrences rencontrées dans les textes sur Internet. Ces co-occurrences qui existent très largement dans la vie réelle se généralisent sur Internet. Bolukbasi et al. (2016) ont ainsi montré que les biais de stéréotype se retrouvent dans l'analyse textuelle des associations

de mots liées aux métiers et au genre, le mot « femme » étant systématiquement associé à des termes tels que « femme de ménage », « nourrice », « réceptionniste », « styliste », ou bien « coiffeuse », tandis que le mot « homme » est associé à « chef », « philosophe », « capitaine », « financier ».

### 3.2 Les biais statistiques

Nous regroupons dans cette partie des biais liés aux données ou des biais plus statistiques comme le biais de variable omise, le biais de sélection, ou encore le biais d'endogénéité.

- **Le biais des données : « Garbage in, garbage out » (GIGO)**

L'acronyme GIGO « Garbage in, garbage out » que l'on pourrait traduire en français par « Foutaises en entrée, Foutaises en sorties » fait référence au fait que, même l'algorithme le plus sophistiqué qui soit, produira des résultats inexacts et potentiellement biaisés si les données d'entrée sur lesquelles il s'entraîne sont inexactes. Le grand danger est donc qu'un algorithme produise des résultats qui semblent bons et utiles, alors qu'ils ne le sont pas. Après tout, il est facile de croire à un score produit par un algorithme propriétaire complexe et qui semble être basé sur des sources multiples. Mais si les données sous-jacentes sont entachées de biais cognitifs, sont inexactes, approximatives, biaisées en entrée (et c'est souvent le cas), le résultat final sera inévitablement biaisé, sans un traitement adapté de l'information et/ou la prise en compte des biais reconnus en entrée. Ce phénomène est renforcé par les méthodes d'apprentissage qui s'auto-alimentent des données créées par l'algorithme.

Un exemple frappant est l'algorithme mis en place à partir de 2015 par Amazon pour faciliter le recrutement de talents.[14] L'algorithme utilisait des données de centaines de milliers de curriculum vitae (CV) reçus par Amazon au cours des dix dernières années en vue de noter de nouvelles candidatures. L'algorithme attribuait une note allant de 1 à 5 étoiles, à l'image de l'évaluation de produits sur Amazon. L'utilisation de l'algorithme a été rapidement suspendue en raison de son incapacité à sélectionner les meilleurs candidats et sa propension à discriminer les femmes. En effet, l'algorithme attribuait fréquemment de mauvaises notes à des profils qualifiés de femme et en adéquation avec les postes à pourvoir, et proposait systématiquement également des candidats sous-qualifiés pour toutes sortes de postes très variés. Il attribuait de mauvaises notes en particulier aux CV contenant les mots « women's » comme dans « women's chess club captain ». Il défavorisait également les diplômées de collèges américains uniquement réservés aux femmes. L'algorithme favorisait à l'inverse des mots tels que « executed » ou « captured » plus présents dans les CV masculins. Dans ce cas particulier, les données en entrée étaient complètement déséquilibrées entre homme et femme, les hommes constituant l'écrasante majorité des cadres recrutés dans le passé, l'algorithme ne laissant du coup aucune chance aux nouvelles candidates pourtant qualifiées.

Ce biais des données est très fréquent également dans les algorithmes de détection faciale (voir l'algorithme de détections de criminels au faciès en Chine (Wu et Zhang, 2016) qui détecte essentiellement les individus qui sourient versus ceux qui font la tête parce qu'ils sont en prison...), mais aussi en médecine ou en génétique, où l'apprentissage est effectué sur des données spécifiques, sur des populations de type « caucasiennes » (Buolamwini et Gebru, 2018).

- Le biais de variable omise

Les algorithmes d'apprentissage produisent des résultats sur la base de modèles qui exploitent de nombreuses variables. Toutes les variables ne sont toutefois pas toujours disponibles pour produire les résultats. Par exemple, certaines compétences humaines dites tacites sont difficiles

[14] Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 10 octobre 2018.

codifier et à incorporer dans des algorithmes de recrutement. Il s'agit de capacités telles que le leardership, l'initiative entrepreneuriale, le travail en équipe ou encore l'intelligence émotionnelle. Si ces capacités ne sont pas prises en compte mais qu'elles sont négativement corrélées aux résultats scolaires, l'algorithme risque de pénaliser sur le marché du travail les personnes avec des résultats scolaires moyens, mais qui demeurent indispensables pour monter une équipe travaillant sur un projet innovant.

L'omission de certaines variables peut donc compromettre la fiabilité des résultats. Žliobaitė and Custers (2016) montrent par exemple que l'omission d'une caractéristique sensible (telle que l'origine ethnique) d'un algorithme peut renforcer la discrimination. Cette analyse est également partagée par Kleinberg et al. (2017a, 2017b) qui montrent que la stratégie consistant à rendre aveugle les algorithmes peut porter atteinte à l'équité. Ce problème est devenu particulièrement prégnant dans le contexte de l'application du règlement général sur la protection des données en Europe.[15] Comment lutter contre la discrimination à l'égard des femmes ou des minorités sexuelles, par exemple, s'il est impossible de collecter des données sur le genre ?

- **Le biais de sélection**

Le biais de sélection peut fortement influencer les résultats des algorithmes. Ce biais apparaît lorsque les caractéristiques de la population étudiée sont différentes de celles de la population générale. James Heckman, dans ses travaux lui ayant valu le « prix Nobel d'économie », a montré que le biais de sélection était de la même nature que le biais de variable omise.

