Statistique et Société
Décembre 2022

Volume 10, Numéro 3

Équité en apprentissage automatique

Sommaire
Statistique et Société

Volume 10, Numéro 3

Sommaire
Statistique et Société

Volume 10, Numéro 3

Statistique et société

Magazine quadrimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de présenter, d'une manière attrayante et qui invite à la réflexion, l'utilisation pratique de la statistique dans tous les domaines de la vie. Il s'agit de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Directrice de publication
Anne Philippe, Présidente de la SFdS

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel DIDIER
Rédacteur en chef de Statistique et Société

Chère lectrice, cher lecteur,

Cette nouvelle livraison de *Statistique et Société* porte presque intégralement sur l'intelligence artificielle. Nous présentons un dossier monté par Bilel Benbouzid qui porte sur l'équité algorithmique. Qu'est-ce à dire ? Comme le montre l'introduction au dossier, il s'agit de l'exigence qui pèse sur les algorithmes de traiter les populations et sous-populations qui constituent une société de façon équitable. Le premier article, dû à Gilbert Saporta, expose les principaux concepts et paradoxes permettant de s'emparer de cette question. Philippe Besse présente dans l'article suivant les principales règlementations européennes dans ce secteur. Arthur Charpentier et Laurence Barry creusent, dans le troisième article, le cas spécifique du champ de l'assurance où le traitement équitable des personnes est évidemment crucial. Enfin, la parole est à nouveau donnée à Bilel Benbouzid qui, dans une communication conclusive, présente des outils venant des *Science and Technology Studies* permettant de faire la théorie des faits qui ont été abordés auparavant.
À la suite de ce dossier, vous trouverez une recension par Jean-Jacques Droesbeke du *Manifeste pour une intelligence artificielle comprise et responsable* de Jean-Paul Aimetti, Olivier Coppet et Gilbert Saporta. La livraison s'achève par une évocation d'André Vanoli qui nous a quittés l'an dernier ; elle est due à Quentin Dufour. André Vanoli était l'un des plus grands défenseurs et connaisseurs de la comptabilité nationale en France, et sans doute au monde. Il était en outre un homme curieux, ouvert, très aimable. Il était essentiel de lui rendre hommage.

Enfin, je souhaite vous annoncer une très grande nouvelle pour votre revue : celle-ci est en train de migrer vers la plateforme d'édition en ligne OpenEdition du CNRS. Cette évolution est extrêmement favorable : elle légitime le travail que nous avons effectué depuis dix ans en renforçant la reconnaissance universitaire. Les articles seront ainsi beaucoup mieux référencés et accessibles – toujours gratuitement.

Bonne lecture, et merci de votre fidélité.

Emmanuel Didier

Introduction au dossier « Équité en apprentissage automatique »

Bilel BENBOUZID[1]
Maître de conférences en sociologie, Université Paris Est, Marne-la-Vallée, Laboratoire Interdisciplinaire, Science, Innovation et Société (LISIS)

« La justice est la première vertu des institutions sociales
comme la vérité est celle des systèmes de pensée. »
John Rawls

Pour celles et ceux qui s'intéressent à la prise en compte de l'équité dans les systèmes d'intelligence artificielle, la célèbre analogie qu'établit Rawls en 1971 dans *Théorie de la justice* peut sonner étrangement (Rawls, 2009). Dans ce domaine de recherche nouveau, à la croisée du *machine learning* et des sciences sociales, les valeurs de justice et vérité semblent moins reposer sur une logique de similitude de forme que sur celle d'une liaison intime. L'analogie de Rawls est d'autant plus frappante aujourd'hui que les machines prédictives se présentent à la fois comme des institutions sociales et des systèmes de pensée où s'entremêlent, de manière indissociable, les valeurs de justice et vérité.

C'est suite à de nombreuses controverses sur les discriminations algorithmiques qu'a émergé ce projet de rendre John Rawls fongible dans les systèmes d'intelligence artificielle, pour ainsi dire. Depuis une dizaine d'années, l'appel à la moralisation des statistiques d'apprentissage s'est traduit par un débat scientifique sur la *mise en nombre* des théories de la justice et du droit antidiscriminatoire. Cette mathématisation de l'équité pose un ensemble de questions épistémologiques déjà bien connu du côté de l'économie du bien-être autour des problèmes d'allocation des ressources comme de celui de l'économie expérimentale et de ses techniques de *testing* des discriminations : comment prendre en compte à la fois les relations entre les phénomènes sociaux et les normes que l'on souhaite respecter ? Peut-on réaliser des prescriptions de bien-être sur une base positiviste, tout en ayant recours à un jugement de valeur ? Sur quels critères statistiques faut-il mesurer et détecter les discriminations ? Et d'une manière plus générale, comment juger et comparer la « qualité » des situations sociales des personnes et les décisions qui leur sont associées ?

Si les questions qui entourent l'évaluation quantitative de l'équité et la détection statistique des discriminations sont anciennes, le domaine de la *fairness* dans le *machine learning* (appelé aussi *FairML*) reste un objet d'étude original pour qui s'intéresse à l'histoire et la sociologie de la quantification. En effet, ce sont désormais les techniques statistiques elles-mêmes, et les

1. bilel.benbouzid198@gmail.com

algorithmes qui leur sont associés, qui sont au cœur du débat public sur la justice sociale et les discriminations. C'est que la statistique d'apprentissage automatique fait de plus en plus partie de notre quotidien au point d'en être devenue un enjeu de régulation juridique. Les algorithmes influencent non seulement nos interactions individuelles sur les plateformes numériques, mais aussi tout un ensemble de décisions administratives qui façonnent la société dans son ensemble. Les machines prédictives affectent si profondément nos vies que nous devons nous assurer que leurs décisions automatisées sont vérifiables, responsables et justes.

Dans ce dossier de *Statistiques et Société*, nous avons souhaité porter à la discussion les travaux relatifs à cette prise en compte de l'équité dans le *machine learning*. Ce projet éditorial est né d'un atelier pluridisciplinaire organisé avec Ruta Binkytė-Sadauskienė, par le Centre Internet et Société (CIS), en partenariat avec le Laboratoire interdisciplinaire Sciences Innovations Sociétés (LISIS), qui a rassemblé à l'école AIVANCITY (Cachan) en 2021 la communauté scientifique française autour de la recherche sur la justice sociale, l'équité et les discriminations dans les systèmes algorithmiques.

Parmi les nombreuses interventions de ces deux journées d'atelier, nous avons retenu pour ce numéro les contributions apportant un regard particulièrement réflexif sur le sujet, notre objectif étant moins de contribuer en substance à ce nouveau domaine foisonnant que de prendre le recul nécessaire pour en cerner les principaux aspects méthodologiques, juridiques et épistémologiques. Sur les quatre articles composant ce numéro, trois sont directement issus de notre rencontre. Nous en avons introduit un quatrième afin d'apporter des éléments introductifs nécessaires à la compréhension des débats scientifiques internes.

L'article de Gilbert Saporta, « Équité, explicabilité, paradoxe et bais », fait office de cette entrée en matière didactique. Il est tout à fait précieux pour ce numéro d'avoir pu bénéficier du regard d'un des pionniers de « l'analyse de données à la française », domaine plus proche des *data sciences* contemporaines que les statistiques dites fréquentistes qui ont longtemps dominé le champ. L'auteur du célèbre manuel plusieurs fois réédité, *Probabilités, analyse des données et statistique*, soutient tout d'abord qu'il faut distinguer les problèmes d'équité de ceux d'explicabilité et d'interprétabilité. Il s'agit là d'un point de vue purement statistique qui envisage l'explicabilité et l'interprétabilité respectivement dans le sens étroit de la capacité à rendre compte des liens entre les variables et de la simplification des modèles. C'est que si l'équité pose des problèmes statistiques, elle implique moins une posture explicative qu'une éthique de la compréhension. Gilbert Saporta propose ensuite un tour d'horizon des approches, métriques et biais qu'il illustre à partir du célèbre cas COMPAS, le logiciel de la prédiction de la récidive accusé par un groupe de data journalistes de discriminer les minorités ethniques. Dans un style simple, Gilbert Saporta parvient à rendre compte d'un débat compliqué sur la diversité des approches. Pour terminer, l'auteur affirme sans détour que « nous ne pouvons pas attendre des algorithmes qu'ils corrigent les inégalités ». Voilà une assertion qui a le mérite d'être claire, mais qui est en fait au cœur des débats de la communauté du FairML. Rappelons que les plus éminents contributeurs du domaine estiment que les systèmes d'IA sont moins des dangers que de véritables opportunités pour mieux maitriser les problèmes de justice sociale et de discrimination (Abebe *et al.*, 2020).

Comme Gilbert Saporta, Philippe Besse revient sur la pluralité des approches dans son article « Conformité européenne des systèmes d'IA : outils statistiques élémentaires », mais l'illustration, plus empirique, repose sur un jeu de données « jouet » sur l'octroi de crédit bancaire. Cette étude de cas est un bon moyen de mettre en perspective les pratiques concrètes du FairML avec les exigences juridiques de l'*Artificial Intelligence Act* (AI ACT), la réglementation émergente au niveau européen. Philippe Besse connaît bien la matière puisqu'avec ses collaborateurs de l'université de Toulouse, ils sont parmi les premiers en France à contribuer au débat scientifique sur la *fairness* dans le *machine learning*. Il est donc particulièrement bien placé pour s'interroger

sur la pertinence de ce texte réglementaire qui suscite un vif débat au niveau international. Quelles seront les conséquences de l'AI Act, se demande Philippe Besse, pour les statisticiens impliqués dans la conception des systèmes d'IA ? Quels sont les outils à leur disposition pour répondre aux obligations à venir de mise en conformité des machines prédictives ? Comme le montre Philippe Besse, la réglementation est loin d'être claire à ce sujet, en particulier en ce qui concerne l'équité. Comment certifier qu'une machine soit *fair* ? La réglementation européenne semble répondre à cette question en visant davantage la protection des « vendeurs » de machine que celle des citoyens – l'administration de la preuve de la discrimination algorithmique n'est pas univoque et dépend toujours du point de vue du système et du contexte d'usage.

C'est sans doute l'hétérogénéité des situations concernées par l'équité des systèmes algorithmiques qui fait obstacle à la normalisation. Chaque secteur (la banque, la justice, la police, les diagnostics médicaux, etc.) a une histoire et un rapport à l'équité et aux discriminations qui lui est propre. C'est pourquoi nous avons choisi d'intégrer à ce numéro un cas d'usage particulier. Dans « L'équité de l'apprentissage machine en assurance », Arthur Charpentier et Laurence Barry reviennent sur le cas particulier de l'assurance, un secteur qui est confronté de longue date au problème de l'équité dans les données et les modélisations. Les auteurs rappellent que « discriminer » est l'essence même de la classification, et qu'en assurance « toute discrimination statistique est susceptible d'être perçue comme une injustice, rejoignant ainsi le langage courant de discrimination sociale ». Les auteurs retracent la longue histoire du traitement des biais tarifaires et montrent ce qui change actuellement avec l'usage de l'apprentissage machine en assurance : les tarifications de plus en plus personnalisées, paradoxalement, exacerbent les biais déjà connus de longue date du monde de l'assurance, tout en limitant leur « contestabilité ».