L'exemple du calcul d'un score de crédit (credit scoring) dans le cas de l'attribution d'un crédit bancaire est frappant. Pour déterminer la catégorie de risque de l'emprunteur, les algorithmes calculent un score en se basant sur les personnes qui ont été éligibles à un emprunt dans un établissement particulier (celui qui établit des scores de crédit en interne). L'algorithme utilise alors seulement les informations disponibles sur une partie de la population et ignore les dossiers de toutes les personnes à qui les banques ont refusé un prêt, celles qui n'ont jamais eu besoin d'emprunter, celles qui ont fini de rembourser leurs emprunts, et enfin celles qui ont des emprunts dans d'autres établissements. En ignorant les autres populations qui ont des caractéristiques différentes de la population étudiée, l'algorithme peut fournir des résultats biaisés.

En outre, des données pour un individu peuvent être manquantes. Ceci soulève également un problème de biais de sélection si les observations disponibles n'ont pas les mêmes propriétés que les observations manquantes. Ce problème est fréquent même à l'heure du Big Data, car il est rare d'avoir des observations répétées sur l'ensemble des individus de la population étudiée.

Le déploiement du machine learning peut lui-même être à l'origine de biais statistiques significatifs liés au fait que l'on n'observe pas l'ensemble de la distribution de probabilité générant les données. Ces biais sont appelés biais de censure et de troncature et sont très proches des biais de sélection. Les applications à la maintenance prédictive par exemple visent à éviter les pannes d'un système en permettant le remplacement de composants, dont le fonctionnement est contrôlé en temps quasi-réel au moyen de capteurs, avant leur défaillance. Il faut alors voir que la mise en œuvre généralisée de telles solutions, si elle permet d'assurer la continuité du système auquel elles s'appliquent, induira une forte sous-estimation des durées de vie. Le caractère massif de l'information ne peut à lui seul y remédier. D'une manière générale, le

[15] Le règlement général sur la protection des données (GDPR) régit le traitement par une personne, une entreprise ou une organisation de données à caractère personnel relatives à des personnes dans l'Union européenne. Le règlement est entré en vigueur le 24 mai 2016 et est applicable depuis le 25 mai 2018. Il a pour objectif de renforcer les droits fondamentaux des particuliers à l'ère numérique et de faciliter les affaires en clarifiant les règles applicables aux entreprises et aux organismes publics dans le marché unique numérique (UE). La Charte des droits fondamentaux stipule que les citoyens de l'UE ont le droit de protéger leurs données à caractère personnel.

machine learning s'incarne aujourd'hui de plus en plus souvent dans des systèmes intelligents qui interagissent avec l'environnement, lequel 'produit' en retour les données que nous observons, et le système construit ainsi son propre plan d'expériences au cours du temps afin d'optimiser sa performance. On parle alors d'apprentissage « par renforcement ». Les systèmes de recommandation par exemple exploitent les données relatives à l'historique courant d'un utilisateur de manière à lui proposer les objets/contenus qu'il est le plus susceptible de consommer. Cette tâche « d'exploitation des données passées » doit naturellement se combiner à une tâche « d'exploration des possibles », permettant de compléter l'information statistique disponible. Or, si la première tâche est souvent perçue comme un levier pour accroître les gains, la seconde est généralement exclusivement associée à un coût. A titre d'exemple, l'apprentissage par renforcement réalisé par les agents conversationnels peut être biaisé lors de la phase exploratoire. Ainsi, l'agent conversationnel Tay lancé par Microsoft a rapidement proféré des insanités à la suite des insultes et des offenses prononcées par des internautes et fut fermé en moins de vingt-quatre heures.

- **Le biais d'endogénéité**

La plupart des algorithmes de machine learning utilisent des données passées pour prédire le futur. Or dans de très nombreuses situations, les anticipations du futur sont plus importantes pour expliquer les événements présents que les comportements et données passés. Ainsi sur les marchés financiers, l'hypothèse de marchés efficients conclut que les prix des actifs financiers doivent fluctuer de manière aléatoire à l'équilibre car les prix contiennent toutes les informations des investisseurs, et il est impossible pour un investisseur particulier d'avoir un avantage informationnel en utilisant des données passées. Les prix reflètent plutôt les anticipations de profits futurs. Dans le cas du credit scoring, il se peut qu'un prospect avec un mauvais historique de remboursement d'emprunt puisse changer de style de vie lorsqu'il décide de fonder une famille. Un être humain peut comprendre ce changement de comportement alors qu'un algorithme ne le peut sans doute pas. Ces deux exemples démontrent l'importance des anticipations pour comprendre les comportements humains. Malheureusement, les algorithmes actuels ne sont pas programmés pour le faire car modéliser les anticipations reste extrêmement difficile.

### 3.3 Les biais économiques

Les algorithmes peuvent être biaisés volontairement ou involontairement pour des raisons économiques, i.e. de prix et de coût. Lambrecht et Tucker (2017) ont étudié, par exemple, comment un algorithme fournissant des annonces publicitaires faisant la promotion d'emplois dans les domaines des sciences, de la technologie, de l'ingénierie et des mathématiques (STEM) peut discriminer les femmes. Les auteurs ont montré qu'un algorithme qui optimise simplement le rapport coût-efficacité de la diffusion d'annonces affiche moins d'annonces destinées aux femmes, car le prix du segment des femmes jeunes est supérieur à celui des hommes jeunes. Il est donc moins couteux pour l'algorithme de servir des publicités aux hommes jeunes qu'aux femmes jeunes. L'efficacité d'un algorithme peut être donc compromise si l'on ne tient pas compte du contexte économique dans lequel il est déployé.

Les biais peuvent être également dus à des manipulations volontaires de la part des entreprises. Ce phénomène est connu sous le nom de « search engine manipulation » (Epstein et Robertson, 2015 et Linskey, 2017). Dans le contexte de l'e-commerce, la commission européenne a par exemple condamné Google en 2017 à une amende de 2,4 milliards d'euros pour avoir favorisé ses propres produits dans les résultats de recherche de Google Shopping au détriment de ses concurrents (IP/17/1784 ; affaire 39740).

## 4. Quelques pistes pour limiter les biais des algorithmes

Nous examinons deux pistes poursuivies par les recherches récentes pour limiter les biais associés aux algorithmes : les pistes statistiques et les pistes algorithmiques.

### 4.1 Les pistes statistiques

Il y a encore quelques décennies, les données étaient essentiellement collectées au travers d'enquêtes réalisées sur des échantillons représentatifs ou non de la population nationale afin d'obtenir des réponses à des questions définies à l'avance. Cela permettait ainsi l'élaboration de plans expérimentaux visant à contrôler certains biais éventuels et renforcer les conclusions pouvant être déduites de l'analyse statistique. Mais la technologie a profondément transformé l'articulation entre acquisition des données et traitement mathématique de l'information, en inversant le processus. Les masses de données numérisées à disposition n'ont généralement pas été acquises au moyen d'un plan d'expérience : on observe simplement les données et on imagine a posteriori les services et les analyses qui pourraient résulter de l'information disponible. Si la technologie nous donne des quantités importantes de données à faible coût, elle nous renseigne souvent beaucoup moins sur les conditions dans lesquelles les données ont été collectées. Il est alors difficile de proposer des méthodes systématiques de correction de biais voire même de justifier l'usage de l'inférence statistique pour produire des estimations fiables. Comme nous venons de le voir, les biais ne s'évanouissent pas nécessairement sous le simple effet de la grande masse de données à disposition sur internet.

Dans cette partie, nous envisageons quelques pistes statistiques liées au traitement de données manquantes, au redressement et à l'échantillonnage, et enfin à l'inclusion de variables dites auxiliaires.

**▪ Compléter l'information**

La première piste statistique pour limiter les biais des algorithmes est sans doute la plus simple à expliquer. Lorsque certaines données (par exemple les caractéristiques des individus figurant dans la base de données) sont omises ou censurées, il s'agit de les reconstituer par l'intermédiaire d'un modèle statistique approprié, ajusté idéalement au moyen de données issues cette fois d'un plan d'expérience contrôlé. L'imputation dite 'hot-deck' reste toutefois très utilisée en pratique : dans ce cas de figure, les valeurs des populations manquantes sont remplacées par la valeur d'individus ou la moyenne de valeurs d'individus ayant des caractéristiques similaires. Cette similarité peut être mesurée par une distance entre caractéristiques ou des modèles de type plus proches voisins. Le choix des caractéristiques n'est pas sans conséquence et l'imputation 'hot-deck' n'est pas exempte d'hypothèses de modélisation. Il est possible de donner des garanties de fonctionnement de cette méthode sous des conditions très restrictives (Chen et Shao, 2000). De manière générale, l'approche qui consiste à imputer des valeurs manquantes a des limites évidentes. L'utilisation d'un modèle pouvant lui même être biaisé par construction, l'omission de variables essentielles, la qualité et la variabilité des observations utilisées pour l'estimation du modèle peuvent aggraver encore les phénomènes de biais, d'autant plus sérieusement que l'imputation est effectuée à grande échelle (Fan, Fang et Han (2014) pour les défis que posent statistiquement les données massives).

**▪ Le redressement de l'algorithme d'apprentissage**

La deuxième piste envisagée repose sur le redressement des données. La théorie des sondages, revisitée récemment dans un cadre mathématique plus adapté aux données massives (Bertail et al. (2017), Clémençon et al. (2017), Boistard et al. (2018)) fournit également des pistes prometteuses pour le redressement des algorithmes d'apprentissage statistique. Il s'agit

essentiellement de comprendre pourquoi certaines classes d'individus sont peu représentées dans la base de données et de modéliser la probabilité qu'un individu aux caractéristiques données y figure, qui est appelée score de « propension » à être observé/sélectionné ou encore « probabilité d'inclusion ». Si le phénomène ou le problème prédictif étudié est indépendant du mécanisme de sélection, les biais peuvent être ignorés. Dans le cas contraire, une correction du biais de sélection est indispensable (voir le principe de non-ignorabilité de Rubin (1987)) et peut s'effectuer en incluant dans l'analyse des variables auxiliaires expliquant la sélection, à condition que le mécanisme de censure soit indépendant du phénomène analysé.

- **Redressement vs. ré-échantillonnage.**

Dans les cas où la complexité du calcul des statistiques en jeu rend difficile l'incorporation des poids des individus, on préférera créer des populations artificielles, ressemblant davantage à la population cible, par des procédures de sous/sur-échantillonnage. Celles-ci permettent de répliquer ou de supprimer certains individus. D'autres techniques consistent à créer des individus virtuels, par exemple par interpolation. Ces méthodes relèvent toutes d'une approche dite de ré-échantillonnage. Mais, si elles sont largement pratiquées du fait de leur simplicité, elles n'ont pas été étudiées sur le plan théorique hors de cadres standards et peuvent conduire à des résultats erronés dans certains cas.