Pour conclure, notre article « L'équité dans la machine ou comment le *machine learning* devient scientifique en tournant le dos au réalisme métrologique », propose d'analyser le domaine du FairML en mobilisant les outils de la sociologie de la quantification. Nous retraçons d'abord comment cette spécialité s'inscrit dans la continuité des travaux sur la *privacy*. Ce sont les chercheurs qui ont œuvré aux problèmes de *privacy* dans la conception des systèmes algorithmiques qui ont mis à l'agenda scientifique les problématiques de *fairness*. La posture de ces chercheurs est tout à fait originale du point de vue de l'histoire des pratiques de quantification : c'est en entretenant une proximité politique avec leur objet de recherche (la *privacy* et la *fairness*) qu'ils parviennent à des énoncés scientifiques qui *tiennent*. Du point de vue de l'analyse des rapports entre statistique et société au cœur de la ligne éditoriale de cette revue éponyme, le FairML, comme nous le soulignons dans notre article, est « un bon observatoire de la politique des statistiques et de leur transformation actuelle ».

Références

Abebe R., Barocas S., Kleinberg J. *et al*. (2020), « Roles for computing in social change », *in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 27 January 2020), FAT* '20, Association for Computing Machinery, pp. 252-260.

Rawls J. (2009), *Théorie de la justice*, Paris, Points.

Équité, explicabilité, paradoxes et biais

Gilbert SAPORTA[1]
Laboratoire Cedric, Conservatoire national des arts et métiers, Paris

TITRE

Fairness, explainability, paradoxes and bias

RÉSUMÉ

L'équité ou fairness des algorithmes suscite une abondante littérature.
Sur un plan qualitatif, on examinera les liens entre équité, explicabilité et interprétabilité. On peut penser qu'il vaut mieux comprendre le fonctionnement d'un algorithme pour savoir s'il est équitable, mais en fait il n'en est rien car la transparence ou l'explicabilité sont relatives à l'algorithme alors que l'équité concerne son application différenciée à des groupes d'individus. Suivant Rudin (2019), on distinguera l'interprétabilité, qui est liée à la simplicité, de l'explicabilité qui est en général *post-hoc* avec des approches globales ou locales, agnostiques ou spécifiques, utilisant souvent des modèles de substitution (Molnar, 2021).
La diversité des mesures d'équité ne simplifie pas son appréhension : Verma et Rubin (2018) en ont dénombré plus de vingt qui conduisent d'ailleurs à des incompatibilités comme l'illustre la controverse sur laquelle nous reviendrons concernant l'application « COMPAS » de prédiction de la récidive.
Les « biais » des algorithmes ne sont souvent que la reproduction de ceux des décisions antérieures que l'on retrouve dans les données d'apprentissage. Mais ce ne sont pas les seuls. On tentera de dresser une typologie des principaux biais : statistiques, sociétaux, cognitifs, etc.

Mots-clés **:** *équité, algorithmes, biais, apprentissage.*

ABSTRACT

The fairness of algorithms is the subject of an abundant literature.
On a qualitative level, we will examine the links between fairness, explainability, and interpretability. One may think that it is better to understand the functioning of an algorithm to know if it is fair, but in fact this is not the case because transparency or explicability are relative to the algorithm whereas fairness concerns its differentiated application to groups of individuals. Following Rudin (2019), we distinguish interpretability, which is related to simplicity, from explainability, which is generally *post-hoc* with global or local, agnostic or specific approaches, often using surrogate models (Molnar, 2021).
The diversity of fairness measures does not simplify its apprehension: Verma and Rubin (2018) have counted more than twenty of them, which moreover lead to incompatibilities as illustrated by the controversy to which we will return concerning the "COMPAS" recidivism prediction application.
The "biases" of the algorithms are often simply the reproduction of those of previous decisions found in the training data. But they are not the only ones. We will try to draw up a typology of the main biases: statistical, societal, cognitive, etc.

Keywords**:** *Fairness, equity, algorithms, bias, machine learning.*

1. gilbert.saporta@cnam.fr

1. Les algorithmes en question

Une vaste littérature de dénonciation accuse de discrimination les algorithmes d'apprentissage couramment utilisés pour accepter des demandes de prêt, sélectionner des réponses à des offres d'emploi, ou prédire la récidive dans la justice pénale. Les livres de O'Neil (2016) et Noble (2018) en sont les exemples les plus connus. Pour des aspects non polémiques, on pourra consulter trois articles récents de revue des questions d'équité des algorithmes : Mitchell *et al*. (2021), Tsamados *et al*. (2022), et Pessac et Shmueli (2023).

1.1 Les risques de discrimination

Les algorithmes sont-ils racistes ou sexistes ? Si l'hypothèse de la malveillance relève du complotisme, les mauvais usages sont fréquents et sont largement documentés, comme la confusion entre des images de noirs américains et de gorilles dans un algorithme de Google (Barr, 2015).

En 2014, dans un rapport sur le Big Data de l'*Executive Office of US President*[2] on peut lire : « L'utilisation croissante d'algorithmes pour prendre des décisions d'éligibilité doit être surveillée de près pour des résultats discriminatoires potentiels pour les groupes défavorisés, même en l'absence d'intention discriminatoire ».

En 2018, le cabinet Gartner[3] prévoyait que : « À l'horizon 2022, 85 % des projets d'intelligence artificielle donneront des résultats erronés en raison de biais dans les données, les algorithmes ou les équipes chargées de les gérer ».

Le rapport établi pour le conseil de l'Europe par le professeur de droit F. Z. Borgesius (2018) dresse un panorama des domaines dans lesquels l'IA suscite des risques de discrimination, y compris des cas potentiels de discrimination délibérée (détection des femmes enceintes), et s'attache aux réponses réglementaires et juridiques.

Ces utilisations discutables soulèvent deux types de considérations : d'une part, les algorithmes, mais surtout, d'autre part, les données.

1.2 Transparence et équité

Depuis les dix principes de la déclaration de Montréal en 2018[4], de nombreux codes et déclarations d'éthique sur l'utilisation de l'intelligence artificielle ont vu le jour, tels ceux de l'OCDE (2019), de l'UE (2019) et de l'UNESCO (2021) qui insistent sur la transparence, l'explicabilité et l'équité. Certains codes y ajoutent la responsabilité des entreprises et l'auditabilité des algorithmes (possibilité d'évaluer des algorithmes, des modèles et des jeux de données ; d'analyser le fonctionnement, les résultats et les effets, même inattendus, des systèmes d'IA).

En suivant Rudin (2019), on distinguera l'interprétabilité, qui est liée à la simplicité, de l'explicabilité, qui est généralement *post-hoc* avec des approches globales ou locales, agnostiques (c'est-à-dire sans modèle particulier) ou spécifiques, utilisant souvent des modèles de substitution (Molnar, 2022). Parmi les modèles interprétables les plus courants, on citera ceux à base de règles, comme les arbres de décision et les modèles de régression linéaire. L'explicabilité des modèles recouvre de nombreuses méthodes telles que les analyses d'importance des variables ou de sensibilité, et l'approximation locale par des modèles interprétables.

2. *Big Data: Seizing Opportunities and Preserving Values*, https://www.hsdl.org/?view&did=752636
3. https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence
4. « Déclaration de Montréal pour le développement responsable de l'IA », Université de Montréal, https://www.declarationmontreal-iaresponsable.com/la-declaration

On peut penser qu'il vaut mieux comprendre le fonctionnement d'un algorithme pour savoir s'il est équitable, mais en fait ce n'est pas le cas car la transparence ou l'explicabilité sont relatives à l'algorithme alors que l'équité concerne son application différenciée à des groupes d'individus.

1.3 Apprentissage et décisions

Les algorithmes de *Machine Learning* ou d'IA apprennent à partir de données et en déduisent des règles qui s'appliqueront à des cas futurs. Il est évident que si les données d'apprentissage ne sont pas représentatives ou présentent d'autres types de biais, ceux-ci seront reproduits.

Une littérature abondante et des conférences spécialisées comme l'*ACM Conference on Fairness, Accountability, and Transparency*[5] se sont récemment développées au sein des communautés en informatique. Les guides de bonnes pratiques et les manifestes fleurissent. Mais la communauté en statistique est souvent absente de ces débats, alors qu'elle a développé depuis près de deux siècles des compétences dans le domaine de la collecte et du traitement des données et en connaît les pièges. Notons toutefois les contributions suivantes : Besse (2020), Besse *et al*. (2018 et 2022), del Barrio *et al*. (2020), Bertail *et al*. (2019).

Nous considérerons essentiellement dans cet article les algorithmes de décision binaire qui peuvent avoir un impact négatif sur la vie des personnes en restreignant leur liberté ou en leur refusant des avantages auxquels elles pourraient avoir droit.

Ces algorithmes utilisent des données d'apprentissage pour prédire un comportement binaire Y, tel que le remboursement ou non d'un prêt, en fonction de caractéristiques X. Ils aboutissent à une règle de décision D (accepter ou refuser). Face à une caractéristique sensible A (couleur de peau, sexe, etc.)[6], on soupçonnera une discrimination ou un algorithme injuste si la probabilité de refus dépend de A. Le refus sera noté $D = 1$, et l'acceptation $D = 0$ dans ce qui suit.

D'autres types d'algorithmes de *Machine Learning*, tels les algorithmes de recommandation, sont moins controversés car ils ne sont pas considérés comme des décisions à fort enjeu. Ils ont également l'avantage de pouvoir valider aisément les prévisions puisque la réponse Y, par exemple achat ou non-achat, peut être observée.

Des travaux récents (Stinson, 2022) attirent cependant l'attention sur les risques de discrimination des algorithmes de recommandation de type filtrage collaboratif par opposition au filtrage par contenu. Ces algorithmes consistent à recommander à un utilisateur les produits choisis par des utilisateurs statistiquement proches. L'hypothèse de base est que personne n'est unique, mais si les utilisateurs les plus originaux ou éloignés des goûts des autres s'avèrent être des personnes qui appartiennent à des groupes minoritaires, le filtrage collaboratif va être biaisé en faveur de la majorité. Ce type de biais peut avoir pour effet de marginaliser encore davantage les personnes déjà marginalisées. Par ailleurs, les recommandations de contenus peuvent avoir un impact significatif sur l'accès à l'information en renforçant l'effet d'enfermement dans des « bulles » (Pariser, 2011) où seules certaines informations, dont des *fake news*, sont partagées. Comme le dit l'article de Stinson cité : « Ces biais ne sont pas le résultat d'ensembles de données biaisées, ni des préjugés personnels des créateurs d'algorithmes ; ils sont le résultat d'hypothèses faites lors de la conception des algorithmes eux-mêmes ».