- Information auxiliaire

Il est souvent fréquent, en particulier dans le cas des grandes bases auto-sélectionnées, que les probabilités d'inclusion soient inconnues. On ne peut alors espérer corriger les biais de sélection sans information additionnelle, sans 'variables auxiliaires'. A condition d'en disposer également pour les populations sous-représentées dans la base de données ou de réaliser une enquête supplémentaire qui va reproduire les conditions de constitution de la base sur un échantillon test, l'information auxiliaire peut permettre de comprendre le mécanisme de biais/sélection en jeu : il s'agit d'expliquer pourquoi un individu est présent dans la base ou pas, en fonction de ces caractéristiques additionnelles. On en déduit un score de propension à être inclus dans la base qui est utilisée pour corriger des biais de sélection selon le principe de l'estimation de Horvitz-Thompson sur l'ensemble des observations de la base originale. Ce principe est mis en œuvre dans Borrajo et Cao (2018) dans un cas non-paramétrique simple où un échantillon externe permet de ré-estimer correctement les poids des données de départ. Dans le cas où les variables auxiliaires sont très fortement corrélées avec la structure de la population statistique étudiée, on pourra aussi chercher à re-pondérer les données au moyen de scores de propension ne dépendant que de l'information auxiliaire, c'est le principe de la post-stratification ou du calage.

- Dérive temporelle

Enfin, avec l'ubiquité de capteurs fonctionnant en temps continu, les données sont de plus en plus souvent collectées de façon séquentielle, parfois disponibles sous forme de flux uniquement et doivent parfois être analysées à la volée ('streaming machine-learning'). Mais l'analyse des données sur des fenêtres temporelles de durée trop limitée a ses limites et conduit à ignorer certaines caractéristiques du phénomène étudié telles que des tendances de long terme, des effets saisonniers ou des ruptures. Incorporer dans les méthodes d'apprentissage des modèles temporels décrivant les éventuels mécanismes d'évolution des phénomènes analysés est là encore indispensable.

Certains praticiens ont pu croire de façon erronée que le caractère massif des données disponibles pouvait permettre de s'affranchir de réfléchir aux conditions d'acquisition des données et ignorer ainsi les biais statistiques. La communauté du machine learning est toutefois aujourd'hui sensibilisée à ces problèmes. Si les méthodes de correction de biais (e.g. censure,

sélection) demeurent encore largement méconnues en son sein, il est possible de les intégrer aux algorithmes d'apprentissage statistique comme l'on montré certains travaux précurseurs. La recherche sur ce sujet en est certes encore à ses prémices mais promise à de nombreux développements dans les prochaines années.

### 4.2 Assurer l'équité : les pistes algorithmiques

Au-delà des pistes statistiques, un domaine de recherche en machine learning se développe autour de ce que l'on appelle l'équité algorithmique. Ces travaux ont pour objectif de concevoir des algorithmes qui répondent à des critères d'équité, par exemple la non-discrimination en fonction d'attributs protégés par la loi comme l'origine ethnique, le genre ou l'orientation sexuelle. Ces travaux de recherche remontent à l'article pionnier de Pedreshi et al. (2008) qui propose une méthode pour déterminer une règle de classification dite éthique. Depuis ce travail, le domaine évolue rapidement, principalement en raison de l'existence de cas de discriminations avérées. Il est important de noter que cette recherche est de plus en plus multidisciplinaire et comprend les sciences des données, l'informatique, la statistique mais également les sciences humaines et sociales avec l'économie, le droit et la philosophie, car les questions d'équité et de discrimination sont au cœur des théories sur la justice sociale.

Trois conclusions peuvent être tirées de ces recherches récentes en machine learning : l'équité est un concept éthique et pluriel ; l'équité peut être formalisée par un algorithme ; les algorithmes d'équité sont incompatibles et non universels.

- L'équité est un concept éthique et pluriel.

L'équité est un concept éthique qui a trait à des conceptions plurielles de la justice entre les individus (Kolm, 1998). Ce concept est au centre des travaux en sciences sociales qui s'intéressent en particulier à la répartition des richesses entre les individus, et à l'intervention possible de l'État (justice sociale).

On prête à Aristote la distinction entre deux critères d'équité, à savoir l'équité horizontale et l'équité verticale. L'équité horizontale correspond au principe selon lequel des individus égaux doivent être traités également, indépendamment de leur origine ethnique ou de leur genre par exemple. L'équité verticale requiert que les individus qui sont par essence inégaux soient traités inégalement. Par exemple, les individus qui ne sont pas égaux en termes de revenus doivent contribuer de manière inégale à un régime fiscal ; on peut demander par exemple à ce que les plus riches contribuent plus que les plus pauvres à un impôt.

On devine immédiatement la difficulté : l'équité est un jugement de valeur qui relève de l'éthique, et son application va varier selon les cultures, les systèmes politiques, les besoins à couvrir, etc.

- L'équité peut être intégrée aux algorithmes sous diverses formes de contrainte.

Ces concepts d'équité peuvent être aisément formalisés en utilisant le langage des probabilités et statistiques, et être intégré dans des algorithmes. Mais, étant données l'origine et la multiplicité des domaines académiques, la manière de les formaliser n'est pas harmonisée.

Corbett-Davies et Goel (2018) proposent trois définitions formelles de l'équité, à savoir « anti-classification », « parité de classification » et « calibration ». L'anti-classification fait référence aux algorithmes qui ne prennent pas en compte les attributs protégés dans les méthodes de classification ou de prédiction; la probabilité d'un résultat est égale pour tous les individus, indépendamment de leur appartenance à un groupe. Une autre classe d'algorithmes est appelée parité de classification, ou parité démographique; la probabilité d'un résultat est égale pour tous les individus appartenant à un même groupe. Enfin, la troisième définition est connue sous le

nom de calibration. Elle requiert que les résultats soient indépendants des attributs protégés après un contrôle du risque estimé. Par exemple, parmi les demandeurs de crédit qui ont 10% de chances de ne pas rembourser le crédit demandé, la méthode de calibration requiert que les taux de défaut de paiement soient les mêmes à travers plusieurs groupes.

Hamilton (2016) contraste l'équité individuelle à celle de groupe. L'équité individuelle garantit que des individus similaires obtiennent des résultats similaires (Dwork et al., 2012); autrement dit, les individus sont traités en fonction de leurs propres mérites (égalité de traitement). L'équité de groupe prend en compte des caractéristiques personnelles telles que l'origine ethnique et le genre, et traite les individus différemment en fonction de leurs différences (égalité de résultats) (Pedreshi et al., 2008).