5. https://facctconference.org/
6. Aux États-Unis, les catégories sensibles correspondent à des groupes légalement protégés et la loi fédérale rend illégale toute discrimination fondée sur : la race, la couleur, l'origine nationale, la religion, le sexe, le handicap, l'âge (40 ans et plus), le statut de citoyen, les informations génétiques. En Californie, il existe 18 groupes protégés : https://www.senate.ca.gov/content/protected-classes

2. Quelques approches naïves

Les premières tentatives pour mesurer l'équité ont porté sur la distribution de la variable de décision D, indépendamment de celle de la variable d'intérêt Y.

2.1 La parité démographique

On peut souhaiter (ou exiger) que la probabilité de refus soit la même pour chaque catégorie d'une variable sensible A, ce qui se traduit mathématiquement par l'égalité

$$P(D = 1 \mid A = a) = P(D = 1 \mid A = a') \quad \text{pour tout } a, a'.$$

On l'appelle *parité démographique* ou encore *parité statistique*.

Cette approche présente divers inconvénients :

- La parité démographique ne garantit pas que les comportements (les catégories de la variable Y) soient identiques pour chaque groupe. Accorder la même proportion de prêts par tranche d'âge ne garantit pas les mêmes probabilités de remboursement.

- La parité démographique peut être équitable au niveau du groupe, mais injuste au niveau individuel. Si, par exemple, les qualifications sont différentes pour une catégorie protégée, imposer la parité démographique peut signifier qu'une personne moins qualifiée pourra être embauchée. Par conséquent, si un grand nombre de candidats masculins non qualifiés est ajouté au vivier de candidats, l'embauche de candidates qualifiées diminuera[7].

- Dans le même ordre d'idée, Bertail *et al*. (2019) écrivent : « Appliquée au cas des admissions dans les collèges, par exemple, l'équité de groupe stipulerait que les taux d'admission sont égaux pour les attributs protégés (sexe, etc.), tandis que l'équité individuelle exigerait que chaque personne soit évaluée indépendamment de son sexe ».

- Si les femmes ont moins accès à l'éducation et à la formation que les hommes, elles ont souvent moins de chances d'être recrutées à des postes à responsabilité. Il existe alors une discrimination à l'embauche, mais imposer des quotas pour rétablir la parité démographique ne permettra pas de s'attaquer à la cause profonde. Par ailleurs, il peut exister des biais sociétaux qui entrainent qu'à compétences égales on préfère choisir un homme à une femme, ou l'inverse, selon le type de métier.

- Constater que la probabilité de refus dépend d'une caractéristique sensible, i.e.

  $$P(D = 1 \mid A = a) > P(D = 1),$$

  n'est pas suffisant pour caractériser la discrimination tant que l'influence de toutes les covariables X (parfois appelées facteurs de confusion) n'a pas été contrôlée. Mais déterminer l'ensemble des facteurs de confusion afin de n'en omettre aucun est un problème difficile, voire impossible.

2.2 L'équité par l'ignorance (unawareness)

Selon ce point de vue, un algorithme est équitable tant que les attributs protégés A ne sont pas explicitement utilisés dans le processus de décision. Voir Dwork *et al*. (2012) et Chen *et al*. (2019).

7. https://ocw.mit.edu/resources/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/fairness-criteria/

Mais le fait de retirer les variables sensibles de la liste des prédicteurs ne les empêche pas d'être influentes si d'autres variables leur sont liées, comme le code postal qui pourrait être un proxy pour la race ou la pauvreté. Il a été démontré au contraire que l'omission de la caractéristique sensible de l'équation ne rend pas le modèle de décision exempt de discrimination et que, pour éliminer les biais, la caractéristique sensible doit être utilisée dans le processus de modélisation. C'est un cas particulier du « biais de la variable omise » (cf. Žliobaitė and Custers, 2016), que l'on pourrait dénommer ici *biais de l'autruche*.

3. De nombreuses (trop nombreuses !) mesures d'équité

Pour aller au-delà de la parité démographique et de l'équité par ignorance, dont nous venons de voir les lacunes, d'autres exigences ont été formulées qui conduisent à une grande diversité de mesures : Verma et Rubin (2018) en ont dénombré plus de vingt tandis que, parmi les outils libres mesurant l'équité des modèles d'IA, *AI Fairness 360* d'IBM (Bellamy *et al*., 2019) en propose 71 !

Elles sont pour la plupart incompatibles et il existe de nombreux résultats d'impossibilité. Nous ne les développerons pas toutes en nous focalisant sur les plus connues qui incluent la variable d'intérêt Y.

Le tableau suivant, adapté de Mitchell *et al*. (2021), résume les différents cas de croisement entre les modalités de la variable binaire Y et de sa prévision D. On parle de valeur positive pour la modalité 1 et de valeur négative pour la modalité 0. Ainsi un faux positif correspond à la décision erronée $D = 1$ alors que $Y = 0$. On retrouve ici des notions familières en épidémiologie.

Tableau 1 – *Matrice de confusion*

	$Y = 1$	$Y = 0$	$P(Y = 1 \mid D)$	$P(Y = 0 \mid D)$
$D = 1$	Vrai positif	Faux positif	$P(Y = 1 \mid D = 1)$ Valeur prédictive positive	
$D = 0$	Faux négatif	Vrai négatif		$P(Y = 0 \mid D = 0)$ Valeur prédictive négative
$P(D = 1 \mid Y)$	$P(D = 1 \mid Y = 1)$ Taux de vrais positifs	$P(D = 1 \mid Y = 0)$ Taux de faux positifs		
$P(D = 0 \mid Y)$	$P(D = 0 \mid Y = 1)$ Taux de faux négatifs	$P(D = 0 \mid Y = 0)$ Taux de vrais négatifs		

3.1 Égalisation des chances ou equalized odds

On peut ainsi exiger que les taux de faux positifs (et donc les taux de vrais négatifs) soient identiques, ou que les taux de faux négatifs (et donc de vrais positifs) soient identiques quel que soit le groupe protégé (*equal opportunity*). La combinaison des deux (appelée chances égalisées ou *equalized odds* ou encore *separation*) reflète une notion d'équité selon laquelle les personnes ayant le même résultat doivent être traitées de la même manière, indépendamment de l'appartenance à un groupe sensible. Exemple : en matière de prêts bancaires, ceux qui feront défaut (resp. ceux qui ne feront pas défaut) devraient avoir la même probabilité de rejet (resp. d'acceptation), quelle que soit par exemple leur couleur de peau. En termes mathématiques :

$$P(D = 1 \mid Y = 0, A = a) = P(D = 1 \mid Y = 0, A = a'),$$
$$P(D = 0 \mid Y = 1, A = a) = P(D = 0 \mid Y = 1, A = a').$$

D est alors indépendant de A, conditionnellement à Y.

3.2 Parités prédictives

Une autre façon de définir l'équité consiste à exiger l'égalité des valeurs prédictives négatives :

$$P(Y = 0 \mid D = 0, A = a) = P(Y = 0 \mid D = 0, A = a'),$$

ou l'égalité des valeurs prédictives positives :

$$P(Y = 1 \mid D = 1, A = a) = P(Y = 1 \mid D = 1, A = a').$$

La combinaison de ces deux parités prédictives est appelée suffisance (*sufficiency*). En d'autres termes, toutes les personnes qui se sont vues refuser un prêt auraient la même probabilité d'être en défaut de paiement si le prêt leur avait été accordé et les personnes appartenant aux groupes favorisés et défavorisés qui se voient accorder un prêt le remboursent avec la même probabilité. Cette propriété reflète une notion d'équité selon laquelle les personnes ayant subi la même décision auraient eu des résultats similaires, quel que soit leur groupe (Mitchell *et al*., 2021). Le conditionnement sur la décision semble plus approprié au fait que la décision intervient avant la réalisation de la réponse Y.

Un résultat d'impossibilité montre que l'on ne peut avoir simultanément les propriétés d'*equalized odds* et de *sufficiency*.

3.3 La controverse COMPAS (Rudin *et al*., 2020 ; Mitchell *et al*., 2021 ; Wang *et al*., 2022)

La controverse autour du modèle COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) illustre ce qui précède. COMPAS fournit un score de prédiction de la récidive fréquemment utilisé par les tribunaux américains. L'association ProPublica a constaté que COMPAS ne satisfaisait pas à l'égalité des taux de faux positifs selon la race : parmi les prévenus qui n'ont pas été réarrêtés, les prévenus noirs étaient deux fois plus susceptibles d'être classés à tort comme étant à haut risque. Ils ont conclu que l'outil était biaisé contre les noirs.

La société ayant développé COMPAS a répliqué que son système n'était pas discriminatoire puisqu'il satisfaisait à des valeurs prédictives positives égales : parmi les personnes dites à haut risque, la proportion de condamnés qui ont été arrêtés à nouveau était approximativement la même, quelle que soit la race.

Ces affirmations toutes deux exactes correspondent à deux définitions incompatibles de l'équité qui ne peuvent être satisfaites simultanément que si (a) le taux de récidive et la distribution des scores sont les mêmes pour tous les groupes raciaux ou (b) certains groupes ne sont jamais susceptibles d'être concernés par certains des résultats (par exemple si les blancs ne sont jamais réarrêtés).

Le choix d'une mesure revient à choisir une définition de l'équité, qui relève en réalité d'un choix éthique et non statistique (Lee *et al*., 2021).

3.4 Autres difficultés

Le caractère opérationnel des mesures d'équité est également problématique, non seulement parce que le résultat Y ne sera souvent observable que bien après la décision, mais surtout en raison d'un problème contrefactuel puisque, dans certains cas, la décision interdit l'observation : on ne saura jamais si un prêt refusé aurait été remboursé.

Dans d'autres cas, comme la sélection de candidats pour un emploi, la variable Y qui consisterait à déterminer si le candidat recruté fait bien son travail n'est pratiquement jamais observable. Les algorithmes ne font alors qu'automatiser les processus antérieurs.

4. Algorithmes injustes ou données biaisées ?

Un algorithme prédictif n'est pas inéquitable en soi. Il est entraîné pour optimiser la prédiction de la réponse sur les données d'apprentissage en supposant que les données futures proviendront de la même distribution. Si l'ensemble d'apprentissage est biaisé, il reproduira ces biais tels des stéréotypes de genre ou ethniques.

4.1 Biais statistiques et remèdes possibles

Les biais d'échantillonnage sont très fréquents lorsqu'on sélectionne certains cas plus que d'autres. Si les probabilités d'inclusion sont connues et non-nulles, ou si l'on dispose de variables de calage, la solution consistant à modifier les poids des observations est généralement efficace.

Les biais de sélection et les données manquantes non aléatoires, comme les erreurs autres que d'échantillonnage (erreurs de couverture), peuvent conduire à des erreurs graves, difficiles à corriger sans modèle. C'est le cas de la notation du risque de crédit basée uniquement sur les candidats acceptés. En effet, les données d'apprentissage qui incluent la variable de comportement Y (bon ou mauvais payeur), ne sont pas représentatives de l'ensemble des demandes de crédit car certaines ont été rejetées d'emblée. Ce problème connu sous le nom de *reject inference* a fait l'objet de nombreux travaux, pas toujours concluants (Hand et Henley, 1993) qui nécessitent de modéliser le processus de rejet avec par exemple des modèles *logit* ou *tobit*, quand cela est possible (dossiers non conservés, décisions humaines subjectives).

4.2 Autres types de biais

Les biais de mesure et biais technologiques : par exemple, la reconnaissance faciale ne parvient pas à reconnaître les personnes de couleur avec autant de précision que les personnes à peau blanche car les paramètres par défaut des caméras ne sont souvent pas optimisés pour capturer les tons de peau plus foncés, ce qui donne lieu à des images de qualité inférieure dans les bases de données sur les noirs américains (Najibi, 2020). Ce biais vient d'ailleurs se rajouter au biais des bases de données d'images qui comportent souvent un nombre insuffisant de personnes de couleur.

Biais « historique » : les données peuvent être représentatives mais reproduire des inégalités. Même le meilleur algorithme prédira des salaires inférieurs pour les femmes si de telles inégalités préexistent.

Biais sociaux et cognitifs : les taux de criminalité reflètent des structures sociales inégales et aussi des inégalités éventuelles dans les décisions de justice.

On trouvera des compléments dans Srinivasan et Chander (2021) et Merhabi *et al*. (2022) qui répertorient un grand nombre de biais en Intelligence Artificielle : biais de mesure, d'étiquetage, d'agrégation (paradoxe de Simpson), de confusion et bien d'autres.

La littérature sur les biais cognitifs concerne plutôt les décisions humaines. On en dénombre des dizaines de types à la suite des travaux de Tversky et Kahneman (1974). Ces biais peuvent entacher la qualité des bases d'apprentissage des algorithmes de décision.

4.3 Données déséquilibrées

Lorsqu'on cherche à prédire un comportement rare, les données d'apprentissage sont souvent très déséquilibrées. Ce n'est pas techniquement un problème de biais d'échantillonnage car les données peuvent respecter les vraies proportions. Cependant, le risque est de trouver un taux élevé de faux positifs lorsque l'on veut prédire avec une bonne probabilité de succès la catégorie rare.

Exemple : nous voulons qu'un algorithme détecte avec une très forte probabilité, disons 0,95, la présence d'une maladie grave comme un mélanome à partir d'images de tumeurs. Si l'algorithme n'est pas très spécifique, on doit s'attendre à confondre de nombreuses tumeurs bénignes avec des mélanomes. En poussant le raisonnement jusqu'au bout, pour ne manquer aucun cas (sensibilité = 100 %), il faudrait prédire que tous les cas sont des mélanomes.

4.4 Bruit et biais

Au-delà des biais, un algorithme peut aussi se révéler injuste s'il est instable, ou, en d'autres termes, non robuste, car la variabilité des prévisions possibles fait courir des risques inattendus. Les exemples abondent en reconnaissance d'image pour des algorithmes très complexes, où des modifications invisibles à l'œil comme la modification d'un seul pixel peuvent conduire à des décisions erronées (Su *et al*., 2019).

4.5 Causalité, corrélation, déterminisme et décisions individuelles

Chaque être humain est unique et prédire son comportement en fonction des caractéristiques des groupes auxquels il appartient est inévitablement une source d'erreur et d'iniquité potentielle.

Ceci est d'autant plus vrai que de nombreux algorithmes sont basés sur des corrélations et non des causalités. L'utilisation de modèles causaux est souhaitable pour éviter l'utilisation de corrélations fallacieuses, mais reste difficile lorsque les cas sont complexes. En termes de comportement humain, les prédictions des modèles causaux ne sont pas plus déterministes que le tabagisme pour le cancer du poumon : même si le lien de causalité a été établi, tous les fumeurs ne développeront pas un cancer. Enfin, ce n'est pas parce qu'un lien de causalité a pu être découvert que la décision qui en découle sera forcément juste au sens éthique : si une disposition interdit l'accès des femmes à certaines professions, il y a bien causalité et la prévision est facile. Il y a clairement discrimination ; qu'elle soit justifiée ou non n'est plus un problème statistique.

5. Conclusion et perspectives

Ce que l'on appelle biais dans les algorithmes est le plus souvent un biais dans les données d'apprentissage car très souvent les algorithmes tentent de reproduire d'anciennes règles de décision, basées sur des données biaisées.

L'optimisation d'une mesure d'équité d'un algorithme est une approche intéressante, mais elle se heurte au problème du choix d'une mesure parmi plusieurs dizaines de mesures connues et à l'impossibilité de satisfaire simultanément plusieurs de ces critères.

Un domaine de recherche émergent est celui de l'équité contrefactuelle : en s'inspirant du modèle causal de Pearl, on considère qu'un modèle est équitable si, pour un individu ou un groupe particulier, la prévision dans le monde réel est la même que celle dans le monde

contrefactuel où l'individu ou le groupe auraient appartenu à une catégorie différente. En d'autres termes si la prévision était la même quelle que soit la catégorie de la variable sensible et toutes choses égales par ailleurs (Kusner *et al*., 2017).

L'utilisation de mesures d'équité est utile pour détecter les discriminations et les abus, mais le concept d'équité n'est ni un concept statistique, ni un concept informatique : c'est un concept éthique qui va bien au-delà et relève plutôt de la philosophie et de l'économie politique (Rawls, 1971 et 2001 ; Kolm, 1971). Nous ne pouvons pas attendre des algorithmes qu'ils corrigent les inégalités.

Remerciements

Je remercie vivement les rapporteurs anonymes pour leurs remarques et commentaires pertinents qui m'ont permis d'améliorer une première version de cet article, ainsi que Rich Timpone (Ipsos-Global Science Organisation) pour nos nombreuses discussions.

Références

Barr A. (2015), « Google mistakenly tags black people as 'gorillas,' showing limits of algorithms », *The Wall Street Journal*, 1(7), https://www.wsj.com/articles/BL-DGB-42522.

Bellamy R. K. *et al*. (2019), « AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias », *IBM Journal of Research and Development*, 63(4/5), pp. 4-1.

Bertail P., Bounie D., Clémençon S. et Waelbroeck P. (2019), « Algorithmes : Biais, Discrimination et Équité », https://hal.telecom-paris.fr/hal-02077745.

Besse P. (2020), « Détecter, évaluer les risques des impacts discriminatoires des algorithmes d'IA », https://hal.archives-ouvertes.fr/hal-02616963.

Besse P., Castets-Renard C., Garivier A. et Loubes J.-M. (2018), « L'IA du quotidien peut-elle être éthique ? Loyauté des Algorithmes d'Apprentissage Automatique », *Statistique et Société*, 6(3), pp. 9-31.

Besse P., del Barrio E., Gordaliza P., Loubes J.-M., and Risser L. (2022), « A survey of bias in machine learning through the prism of statistical parity », *The American Statistician*, 76(2), pp. 188-198.

del Barrio E., Gordaliza P., and Loubes, J.-M. (2020), « Review of mathematical frameworks for fairness in machine learning », *arXiv preprint*, arXiv:2005.13755.

Borgesius F. Z. (2018), « Discriminations, intelligence artificielle et décisions algorithmiques », Direction générale de la Démocratie, Conseil de l'Europe, https://rm.coe.int/etude-sur-discrimination-intelligence-artificielle-et-decisions-algori/1680925d84.

Chen J., Kallus N., Mao X., Svacha G., and Udell M. (2019), « Fairness under unawareness: Assessing disparity when protected class is unobserved », *in Proceedings of the conference on fairness, accountability, and transparency*, pp. 339-348.

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*, pp. 214-226.

Hand D. J. and Henley W. E. (1993), « Can reject inference ever work? », *IMA Journal of Management Mathematics*, 5(1), pp. 45-55.

Kolm S.-C. (1971), *Justice et équité*, Paris, Cepremap, Réédition CNRS (1972).

Kusner M. J., Loftus J., Russell C., and Silva R. (2017), « Counterfactual fairness », *Advances in neural information processing systems*, 30.

Lee M. S. A., Floridi L., and Singh J. (2021), « Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics », *AI Ethics*, 1, pp. 529-544.

Mehrabi N., Morstatter F., Saxena N., Lerman K., and Galstyan A. (2021), « A survey on bias and fairness in machine learning », *ACM Computing Surveys (CSUR)*, 54(6), pp. 1-35.

Mitchell S., Potash E., Barocas S., D'Amour A., and Lum K. (2021), « Algorithmic Fairness: Choices, Assumptions, and Definitions », *Annual Review of Statistics and Its Application*, 8, pp. 141-163.

Molnar C. (2022), I*nterpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.), https://christophm.github.io/interpretable-ml-book.

Najibi A. (2020), « Racial discrimination in face recognition technology », *Harvard Online: Science Policy and Social Justice*, 24, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

O'Neil C. (2016), *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books. Traduction française, 2018, *Algorithmes : la bombe à retardement*, Paris, Les Arènes.

Noble S. U. (2018), *Algorithms of oppression : How Search Engines Reinforce Racism*, New York University Press.

OCDE (2019), « Recommandation du Conseil sur l'intelligence artificielle », https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449.

Pariser E. (2011), *The filter bubble: What the Internet is hiding from you*, Penguin UK.

Pessach D. and Shmueli E. (2022), « A Review on Fairness in Machine Learning », *ACM Computing Surveys (CSUR)*, 55(3), pp. 1-44, https://doi.org/10.1145/3494672.

Rawls J. (1971), *A Theory of Justice*, Harvard University Press.

Rawls J. (2001), *Justice as Fairness: a Restatement*, Harvard University Press.

Rudin C. (2019), « Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead », *Nature Machine Intelligence*, 1(5), pp. 206-215.

Rudin C., Wang C., and Coker B. (2020), « The Age of Secrecy and Unfairness in Recidivism Prediction », *Harvard Data Science Review*, 2(1).

Su J., Vargas D. V., and Sakurai K. (2019), « One pixel attack for fooling deep neural networks », *IEEE Transactions on Evolutionary Computation*, 23(5), pp. 828-841.

Stinson C. (2022), « Algorithms are not neutral », *AI Ethics*, 2, pp. 763-770, https://doi.org/10.1007/s43681-022-00136-w.

Srinivasan R. and Chander A. (2021), « Biases in AI Systems: A survey for practitioners », *Queue*, 19(2), pp. 45-64.

Tsamados A., Aggarwal N., Cowls J. *et al*. (2022), « The ethics of algorithms: key problems and solutions », *AI & Society*, 37(1), pp. 215-230.

Tversky A. and Kahneman D. (1974), « Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty », *Science*, 185(4157), pp. 1124-1131.

UE (2019), « Lignes directrices en matière d'éthique pour une IA digne de confiance », https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60427.

UNESCO (2021), « Recommandation sur l'éthique de l'intelligence artificielle », https://unesdoc.unesco.org/ark:/48223/pf0000380455_fre.