D'un point de vue algorithmique, assurer l'équité revient généralement à intégrer un certain nombre de contraintes dans le programme d'optimisation permettant d'apprendre une règle de décision à partir des données. Dans le domaine de la biométrie par exemple, l'équité pourrait conduire à souhaiter que le taux de faux positifs d'un logiciel de reconnaissance faciale soit comparable pour tous les groupes ethniques par exemple. Mais il faut aussi bien voir que l'intégration d'une telle contrainte dans l'algorithme d'apprentissage pourrait détériorer les capacités prédictives du moteur d'identification appliqué à certains groupes. La difficulté intrinsèque du problème de reconnaissance peut en effet très bien varier selon les groupes sans que les différences de performance ne soient imputables à d'éventuels biais présentés par la base de données d'apprentissage.

- Les règles formelles d'équité sont incompatibles et non universelles

Une implication directe des définitions précédentes est d'une part qu'il est impossible de construire un algorithme universel pour empêcher simultanément toutes les formes de discrimination (genre, origine ethnique, etc.). D'autre part, les définitions formelles de l'équité sont incompatibles entre elles (Berk, 2017, Friedler et al., 2016; Loftus et al., 2018). Un exemple simple peut être donné en utilisant les concepts d'équité individuelle et celle de groupe. Appliquée au cas des admissions dans des collèges par exemple, l'équité de groupe stipulerait que les taux d'admission soient égaux pour des attributs protégés (le genre, etc.), alors que l'équité individuelle exigerait que chaque personne soit évaluée indépendamment de son genre.

Une autre illustration des définitions contradictoires et incompatibles de l'équité est également donnée dans l'exemple de la prévision algorithmique en matière de justice pénale. Aux États-Unis, les juges ont de plus en plus recours à des algorithmes pour évaluer la probabilité de récidive d'une personne inculpée. Un algorithme calcule un score (allant de 1 à 10) pour chaque accusé; ce score est supposé mesurer la probabilité de récidive de l'accusé. Ce score est calculé sur la base de plus de 100 questions posées à l'accusé comme son âge, son genre et ses antécédents criminels. L'origine ethnique n'est pas utilisée par l'algorithme. L'objectif de l'algorithme est alors de classifier le risque de récidive indépendamment de l'origine ethnique de la personne (égalité de traitement). Trois catégories de profils sont construites à partir des scores : personne à risque faible (1 à 4), à risque moyen (5 à 7) et à risque élevé (8 à 10). Les personnes présentant un risque moyen ou élevé de récidive ont une probabilité plus faible d'être remis en liberté.

ProPublica, organisation indépendante à but non lucratif, a évalué l'un des algorithmes de la société Northpoint dans le cadre d'une célèbre analyse intitulée COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). ProPublica a comparé sur l'ensemble d'un comté pendant deux ans, les taux de risque de récidive prédits par l'algorithme à ceux réellement observés lorsque les délinquants étaient remis en liberté (Angwin et al., 2016). L'algorithme prédit correctement le risque de récidive dans 61% des cas (59% des cas pour les Afro-américains et 63% des cas pour les blancs). Mais quand l'algorithme se trompe,

il se trompe plus fréquemment pour les afro-américains que pour les blancs. Les accusés blancs sont souvent prédits moins risqués qu'ils ne le sont : les accusés blancs qui avaient récidivé dans les deux ans avaient été considérés à tort comme à faible risque presque deux fois plus souvent que les récidivistes noirs (48% contre 28%). Les accusés noirs sont souvent prédits plus risqués qu'ils ne le sont ; les accusés noirs qui n'ont pas récidivé dans les 2 ans sont plus fréquemment classés à tort dans la classe risque élevé que les blancs (45% contre 23%). En conclusion, l'algorithme surévalue le risque de récidive des afro-américains et sous-estime ce risque pour les blancs.

## 5. Trois enjeux de société autour des biais des algorithmes

L'acceptation sociale des algorithmes et de l'intelligence artificielle dépendra de la capacité de tous les acteurs, des scientifiques aux décideurs politiques, à répondre aux défis posés par les données, les algorithmes et les pratiques. Plusieurs questions sensibles se posent : faut-il expliquer les résultats des algorithmes ? Faut-il rendre transparents les algorithmes ? Faut-il les auditer ? Qui est responsable du préjudice lié à la discrimination ?

▪ De l'interprétabilité à l'explicabilité des algorithmes

Les progrès récents de l'apprentissage machine et de l'apprentissage profond[16] offrent de nouvelles perspectives pour tout un ensemble de secteurs (santé, éducation, emploi). Mais, les bénéfices de ces progrès s'accompagnent de défis à résoudre. Tout d'abord, si les théories mathématiques sous-jacentes aux modèles utilisés sont bien comprises, il est délicat pour ne pas dire souvent impossible de comprendre le fonctionnement interne de certains modèles. Dit autrement, les règles de décisions extraites à partir des données en étudiant les valeurs des paramètres sont difficilement *interprétables*. C'est le cas bien souvent de certains modèles tels que les machines à vecteurs de support, les forêts aléatoires, les arbres améliorés par gradient, et les algorithmes d'apprentissage profonds tels que les réseaux de neurones artificiels, les réseaux de neurones convolutifs et les réseaux de neurones récurrents.[17]

Le concept d'interprétabilité (ou « explainable AI » (XAI) en anglais), parfois désigné par intelligibilité, est un thème de recherche en informatique en plein essor.[18] Il est en particulier soutenu par un programme ambitieux de l'agence du département de la Défense des États-Unis (DARPA). Les recherches s'orientent sur le développement de méthodes qui aident à mieux comprendre ce que le modèle a appris ainsi que des techniques pour expliquer les prédictions individuelles (Samek et al. 2017).