Verma S. and Rubin J. (2018), « Fairness definitions explained », *in 2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1-7.

Wang C., Han B., Patel B., and Rudin C. (2022), « In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction », *Journal of Quantitative Criminology*, pp. 1-63.

Žliobaitė I. and Custers B. (2016), « Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models », *Artificial Intelligence and Law*, 24(2), pp. 183-201.

Conformité européenne des systèmes d'IA : outils statistiques élémentaires

Philippe BESSE[1]
Université de Toulouse – INSA, Institut de Mathématiques – UMR CNRS 5219,
Université Laval –OBVIA

TITLE
European Compliance of AI Systems: Basic Statistical Tools

RÉSUMÉ
Suite à la publication du livre blanc pour une approche de l'IA basée sur l'excellence et la confiance, la Commission Européenne (CE) a publié de nombreuses propositions de textes réglementaires dont un AI Act (CE, 2021) établissant des règles harmonisées sur l'intelligence artificielle (IA). Quels seront les conséquences et impacts de l'adoption à venir de ce texte du point de vue d'un mathématicien ou plutôt statisticien impliqué dans la conception de système d'intelligence artificielle (IA) à haut risque au sens de la CE ? Quels outils et méthodes permettent de répondre aux obligations à venir de conformité : analyse rigoureuse et documentée des données traitées, des performances, robustesse, résilience de l'algorithme, de son explicabilité, des risques pour les droits fondamentaux, de biais discriminatoires ? Ces questions sont illustrées par un exemple numérique analogue à un score de crédit (cf. tutoriel) à la recherche d'un moins mauvais compromis entre toutes les contraintes. Nous concluons sur les avancées et limites du projet de règlement pour les systèmes d'IA à haut risque.

Mots-clés **:** *intelligence artificielle, apprentissage automatique, discrimination, effet disproportionné, AI Act, réglementation européenne.*

ABSTRACT
Following the publication of the white paper for an excellence and trust-based approach to AI, the European Commission (EC) has published numerous regulatory proposals including an AI Act (EC, 2021) establishing harmonized rules on artificial intelligence (AI). What will be the consequences and impacts of the upcoming adoption of this text from the point of view of a mathematician or rather a statistician involved in the design of high-risk AI systems as defined by the EC? What tools and methods can be used to reach future compliance obligations? Rigorous and documented analysis of the data and performance, robustness, resilience of the algorithm, its explicability, and the risks of discriminatory bias for fundamental rights. These questions are illustrated by a numerical example analogous to a credit score (cf. tutorial) in search of the least bad compromise between all the constraints. We conclude on the advances and limitations of the proposed regulation for high risk AI systems.

Keywords**:** *artificial intelligence, machine learning, bias, discrimination, disparate impact, AI Act, european regulation.*

1. philippe.besse@insa-toulouse.fr

1. Introduction

L'adoption en 2018 du Règlement Général de la Protection des Données (RGPD) a profondément modifié les comportements et pratiques des entreprises dans leurs gestions des données, messageries et sites internet. Néanmoins, les condamnations récurrentes des principaux acteurs du numérique, notamment pour abus de position dominante, apportent les preuves de l'inutilité des chartes (*softlaw*) et résolutions éthiques (*ethical washing*). En conséquence et résistant aux accusations fallacieuses de freiner la recherche, l'Europe poursuit sa démarche visant à harmoniser réglementations et innovations technologiques pour le respect des droits humains fondamentaux, mais aussi la défense des intérêts commerciaux de l'Union.

La publication par la Commission Européenne (CE) d'un livre blanc sur l'*Intelligence Artificielle : une approche européenne axée sur l'excellence et la confiance*[2] (CE, 2020) fait suite au guide pour une IA digne de confiance rédigé par un groupe d'experts (CE, 2019). L'étape suivante est la publication de propositions de règlements dont certains en cours d'adoption :

- *Digital Market Act*[3] (2020) : recherche d'équité dans les relations commerciales et risques d'entraves à la concurrence à l'encontre des entreprises européennes ;
- *Digital Services Act*[4] (2020) : sites de service intermédiaire, d'hébergement, de plateforme en ligne et autres réseaux sociaux ; comment contrôler les contenus illicites et risques des outils automatiques de modération ;
- *Data Governance Act*[5] (2020) : contractualisation des utilisations, réutilisations, des bases de données tant publiques que privées (fiducie des données) ;
- *Artificial Intelligence Act*[6] (CE, 2021) : proposition de règlement établissant des règles harmonisées sur l'intelligence artificielle.

S'ajoutant au RGPD pour la protection des données à caractère personnel, l'adoption européenne à venir de ce dernier texte (*AI Act*) va profondément impacter les conditions de développements et d'exploitations des systèmes d'Intelligence Artificielle (systèmes d'IA). Cette démarche fait passer d'une IA souhaitée éthique (*ethical AI*), à une *obligation de conformité* (*lawfull AI*) qui confère le marquage « CE » ouvrant l'accès au marché européen. La CE veut ainsi manifester son *leadership* normatif à l'international afin que ce pouvoir de l'UE sur la réglementation et le marché lui confère un avantage concurrentiel dans le domaine de l'IA.

En conséquence, le présent document propose une réflexion sur la prise en compte méthodologique de ce projet de réglementation concernant plus spécifiquement les compétences usuelles en Statistique, Mathématiques, des équipes de développement d'un système d'IA, notamment ceux jugés à haut risque selon les critères européens. Il cible plus particulièrement certaines des sept exigences citées dans le guide des experts (CE, 2019), reprises dans le livre blanc (CE, 2020) et identifiées comme risques potentiels (Besse *et al.*, 2019) : 1. confidentialité et analyse des données ; 2. précision, robustesse, résilience ; 3. explicabilité ; 4. non-discrimination.

La section 2 suivante extrait de l'*AI Act* les éléments clefs impactant les choix et développements méthodologiques, puis la section 3 en commente les conséquences tout en proposant les outils statistiques bien connus de niveau Master et bagage d'un futur *scientifique des données*. Ceux-ci semblent adaptés voire suffisants pour satisfaire aux futures obligations réglementaires de contrôle des risques afférents aux systèmes d'IA. Enfin, la section 4 déroule un cas d'usage numérique analogue à la prévision d'un score de crédit sur un jeu de données concret. Cet exemple, extrait d'un

2. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_fr.pdf
3. https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52020PC0842&from=fr
4. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment{_}fr
5. https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:52020PC0767&from=FR
6. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

tutoriel[7] dont le code est librement accessible, permet d'illustrer la démarche de recherche d'un moins mauvais compromis à élaborer entre confidentialité, performance, explicabilité et sources de discrimination. Il souligne les difficultés soulevées par la rédaction de la documentation qui devra accompagner tout système d'IA à haut risque. En conclusion, nous proposons une synthèse des principales avancées de ce projet d'*AI Act* et en relevons, dans la version d'avril 2021, les principales limites.

2. Impacts techniques de l'*AI Act*

2.1. Structure du projet de règlement

Castets-Renard et Besse (2022) détaillent une analyse du régime de responsabilité *ex ante*[8] proposé dans les 89 considérants[9] et 85 articles structurés en 12 titres de l'*AI Act* : entre auto-régulation, certification, normalisation, pour définir des règles de conformité notamment pour la défense des droits fondamentaux. L'objectif du présent article est plus spécifique, il est focalisé sur les éléments du projet de réglementation concernant directement le statisticien ou scientifique des données impliqué dans la conception d'un système d'IA jugé à haut risque car impactant des personnes physiques.

De façon générale, les considérants, introductifs au projet, listent donc les principes retenus par la CE et qui ont prévalu à la rédaction des articles. La CE insiste sur la nécessité de la construction de normes internationales en priorisant le respect des droits fondamentaux dont la non-discrimination. Consciente de la place occupée par les algorithmes d'apprentissage statistiques, elle souligne la nécessité de la représentativité statistique des données d'entraînement et l'importance d'une documentation exhaustive à propos de ces données et des performances d'un système d'IA. Consciente également de l'opacité de ces algorithmes, elle demande que les capacités d'interprétation de leurs sorties ou décisions en découlant soient à jour des recherches scientifiques en cours et qu'un suivi puisse être assuré grâce à une journalisation ou archivage des décisions et données afférentes.

2.2. Articles les plus concernés

La définition adoptée de l'IA (art. 3) est pragmatique et très flexible en se basant sur la liste exhaustive des algorithmes concernés (annexe I). Les algorithmes d'apprentissage automatique supervisés ou non, par renforcement, constituent actuellement l'essentiel des applications quotidiennes de l'IA. La représentation de connaissances, la programmation inductive et plus généralement les systèmes experts très développés dans les années 70s, restent présents dans certains domaines. Le troisième type d'algorithme cité cible les approches statistiques, inférences bayésiennes et méthodes d'optimisation. Les approches statistiques bayésiennes ou non conduisant très généralement à des prévisions pour l'aide à la décision peuvent être incluses dans la grande famille de l'apprentissage fondée sur des données. En revanche, les méthodes d'optimisation comme, par exemple, celles d'allocation optimale de ressources des sites d'intermédiation (*e.g.* Uber, ParcourSup, ...) nécessitent une approche particulière. Cette liste peut être facilement adaptée en fonction des évolutions technologiques. Ces définitions reconnaissent la place prépondérante de l'apprentissage statistique et donc des données exploitées pour leur construction. Ils laissent de côté les algorithmes procéduraux basés sur les règles logiques d'une législation comme par exemple ceux présidant aux calculs des montants d'allocations.

7. https://github.com/wikistat/Fair-ML-4-Ethical-AI
8. Par opposition à *ex-post*, *ex-ante* signifie ici que l'analyse ou audit de conformité d'un algorithme d'IA afin de valider sa certification (marquage "CE") est considérée ou effectivement réalisée *avant* sa diffusion ou commercialisation, et donc avant sa mise en exploitation.
9. Les considérants sont une liste de principes qui motivent un décret, une loi ou un règlement, et qui en précèdent le texte contenu dans la liste des articles.

Les articles 5 et 6 adoptent également le principe de définitions pragmatiques en listant explicitement les applications prohibées (art. 5) et celles à haut risque de l'IA facilement adaptables en fonction des évolutions technologiques. L'article 6 fait la différence entre les systèmes faisant déjà l'objet d'une réglementation européenne (annexe II : systèmes de transports et de soins) qui nécessitent une certification *ex-ante* par un tiers, organisme de notification, contrairement aux autres (annexe III) impactant également des personnes physiques mais dont le processus de mise en conformité est seulement déclaratif. *Attention* : la consultation attentive de ces annexes, de leur évolution, est importante pour bien distinguer les systèmes à haut risque des autres. Les scores de crédit bancaire sont concernés (cf. exemple numérique section 4) ainsi que les évaluations *individuelles* de « police prédictive » ou les scores de récidive (justice), mais pas explicitement celles concernant des évaluations de risques de délits par bloc géographique telles *Predpol* ou *Paved* en France. Pour les applications dans le domaine de la justice, seuls sont concernés les systèmes d'IA à l'usage des autorités judiciaires (magistrats) tels le projet abandonné DataJust[10], mais pas ceux à l'usage des cabinets d'avocats (*e.g. case law analytics*[11]).