Ce concept d'interprétabilité est très proche de celui d'explicabilité, qui se réfère compte tenu de la complexité des modèles, à la nécessité d'expliquer aux utilisateurs finaux comment et pourquoi un résultat a été obtenu. L'explicabilité de l'IA est un critère prépondérant par exemple dans le rapport Villani relatif à la « stratégie nationale de recherche en intelligence artificielle ». L'explicabilité a pour objectif d'ouvrir la boite noire de l'IA pour plusieurs raisons. La première est liée à la vérifiabilité des résultats. Un médecin, qui n'est pas un spécialiste d'IA, doit pouvoir comprendre le résultat d'un algorithme et le réfuter le cas échéant si le résultat est la conséquence d'un biais quelconque (une corrélation malheureuse entre variables par exemple). La deuxième est la progression des connaissances. Un résultat nouveau

[16] On parle d'apprentissage profond (deep learning) lorsque l'apprentissage statistique (machine learning) opère sur des classes de modèles (prédictifs ou représentatifs) d'une grande complexité, représentés par exemple par des réseaux de neurones dont l'architecture présente un grand nombre de 'couches'. Ces techniques sont à l'origine des progrès spectaculaires obtenus récemment dans le domaine de l'analyse des signaux audio et des images par exemple.

[17] Ces modèles sont supposés plus difficilement interprétables que des algorithmes de machine learning de type régression linéaire, régression logistique, classifieur bayésien naïf ou arbre de décision. Ces derniers ont toutefois également leurs limites lorsqu'ils mobilisent de nombreuses variables explicatives dans les régressions, que la profondeur des arbres de décision est importante, etc. (Lipton, 2016).

[18] Explainable Artificial Intelligence (XAI).

prédit par un algorithme peut être au contraire le résultat d'une découverte nouvelle pour la science ; la recherche doit donc pouvoir comprendre l'origine de ces découvertes. La troisième est la conformité au droit et à la régulation. Un résultat doit pouvoir être contesté en cas de défaut, de discrimination et autres. Depuis l'entrée en vigueur du Règlement Général sur la Protection des Données (RGPD) en Europe, les personnes ont le « droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire ». Le RGPD stipule également que, pour les décisions automatisées basées sur des données à caractère personnel, les personnes ont le droit de « demander une explication de la décision [algorithmique] prise à l'issue de cette évaluation et de contester la décision. » Enfin, la quatrième et dernière est la confiance. Une technologie doit être garantie et pouvoir expliquer ses résultats pour entretenir la confiance dans l'économie numérique, et contribuer au développement économique.

Cette dernière assertion n'est toutefois pas valable pour tous les secteurs d'activité. Par exemple, dans le secteur des jeux, et des médias, la performance d'une prévision ou d'une recommandation obtenue dans le cadre d'un algorithme complexe peut être privilégiée même si le résultat n'est pas explicable (l'objectif étant simplement de gagner par exemple dans un jeu d'échec). Les recherches récentes montrent en effet qu'il existe une relation négative entre la performance d'un algorithme d'apprentissage automatique (précision prédictive) et son explicabilité. Les méthodes les plus performantes (par exemple, l'apprentissage profond) sont souvent les moins transparentes, et les méthodes les plus transparentes (les arbres de décision, par exemple) sont parfois moins précises (Bologna et Hayashi, 2017).

Ces questions d'interprétabilité et d'explicabilité sont fondamentales pour le développement de l'IA et de nombreuses questions restent en suspens. Par exemple, que signifie réellement expliquer et que faut-il expliquer ? Quel type d'explicabilité faut-il fournir et pour quels services ? L'explicabilité parfaite existe-t-elle ? En effet, les modèles étant de plus en plus complexes, il va devenir très difficile de déterminer des règles simples et interprétables décrivant le résultat d'un algorithme.

- De la transparence à l'auditabilité des algorithmes

Aux concepts d'interprétabilité et d'explicabilité sont associés ceux de transparence et d'auditabilité des algorithmes. L'idée est de rendre public, ou bien de mettre sous séquestre, des algorithmes en vue les auditer pour étudier des difficultés potentielles. Par exemple, en dépit des controverses relatives à l'usage des algorithmes dans la justice américaine, la société Northpoint qui commercialise l'algorithme a refusé de divulguer les détails de son algorithme propriétaire. Par conséquent, il est impossible aux universitaires et aux autres organisations, y compris aux accusés, d'évaluer dans quelle mesure l'algorithme est potentiellement inéquitable. Cette situation conforte l'idée selon laquelle les algorithmes sont opaques et des boîtes noires (Pasquale, 2015).

Keats Citron et Pasquale (2014) ont très tôt défendu cette position. À propos d'une étude liée à la notation du crédit dans les banques (credit scoring): « Les experts en technologie de la FTC pourraient tester les systèmes de notation pour rechercher les biais, l'arbitraire et les erreurs de caractérisation injustes. Pour ce faire, ils auraient besoin non seulement d'afficher les jeux de données extraits par les systèmes de notation, mais également le code source et les notes du programmeur décrivant les variables, les corrélations et les inférences intégrées aux algorithmes des systèmes de notation ». Diakopoulos (2016) et Chander (2017) partagent cette position : « Fournir de la transparence et des explications sur les sorties algorithmiques pourrait servir un certain nombre d'objectifs, y compris la contrôlabilité, la confiance, l'efficacité, la persuasion, l'efficacité et la satisfaction ».