L'article 10 est fondamental ; il insiste sur l'importance d'une exploration statistique préalable exhaustive des données avant de lancer les procédures largement automatiques d'apprentissage et optimisation. Il évite une forme d'hypocrisie en autorisant, sous réserve de précautions avancées pour la confidentialité, la constitution de bases de données personnelles sensibles permettant par exemple des statistiques ethniques. Cela autorise la mesure directe des biais statistiques, sources potentielles de discrimination.

L'article 11 impose la rédaction d'une documentation qui est essentielle pour ouvrir la possibilité d'audit *ex-ante* d'un système d'IA à haut risque relevant de l'annexe II ou celui d'un contrôle *ex-post* pour ceux relevant de l'annexe III. Avec un reversement de la charge de preuve, c'est au concepteur de montrer qu'il a mis en œuve ce qu'il était techniquement possible en matière de sécurité, qualité, explicabilité, non-discrimination, pour atteindre les objectifs attendus de conformité.

L'article 12 impose un archivage ou journalisation du fonctionnement d'un système d'IA à haut risque. Cette obligation est nouvelle par rapport aux textes européens précédents. Elle est indispensable pour assurer le suivi des mesures de performances, de risques et donc pour être capable de détecter des failles nécessitant des mises à jour voire un ré-entraînement du système ou même son arrêt. Les conditions d'archivage sont précisées dans l'article 61 (*post-market monitoring*).

Selon l'article 13, un utilisateur devrait pouvoir interpréter les sorties, et doit être clairement informé des performances, éventuellement en fonction des groupes concernés, ainsi que des risques notamment de biais et donc de discrimination. Il s'agit ici d'un point sensible directement dépendant de la complexité des systèmes d'IA à base d'algorithmes sophistiqués, donc opaques d'apprentissage statistique. Le choix des métriques de biais est laissé à l'initiative du concepteur. De plus, le manque de recul sur les recherches en cours en matière d'explicabilité d'une décision algorithmique laisse beaucoup de latitude à l'interprétation de cet article qui devra être adaptée à l'évolution des recherches très actives sur ce thème. L'article 14 complète ces dispositions en imposant une surveillance humaine visant à prévenir ou minimiser les risques pour la santé, la sécurité ou les droits fondamentaux.

L'article 15 comble une lacune importante par l'obligation de déclaration des performances (précisions, robustesse, résilience) d'un système d'IA à haut risque. Il concerne également les algorithmes d'apprentissage par renforcement soumis à des risques spécifiques : dérives potentielles (biais) et attaques malveillantes (cybersécurité) comme ce fut le cas pour le *chatbot Tay*[12] de *Microsoft*.

Les articles des chapitres suivants du Titre III notifient des obligations sans apporter de précisions techniques ou méthodologiques : obligations faites au fournisseur (art. 16), obligation de mise en

10. https://www.justice.fr/donnees-personnelles/datajust
11. https://www.caselawanalytics.com/
12. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

place d'un système de gestion de la qualité (art. 17), notamment de toute la procédure de gestion des données de la collecte initiale à leurs mises à jour en exploitation, ainsi que de la maintenance post-commercialisation ; obligation de documentation technique (art. 18), d'évaluation de la conformité (art. 19), obligation des utilisateurs (art. 29), ...

Les États membres sont, par ailleurs, invités à désigner une autorité notifiante comme responsable du suivi des procédures relatives aux systèmes à haut risque et un organisme notifié (art. 30 à 39) indépendant, tout à fait classique des mécanismes de certification déjà en œuvre. Un marquage « CE » sera délivré aux systèmes conformes (art. 49).

Ce processus de marquage « CE » est essentiel pour les systèmes d'IA à haut risque de l'annexe II ; il repose sur un audit *ex-ante* requérant, dans le cas d'une évaluation externe, des compétences très élaborées de la part de l'organisme qui en porte la responsabilité afin d'être à même de pouvoir déceler des manquements intentionnels ou non. Sans évaluation externe, pour les systèmes d'IA de l'annexe II, c'est à l'utilisateur de prendre ses responsabilités vis-à-vis du respect, entre autres, des droits fondamentaux afin de pouvoir faire face à un contrôle si l'État membre désigne une autorité compétente à ce sujet et lui en fournit les moyens.

2.3. Conséquences

L'analyse de ces quelques articles amène des commentaires ou questions, notamment sous le prisme d'une approche mathématique ou statistique de conception d'un système d'IA.

Projet Le projet de règlement (*AI Act*) entre dans un long processus (3 ou 4 ans comme le RGPD ?) de maturation avant une adoption européenne et une application par les États membres. Les amendements à venir devront être successivement pris en considération pour en analyser les conséquences en espérant que des réponses, précisions, corrections, seront apportées aux points ci-dessous. Néanmoins et compte tenu des temps et coûts de conception d'un système d'IA, il est important d'anticiper dès maintenant l'adoption de ce cadre réglementaire.

Exigences essentielles À la suite du guide des experts, le livre blanc appelle à satisfaire sept *exigences essentielles* dont celles de non discrimination et équité, bien être sociétal et environnemental.

Environnement La prise en compte de l'impact environnemental reste anecdotique, simplement évoquée dans les considérants (28) et (81), puis l'article 69 (*codes de conduite*) 2., sans aucune obligation formelle de calculer une balance bénéfices / risques (environnementaux ou autres) d'un système d'IA. Ainsi, l'obligation de l'archivage des données de fonctionnement d'un système d'IA génère un coût environnemental qui mériterait d'être pris en compte dans les risques afférents à son déploiement au regard de son utilité.

Équité La demande exprimée qu'un système d'IA satisfasse au respect des droits fondamentaux en référence à la charte de l'UE, notamment celui de non-discrimination, est très présente dans le livre blanc (cité 16 fois), comme dans les considérants (15, 17, 28, 39) de la proposition de règlement. En revanche, ce principe n'apparaît plus explicitement dans les articles. Est-ce sa présence dans des textes de plus haut niveau comme la Charte des Droits Fondamentaux de l'UE qui n'a pas justifié ici une répétition ou encore un manque d'harmonisation entre les États membres à ce propos ? Il n'y a donc pas de précision sur les façons de « mesurer » une discrimination ou la nécessité de l'atténuer. En revanche, les recherches et documentations des biais potentiels sont clairement explicitées.

Normes Le considérant (13) appelle à la définition de normes internationales notamment à propos des droits fondamentaux. En l'absence d'une définition juridique de l'équité d'un algorithme, celle-ci est définie en creux par l'*absence de discrimination* interdite explicitement. Le souci est

que la littérature regorge de dizaines de définitions de biais statistiques pouvant être à l'origine de sources de discrimination ; lesquels considérer en priorité ? Il est peu probable que les autorités compétentes se prononcent à ce sujet ; elles se focalisent (LNE, 2021) sur les mesures de performances des systèmes d'IA de l'annexe II, notamment les systèmes de transport et les dispositifs de santé en vue de leur certification (marquage « CE »).

Néanmoins, la recherche d'un biais systémique ou de société est requise dans l'analyse préalable des données (art. 10, 2. (f)), ainsi que l'obligation de détailler les performances (précision) par groupe ou sous-groupe d'un système d'IA (art. 13, 3., (b) iv). Ceci permet de prendre en compte certains types de biais, donc de discriminations spécifiques, même en l'absence de définitions normatives. Des indicateurs statistiques de biais devenus relativement consensuels dans la communauté académique sont proposés dans la section suivante.

En revanche, il est regrettable qu'aucune indication, recommandation, contrainte, ne vienne ensuite préciser ce qui pourrait ou devrait être fait pour atténuer ou éliminer un biais discriminatoire. Ceci est laissé au libre arbitre du concepteur d'un système d'IA en espérant que les choix opérés soient explicitement détaillés en toute transparence pour le fournisseur qui en assume la responsabilité et pour l'utilisateur en relation avec les usagers. L'exemple numérique illustre une telle démarche.

Utilisateur & Usager Le règlement traite en priorité les considérations commerciales, donc des risques de défaillance inhérents de l'acquisition des données à la mise en exploitation d'un système d'IA. Tout système doit satisfaire aux exigences de performance annoncées selon un principe de sécurité des produits ou responsabilité du fait des produits défectueux. En revanche, l'usager final, les dommages auxquels il peut être confronté, ne sont pas du tout pris en compte. L'obligation d'information (art. 13) est ainsi au profit de l'utilisateur et pas à celui de l'usager, personne physique impactée, qui ne semble donc protégé à ce jour que par les seules obligations de l'article 22 du RGPD. Il est informé de l'usage d'un système d'IA le concernant, il peut en contester la décision auprès de l'utilisateur humain mais l'explication de la décision, des risques encourus, sont soumises aux compétences et à la déontologie professionnelle de cet utilisateur : conseiller financier pour un client, magistrat pour un justiciable, responsable des ressources humaines pour un candidat, à moins d'un cadre juridique spécifique (*e.g.* code de santé public).

Données Le règlement reconnaît le rôle prépondérant des algorithmes d'apprentissage automatique et donc de la nécessité absolue (considérant 44) de qualité et pertinence des données conduisant à leur entraînement. L'article 10 impose en conséquence des compétences en Statistique pour conduire les études préalables à l'entraînement d'un algorithme. Nous assistons à un renversement de tendance, un retour de balancier, du tout automatique à une approche raisonnée sous responsabilité humaine de cette phase d'analyse des données longue et coûteuse, mais classique du métier de statisticien.

Responsabilité De façon générale, l'objectif essentiel n'est plus la performance absolue comme dans les concours de type *Kaggle* et conduisant à des empilements inextricables d'algorithmes opaques, mais de satisfaire à un ensemble de contraintes pour la mise en conformité, dont celle de transparence, sous la responsabilité du fournisseur du système d'IA. L'analyse des responsabilités en cas de défaillance ou de produit défectueux sera l'objet d'un autre texte.

Documentation Tous les choix opérés lors de la conception d'un système d'IA – ensembles de données, algorithmes, procédures d'apprentissage et de tests, optimisations des paramètres, compromis entre confidentialité, performances, interprétabilité, biais... – doivent (art. 11 et annexe IV) être explicitement documentés en vue d'un audit *ex-ante* des systèmes de l'annexe II ou d'un contrôle *ex-post* d'un système de l'annexe III. C'est un renversement de la charge de preuve sous la responsabilité du fournisseur qui doit pouvoir montrer que le concepteur a

mis en œuvre ce qui était techniquement possible pour satisfaire aux obligations (conformité) légales de sécurité, transparence, performances et non discrimination.

Autorité notifiante (Chapitre 4 Titre III) Chaque pays va se doter ou désigner (art. 30) un service chargé entre autres de superviser l'audit *ex-ante* d'un système d'IA à haut risque de l'annexe II avant son déploiement, qu'il soit commercialisé ou non. L'autorité notifiante désigne l'*organisme de notification* qui exécutera l'audit. De façon assez étonnante, un système d'ascenseur élémentaire, n'embarquant qu'une « IA » logique rudimentaire mais dépendant de l'annexe II, est plus contraint par l'obligation de certification par un organisme tiers, au contraire d'applications des systèmes d'IA de l'annexe III (justice, emploi, crédit...) impactant directement des personnes physiques avec des risques réels envers les droits fondamentaux. Il faudra donc être attentif à l'interprétation que fera un État membre de cette situation afin d'évaluer les possibilités de saisine et compétences de contrôle d'un système à haut risque de l'annexe III.

Archivage & confidentialité Le règlement cible donc, en première lecture, les obligations commerciales du fournisseur plutôt que celles, étiques ou déontologiques, envers l'usager. Néanmoins, le règlement apporte la possibilité de prendre en compte des données sensibles (art. 10, 5.), les obligations d'archivage des décisions (art. 12), de suivi des performances selon les groupes (art. 13), une surveillance humaine (art. 14) pendant toute la période d'utilisation et de correction rétro-active des biais (art. 15). Cette obligation d'archivage et de surveillance du fonctionnement, notamment à destination des groupes sensibles, oblige implicitement à l'acquisition, en toute sécurité (cryptage, anonymisation, pseudonymisation...), de données confidentielles (*e.g.* origine ethnique). Cela ne rend-il pas indispensable, selon le domaine d'application, la mise en place d'un protocole explicite de consentement libre et éclairé, d'un engagement éthique, entre l'utilisateur et l'usager, protégé par le RGPD ? Comment sont évalués les risques encourus d'un usager ou groupe d'usager par le recueil et l'exploitation de leurs données sensibles lors de l'exploitation d'un système d'IA face aux bénéfices attendus pour eux mêmes ou l'intérêt public ?

3. Prise en compte méthodologique de l'*AI Act*

3.1. Quels algorithmes ?

Dans l'attente d'une adoption effective du texte final qui risque d'être amendé, il est néanmoins prudent, compte tenu des investissements en jeu, d'anticiper des réponses techniques à certaines contraintes ou obligations faites aux systèmes d'IA désignés à haut risque. Cet article laisse volontairement de côté certaines classes d'algorithmes mentionnées ou non dans l'annexe I dont la liste finale reste l'objet de débats entre les instances européennes.

Un système expert est l'association d'une base de règles logiques ou base de connaissances construites par des experts du domaine concerné, d'un moteur d'inférence et d'une base de faits observés pour une exécution en cours. Le moteur d'inférence recherche la séquence de règles logiquement applicables à partir des faits observés de la base qui s'incrémente comme conséquences du déclenchement des règles. Le processus itère jusqu'à l'obtention ou non d'une décision recherchée et expliquée par la séquence de règles y conduisant. Très développée dans les années 70, la recherche a marqué le pas face à un problème dit *NP-complet*, c'est-à-dire de complexité algorithmique exponentielle en la taille de la base de connaissances (nombre de règles). Supplantée par la ré-émergence des réseaux de neurones (années 80), puis plus largement par l'apprentissage automatique, la recherche dans ce domaine dit d'IA symbolique est restée active. Elle connaît un renouveau motivé par les capacités d'explicabilité des systèmes experts.

Les approches statistiques bayésiennes ou non basées sur des données sont associées implicitement aux méthodes par apprentissage. En revanche, les algorithmes d'allocation optimale de ressources prennent une place à part. Si les principes d'allocation en tant que tels ne soulèvent pas de problème, ceux d'ordonnancement ou de tri des ressources peuvent amener des risques réels de discrimination indirecte. C'est notamment le cas de l'algorithme ParcourSup lorsque les établissements d'enseignement supérieur introduisent des pondérations selon le lycée d'origine des candidats : lycée de centre ville *vs.* lycée de banlieue. Cette situation rejoint alors le cas des algorithmes déterministes ou procéduraux. Il s'agit d'algorithmes décisionnels (*e.g.* calcul de taxes, impôts, allocations ou prestations sociales, ...) basés sur un ensemble de règles de décision déterministes qui peuvent tout autant présenter des impacts, désavantages ou risques de discrimination indirecte, malgré une apparente neutralité. La Défenseure des Droits (2020) est très attentive en France à l'analyse et détection de ces risques[13]. Celle-ci relève de l'analyse experte des règles de décision codées dans l'algorithme qui, en l'état, ne sont pas concernées par le projet de règlement. Néanmoins, la complexité de l'algorithme peut être telle qu'une analyse experte *ex-post* ne sera pas en mesure d'évaluer l'étendue des risques indirects. Aussi, un algorithme déterministe complexe peut être analysé avec les mêmes outils statistiques que ceux adaptés à un algorithme d'apprentissage automatique.

Nous insistons donc tout particulièrement sur les systèmes d'IA basés sur des algorithmes d'apprentissage supervisé ou statistique, ou IA empirique, par opposition à l'IA dite symbolique des systèmes experts. Ce sont très majoritairement les plus répandus au sein de ceux désignés à haut risque (art. 6) car susceptibles d'impacter directement des personnes physiques.

Même sans obligation de certification *ex-ante* par un organisme notifié, une documentation exhaustive (art. 11) d'un système d'IA à haut risque doit être produite et fournie à l'utilisateur. Cette section propose quelques indications méthodologiques pour répondre à cette attente.

3.2. Les données

Tout système d'IA basé sur un algorithme d'apprentissage statistique nécessite la mise en place d'une base de données d'entraînement fiable et représentative du domaine d'application visé, qui doit en tout premier lieu satisfaire aux exigences de confidentialité du RGPD. Puis, le travail d'exploration statistique, généralement long et fastidieux, d'acquisition, vérification, analyse, préparation, nettoyage, enrichissement, archivage sécurisé des données, est essentiel à l'élaboration d'un système d'IA performant, robuste, résilient et dont les biais potentiels sont sous contrôle. Construire de nouvelles caractéristiques (*features*) adaptées à l'objectif, traquer et gérer éventuellement par imputation des données manquantes, identifier les anomalies ou valeurs atypiques (*outliers*) sources de défaillances, les sources de biais – classes ou groupes sous représentés, biais systémiques –, nécessitent compétences et expériences avancées en Statistique.

Ces compétences sont indispensables pour répondre aux attentes de l'article 10 ainsi qu'aux besoins de la documentation (annexe IV) imposée par l'article 11.

3.3. Qualité, précision et robustesse

Les articles 13 et 15 imposent clairement de devoir documenter les performances et risques d'erreur, éventuellement en fonction de groupes sensibles et protégés, ou de défaillances d'un système d'IA. Cela rend indispensable l'explicitation de choix, notamment des métriques utilisées.

13. https://www.defenseurdesdroits.fr/fr/rapports/2020/05/algorithmes-prevenir-lautomatisation-des-discriminations

Choix de métrique

L'évaluation de la qualité d'une aide algorithmique à la décision est essentielle à la justification du déploiement d'un système d'IA au regard de sa balance bénéfices / risques. Dans le cas d'un système IA empirique ou par apprentissage automatique, il s'agit d'estimer la *précision* des prévisions, dont les mesures sont bien connues et maîtrisées, partie intégrante du processus d'apprentissage. Néanmoins, parmi un large éventail des possibles, le choix, précisément justifié, doit être adapté au domaine, au type de problème traité, aux risques spécifiques encourus quel que soit le modèle ou le type d'algorithme d'apprentissage utilisé. Citons par exemple les situations de :

– *Régression* ou modélisation, et prévision d'une variable cible Y quantitative. Elle est généralement basée sur l'optimisation d'une mesure quadratique (norme L_2) pouvant intégrer, à l'étape d'entraînement, différents types de pénalisation dont celles de parcimonie (*ridge, Lasso*), afin de contrôler la complexité de l'algorithme et éviter les phénomènes de sur-apprentissage. Un autre type de fonction objectif basée sur une perte en norme L_1 ou valeur absolue, moins sensible à la présence de valeurs atypiques (*outliers*) que la norme quadratique, permet des solutions plus robustes car tolérantes à des observations atypiques.

– *Classification* ou modélisation, prévision d'une variable Y qualitative. Le choix d'une mesure d'erreur doit être opéré parmi de très nombreuses possibilités : taux d'erreur, AUC (*area under the ROC Curve* pour une variable Y binaire), score F_β, risque bayésien, entropie... avec la difficile prise en compte des situations de classes déséquilibrées qui oriente le choix du type de mesure et nécessite des précautions spécifiques dans l'équilibrage de la base d'apprentissage ou les pondérations de la fonction objectif en prenant en compte une matrice de coûts de mauvais classement éventuellement asymétrique.

Limites de la précision

Besse (2021) rappelle que les performances de l'IA sont largement surévaluées par le battage médiatique dont bénéficient ces technologies. Ces performances sont d'autant plus dégradées lorsque la décision concerne la prévision d'un comportement (achat, départ, embauche, acte violent, pathologie...) individuel humain dépendant potentiellement d'un très grand nombre de variables explicatives ou facteurs dont certains peuvent ne pas être observables. Il importe de distinguer les systèmes d'IA développés dans un domaine bien déterminé (*e.g.* process industriel sous-contrôle), où le nombre de facteurs ou dimensions est raisonnable et identifié, des systèmes d'IA où opère le fléau ou malédiction de la dimension (*curse of dimensionality*), lorsque celle-ci est très grande, voire indéterminée.

L'histoire de la littérature statistique puis d'apprentissage automatique peut être lue comme une succession de stratégies pour le contrôle du nombre de variables et ainsi de paramètres estimés dans un modèle statistique ou entraînés dans un algorithme. Il s'agit par exemple de contrôler le conditionnement d'une matrice en régression : PLS (*partial least square*), sélection de variables (critères AIC, BIC), pénalisations *ridge* ou *Lasso*, et ainsi l'explosion de la variance des prévisions. Plus généralement, c'est aussi le contrôle du risque de sur-ajustement qui doit être documenté comme résultat de l'optimisation des hyperparamètres : nombre de plus proches voisins, pénalité en machines à vecteurs supports, nombre de feuilles d'un arbre, de variables tirées aléatoirement dans une forêt d'arbres, profondeur des arbres et nombre d'itérations en *boosting*, structures des couches convolutionnelles et *drop out* des réseaux de neurones en reconnaissance d'images. Même si les stratégies d'optimisation de ces hyperparamètres par validation croisée ou échantillon de validation sont bien rodées, le fléau de la dimension peut s'avérer rédhibitoire (*e.g.* Verzelen, 2012).