Mais cet idéal de transparence a aussi des détracteurs. Ananny et Crawford (2016) passent en revue dix limites de l'idéal de transparence telles que la confidentialité, les investissements et la protection des secrets commerciaux, et les limites techniques. Cette position est partagée également par Dana Boyd, sociologue à Microsoft Research, pour qui la transparence est insoutenable et doit être délaissée au profit de la responsabilité (accountability). Cowgill et Tucker (2017) suggèrent également que la transparence n'est pas un problème pertinent car il est difficile de connaître le fonctionnement de certains algorithmes (comme par exemple les réseaux de neurone). Ils suggèrent d'appliquer une évaluation contrefactuelle, une technique utilisée pour quantifier les changements de biais.

Pour limiter les problèmes de secrets commerciaux régulièrement avancés par les entreprises privées, Pasquale (2010) a proposé la solution d'un tiers de confiance, ce qui permet un examen de l'intérêt public sans toutefois permettre la publication de l'algorithme. Cette proposition est conforme à Sandvig et al. (2014) qui ont également proposé l'idée d'audits d'algorithmes. La CNIL en France, l'autorité administrative indépendante qui exerce ses fonctions conformément à la loi française sur la protection des données, appuie également ce point de vue. La CNIL recommande de mettre en place une plateforme nationale d'audit des algorithmes: « ces audits pourraient être réalisés par un organisme public d'experts en algorithmes qui surveilleraient et testeraient les algorithmes (en vérifiant par exemple qu'ils ne pratiquaient pas la discrimination). Compte tenu de la taille du secteur à auditer, une autre solution pourrait impliquer les autorités publiques accréditant des cabinets d'audit privés sur la base d'un cadre de référence ».

Le concept de transparence des algorithmes est souvent associé à celui de responsabilité, car une plus grande transparence des algorithmes est supposée faciliter la responsabilité.

- La responsabilité algorithmique

La responsabilité est le principe selon lequel une personne ou une organisation légalement responsable du préjudice doit fournir une explication ou une compensation au préjudice subi. La prise de décision par un algorithme a soulevé plusieurs questions intéressantes au sujet du caractère intentionnel ou non de la discrimination.

Le cas de la discrimination intentionnelle est probablement le plus simple et le moins fréquent, car les algorithmes n'utilisent pas d'attributs directement protégés pour discriminer. Dans ce cas simple, juridiquement, la responsabilité du programmeur ou de l'entreprise qui commercialise l'algorithme est engagée dans la mesure où la loi interdit la discrimination intentionnelle. Cependant, les algorithmes actuels peuvent discriminer involontairement des personnes sans utiliser d'attribut protégé (lorsque l'attribut protégé est corrélé à d'autres attributs, par exemple). Dans ce cas, la démonstration de la preuve est plus délicate. En effet, une entreprise devrait-elle être tenue responsable de la nature discriminante mais non intentionnelle de son algorithme ?

Il n'est pas toujours facile non plus de savoir qui devrait réparer les dommages. Dans un rapport intitulé « Responsabilité algorithmique », la World Wide Web Foundation prend l'exemple de Facebook et le problème des fausses informations diffusées à partir de plateformes de réseaux sociaux: « Bien que Facebook ait la responsabilité de prendre en compte les fausses informations au sein de son système, la modification des structures d'incitation qui génèrent de fausses informations ainsi que la réparation des torts qui pourraient en résulter est une tâche bien plus importante qui incombe à Facebook mais aussi à d'autres acteurs. » Ils suggèrent par conséquent un autre concept, la justice algorithmique, à savoir la possibilité de réparer les torts causés.

La question de la responsabilité algorithmique a finalement conduit les avocats et les décideurs politiques à suggérer la création d'un nouveau cadre éthique et juridique pour traiter les algorithmes comme des « personnes électroniques ».

## 6. Conclusion

L'intelligence artificielle est souvent vue comme une menace pour l'emploi, le respect de la vie privée ou le contrôle des décisions prises par des systèmes perçus comme des boîtes noires. Les systèmes automatisés fondés sur ce corpus de méthodes sont ainsi parfois accusés de produire des résultats erronés, les données sur lesquelles reposent leurs décisions pouvant par exemple être biaisées par des erreurs de mesure ou 'contaminées' avec une volonté de nuire, et même d'accroître certains types de discrimination, ainsi qu'en attestent les polémiques récentes autour du caractère supposé « raciste » ou « sexiste » de certaines algorithmes incorporés dans des agents conversationnels, des logiciels de police prédictive ou des programmes de recommandation.

Les risques sont en effet bien réels, d'autant plus que les interfaces homme-machine modernes permettent un usage immédiat de certaines solutions logicielles. Mais l'intelligence artificielle ne tiendra ses promesses que si les enjeux d'équité, d'interprétabilité, d'explicabilité et de responsabilité sont considérés au même niveau que la recherche d'efficacité. S'il est encore difficile de savoir comment concevoir une régulation efficace sans brider l'innovation, nul doute que la maîtrise des risques passe en partie par l'éducation, la formation et la diffusion d'une « culture des données et des algorithmes » auprès d'un large public.

Au-delà des approfondissements théoriques et méthodologiques nécessaires à l'élaboration de techniques d'apprentissage statistique plus fiables et plus robustes, d'une plus grande transparence quant à l'élaboration et au fonctionnement des modèles produits par le machine learning, il apparaît indispensable que l'inflation de « solutions technologiques » fondées sur l'analyse des masses de données, s'accompagne d'une plus grande diffusion de la culture probabiliste et statistique dans la plupart des cursus universitaires, et pas seulement dans celui de ces nouveaux spécialistes, les « data scientists ». La transmission de cette culture est sans aucun doute l'une des clefs pour que le citoyen ne se sente pas dépossédé de choix importants qui en appellent aux données, et puisse résister aux éventuelles dérives. Enfin, il serait souhaitable d'élargir cette réflexion en incluant des représentants d'entreprises, des acteurs de la société et du monde politique, autour d'exemples d'utilisations concrètes des algorithmes et de l'intelligence artificielle, pour réfléchir collectivement aux meilleures manières de traiter les aspects de discrimination et d'équité.

## Références

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. ProPublica.

Ananny, M. and Crawford, K. 2016. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New media & society, 1-17.

Bareinboim, E. Tian, J. and Pearl J. 2014. Recovering from Selection Bias in Causal and Statistical Inference. Proceedings of the 28th National Conference on Artificial Intelligence (AAAI), CE Brodley and P. Stone, eds., AAAI Press, Menlo Park, CA p2410-2416.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. 2017. Fairness in criminal justice risk assessments: The state of the art.

Bertail, P., Chautru, E., and Clémençon, S. (2017) Empirical Processes in Survey Sampling with (Conditional) Poisson Designs. Scand J Statist, 44: 97-111. doi: 10.1111/sjos.12243.

Block C. J., Koch, S.M., Liberman, B.E., Merriweather, T.J., and Roberson, L. 2011. Contending With Stereotype Threat at Work: A Model of Long-Term Responses, The Counseling Psychologist 39(4), 570-600.

Buolamwini, J. et Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:77-91.

Boistard, H. Rik Lopuhaä et Anne Ruiz-Gazen. 2017. Functional central limit theorems for single-stage sampling designs, Annals of Statistics, 45(4): 1728–1758.

Bologna, G. et Hayashi, Y. 2017. Characterization of symbolic rules embedded in deep dimlp networks: A challenge to transparency of deep learning. Journal of Articial Intelligence and Soft Computing Research, 7(4):265.

Bolukbasi T., Chang K-W., Zou J., Saligrama V., Kalai, A. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Borrajo, L., Cao, R. 2018. Nonparametric Mean Estimation for Big-But-Biased Data. In The Mathematics of the Uncertain, Studies in Systems, Decision and Control, Springer: Cham, Switzerland, 142, 55-65.

Chander, A. 2017. The Racist Algorithm? Michigan Law Review, 115(6).

Chen, J. et Shao, J. 2000. Nearest Neighbor Imputation for Survey Data, Journal of Official Statistics, 16(2), 113-131.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2):153-163.

Clémençon, S., Bertail P. and Chautru P. 2017. Sampling and empirical risk minimization, Statistics, 51, 1, (30).

Corbett-Davies, S., and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Cortes, C., Mohri, M., Riley, M. et Rostamizadeh, A. 2008. Sample selection bias correction theory. In Proceedings of the 19th International Conference on Algorithmic Learning Theory, ALT '08, 38–53. Berlin, Heidelberg: Springer-Verlag.

Cowgill, B. and Tucker, C. 2017. Algorithmic Bias: A Counterfactual Perspective. NSF Trustworthy Algorithms, December 2017, Arlington, VA.

Diakopoulos, N. 2016. Accountability in algorithmic decision-making. Communications of the ACM 59(2): 56-62.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel R. 2012. Fairness Through Awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, pages 214-226, New York, NY, USA, 2012. ACM.

Edelman, B., Luca, M. and Svirsky, D. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. American Economic Journal: Applied Economics, 9(2).

Fan, J., Fang H., and Han L. 2014. Challenges of Big Data Analysis. National Science Review, 1(2), 293-314.

Friedler, S., Scheidegger, C. and Venkatasubramanian, S. 2016. On the (im)possibility of fairness.

Hall P., Tajvidi, N. 2002. Permutation tests for equality of distribution in high dimension, Biometrika, 89, 359-374.

Hamilton, E. 2016. Benchmarking Four Approaches to Fairness-Aware Machine Learning.

Josse, J. F. Husson, and V. Audigier. 2016. Mimca: Multiple imputation for categorical variables with multiple correspondence analysis. Statistics and Computing, 27: 501-518.

Kahneman D. and Tversky, A. 1974. Judgment under Uncertainty: Heuristics and Biases. Science, New Series, 185(4157): 1124-1131.

Keats Citron D. and Pasquale, F. 2014. The Scored Society: Due Process for Automated Predictions, Washington Law Review.

Kolm, S. C. 1998. Justice and Equity. MIT Press.

Kosinski M., Wang Y. 2018. Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images Journal of Personality and Social Psychology, 114(2): 246-257.

Lambrecht, A. and Tucker, C. 2018. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads.

Lipton, Z. C. 2016. The Mythos of Interpretability. In Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning. arXiv preprint arXiv:1606.03490.

Loftus, J. R., Russell, C., Kusner M. J., and Silva, R. 2018. Causal Reasoning for Algorithmic Fairness.

Megan, P. and Ball. P. 2014. Big Data, Selection Bias, and the Statistical Patterns of Mortality in Conflict. SAIS Review of International Affairs, 34(1), 9-20.

Pasquale, F. 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

Pasquale, F. 2010. Restoring Transparency to Automated Authority, Journal of High Technology and Telecommunications Law.

Pedreshi, D., Ruggieri, S., and Turini, F. 2008. Discrimination-aware Data Mining. InProceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ‘08, pages 560-568, New York, NY, USA, 2008. ACM.

Samek, W., Wiegand, T. et Müller, K-R. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv:1708.08296v1.

Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry.

Szekely G.J., Rizzo, M.L. 2004. Testing for equal distributions in high dimension. InterStat 5 (16.10), 1249-1272.

Tufekci, Zeynep. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In ICWSM ‘14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. 504-514.

Wu X., Zhang X. 2016. Automated inference on Criminality using Face Images, arxiv: 1611.04135v1.