Échantillon test

En tout état de cause, il est indispensable de mettre en place une démarche très rigoureuse pour conduire à l'évaluation de la précision et donc des performances d'un système d'IA basé sur un algorithme d'apprentissage. Comme énoncé dans l'article 3, 31., ce sont des *données de test indépendantes* de celles d'apprentissage qui sont utilisées à cet effet. Attention néanmoins d'évaluer les performances sur des données telles qu'elles se présenteront *réellement* en exploitation, avec leurs défauts, et pas un simple sous-ensemble aléatoire de la base d'apprentissage comme c'est trop souvent le cas en recherche académique. En effet, cet ensemble de données peut bénéficier d'une homogénéité d'acquisition (*e.g.* même technologie, même opérateur) et de prétraitements qui peuvent faire défaut à de réelles données d'entrée à venir en exploitation. Cela demande donc une extrême rigueur dans la constitution d'un échantillon test pour éviter ces pièges bien trop présents en recherche académique (*e.g.* Liu *et al.*, 2019 ; Roberts *et al.*, 2021) et conduisant, sous la pression de publication, à beaucoup trop de résultats non reproductibles et des algorithmes non certifiables. Enfin, une surveillance (art. 14) toute la durée de vie du système d'IA est indispensable afin d'en détecter de possibles dérives ou dysfonctionnements (art. 12 et 15) affectant la robustesse ou la résilience des décisions.

Robustesse

L'évaluation de la *robustesse* est liée aux procédures de contrôle mises en place pour détecter des valeurs atypiques (*outliers*) ou anomalies dans la base d'apprentissage et au choix de la fonction perte de la procédure d'entraînement de l'algorithme. Impérativement, surtout dans les applications sensibles pouvant entraîner des risques élevés en cas d'erreur, la détection d'anomalie doit également être intégrée en exploitation afin de ne pas chercher à proposer des décisions correspondant à des situations atypiques, étrangères à la base d'apprentissage.

Résilience

La *résilience* d'un système d'IA est essentielle pour les dispositifs critiques (dispositifs de santé connectés, aide au pilotage). Cela concerne par exemple la prise en compte de données manquantes lors de l'apprentissage comme en exploitation. Il s'agit d'évaluer la capacité d'un système d'IA à assurer des fonctions pouvant s'avérer vitales en cas, par exemple, de panne ou de fonctionnement erratique d'un capteur : choix d'un algorithme tolérant aux données manquantes, imputation de celles-ci, fonctionnement en mode dégradé, alerte et arrêt du système.

3.4. Explicabilité

Une recherche active

Il est bien trop tôt pour tenter un résumé opérationnel de ce thème et fournir des indications claires sur la démarche à adopter pour satisfaire aux exigences réglementaires (art. 13, 15). Il faut pour cela attendre que la recherche ait progressé et qu'une sélection « naturelle » en extrait les procédures les plus pertinentes parmi une grande quantité de solutions proposées ; un article de revue sur ce sujet (Barredo Arrieta *et al.*, 2020) listait plus de 400 références.

Arbre de choix

Tentons de décrire les premiers embranchements d'un arbre de décision en répondant à quelques questions rudimentaires qu'il faudrait en plus adapter au domaine d'application, car le type de réponse à apporter n'est évidemment pas le même s'il s'agit d'expliquer le refus d'un prêt ou les conséquences d'une aide automatisée au diagnostic d'un cancer.

Il importe de bien distinguer les niveaux d'explication : concepteur, utilisateur ou usager, même si ce dernier n'est pas directement concerné par le projet de règlement. De plus, l'explication peut s'appliquer soit au fonctionnement général de l'algorithme, soit à une décision spécifique.

Il y a schématiquement deux types d'algorithmes dont ceux relativement transparents : modèles linéaires et arbres de décision. L'explication est dans ce cas possible à condition que le nombre de variables et d'interactions prises en compte, ou le nombre de feuilles d'un arbre, reste raisonnable. Toutes les autres classes d'algorithme d'apprentissage, systématiquement non linéaires et complexes, sont par construction opaques. Il s'agit alors de construire une explication par différentes stratégies comme une approximation explicable par un modèle linéaire, un arbre ou un ensemble de règles de décision déterministes. Une autre stratégie consiste à fournir des indications sur l'*importance des variables* en mesurant l'effet d'une permutation aléatoire de leurs valeurs (*mean decrease accuracy* : Breiman, 2001), en stressant l'algorithme (Bachoc *et al.*, 2020) ou en réalisant une analyse de sensibilité par indices de Sobol (Bénesse *et al.*, 2021).

Le concepteur d'un algorithme s'intéresse également à l'explication d'une décision spécifique afin d'identifier la cause d'une erreur, y remédier par exemple en complétant la base d'apprentissage d'un groupe sous-représenté avant de ré-entraîner l'algorithme. L'utilisateur d'un système d'IA doit être au mieux informé (art. 13, 15) des possibilités d'expliquer une décision qu'il pourra retranscrire à l'usager (client, patient, justiciable, citoyen...) selon sa propre déontologie, son intérêt commercial ou une contrainte légale par exemple pour des décisions administratives. Pour ce faire quelques stratégies sont proposées comme une *approximation locale* par un modèle explicable (linéaire, arbre de décision) ou par une liste d'exemples *contrefactuels*, c'est-à-dire des situations les plus proches, en un certain sens, qui conduiraient à décision contraire, généralement plus favorable (attribution d'un prêt). Lorsque cela s'avère impossible, comme par exemple dans le cas d'un diagnostic médical impliquant un nombre important de facteurs opaques, il importe d'informer précisément l'utilisateur et donc le patient sur les risques d'erreur afin que le consentement de ce dernier soit effectivement libre et éclairé.

Quelques démonstrations de procédures explicatives sont proposées sur des sites en accès libre[14].

Réalité complexe

Ne pas perdre de vue que l'impossibilité ou simplement la difficulté à formuler une explication provient certes de l'utilisation d'algorithmes opaques, mais dont la nécessité est inhérente à la complexité même du réel. Un réel complexe (*e.g.* les fonctions du vivant) impliquant de nombreuses variables, leurs interactions, des effets non linéaires voire des boucles de contre-réaction, est nécessairement modélisé par un algorithme complexe afin d'éviter des simplifications abusives pouvant gravement nuire aux performances. C'est tout d'abord le réel qui s'avère complexe à expliquer.

3.5. Biais & discrimination

Bien que très présente dans les textes préliminaires (livre blanc (CE, 2021), considérants de l'AI Act), la référence au risque de discrimination ne l'est pas de façon explicite dans les projets d'articles. Ap-

14. Citons : gems-ai.com, aix360.mybluemix.net, github.com/MAIF/shapash

parait néanmoins l'obligation de détecter des biais dans les données (art. 10) ainsi que celle d'afficher des performances ou risques d'erreur par groupe (art. 13). Quelles en sont les conséquences au regard des difficultés de définir, détecter une discrimination, qu'elle soit humaine ou algorithmique ?

Détecter une discrimination

Formellement, la stricte équité peut s'exprimer par des propriétés d'indépendance en probabilité entre la variable cible Y qui exprime une décision et la variable dite sensible S par rapport à laquelle une discrimination est en principe interdite. Cette variable peut être quantitative (*e.g.* âge) ou qualitative à deux ou plusieurs classes (*e.g.* genre ou origine ethnique) ou, de façon plus complexe, la prise en compte d'interactions entre plusieurs variables sensibles. Néanmoins, cette définition théorique de l'équité n'est pas concrètement praticable pour détecter, mesurer, atténuer des risques de biais. De plus, les textes juridiques font essentiellement référence à un groupe de personnes sensibles par rapport aux autres. En conséquence, et pour simplifier cette première lecture pédagogique de la détection des risques de discrimination, nous ne considérons qu'une variable sensible à 2 modalités : jeune *vs.* vieux, femme *vs.* homme, ...

Une façon bien établie de détecter une décision humaine discriminatoire consiste à opérer par *testing*. Dans le cas d'une présomption de discrimination à l'embauche, la procédure consiste à adresser deux CV comparables, à l'exception (*counterfactual example*) de la modalité de la variable sensible (*e.g.* genre, origine ethnique associée au nom) afin de comparer les réponses : proposition ou non d'entretien. Cette démarche individuelle est rendue systématique (Rich, 2014) dans une enquête par l'envoi de milliers de paires de CV. C'est en France la doctrine officielle promue par le Comité National de l'Information Statistique[15] et commanditée périodiquement par la DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques) du Ministère du travail.

Des indicateurs statistiques peuvent être estimés à l'issue de cette enquête mais, comme il n'existe pas de définition juridique de l'équité, qui devient par défaut l'absence de discrimination, le monde académique a proposé quelques dizaines d'indicateurs (*e.g.* Zliobaitė, 2017) afin d'évaluer des biais potentiels, sources de discrimination. Il est nécessaire d'opérer des choix parmi tous les critères de biais en remarquant que beaucoup de ces indicateurs s'avèrent être très corrélés ou redondants (Friedler *et al.*, 2019). Empiriquement et après avoir consulté une vaste littérature sur l'IA éthique ou plutôt sur les risques identifiés de discrimination algorithmique, un consensus émerge sur le choix en priorité de trois niveaux de biais statistique. Sont finalement considérés dans cet article élémentaire trois types de rapports de probabilités (égaux à 1 en cas d'indépendance stricte) dont Besse *et al.* (2021) proposent des estimations par intervalle de confiance afin d'en contrôler la précision.

Parité statistique et effet disproportionné

Le premier niveau de risque de discrimination algorithmique s'illustre simplement : si un algorithme est entraîné sur des données biaisées, il apprend et reproduit très fidèlement ces biais systémiques, de société ou de population, par lesquels un groupe est historiquement désavantagé (*e.g.* revenu des femmes) ; plus grave, l'algorithme risque même de renforcer le biais en conduisant à des décisions explicitement discriminatoires. Il importe donc de pouvoir détecter, mesurer, atténuer voire éliminer ce type de biais. L'équité ou parité statistique (ou *demographic equality*) serait l'indépendance entre la ou les variables sensibles S (*e.g.* genre, origine ethnique) et la variable de prévision $\widehat{Y}$ de la décision. Historiquement, l'écart à l'indépendance pour mesurer ce type de biais est évalué aux USA dans les procédures d'embauche depuis 1971 par la notion d'effet disproportionné ou *disparate impact* et maintenant reprise systématiquement (Barocas et Selbst, 2016) pour l'évaluation de ce type

15. https://www.cnis.fr/wp-content/uploads/2018/03/Chroniques14.pdf

de discrimination dans un algorithme. L'effet disproportionné consiste à estimer le rapport de deux probabilités : probabilité d'une décision favorable ($\widehat{Y} = 1$) pour une personne du groupe sensible ($S = 0$) au sens de la loi sur la même probabilité pour une personne de l'autre groupe ($S = 1$) :

$$DI = \frac{\mathbb{P}(\widehat{Y} = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 1)}.$$

Cet indicateur est intégré au *Civil Rights act & Code of Federal Regulations (Title 29, Labor : Part 1607 Uniform guidelines on employee selection procedures)*[16] depuis 1978 avec la règle dite des 4/5ème ; si DI est inférieur à $0,8$, l'entreprise doit en apporter les justifications économiques. Les logiciels commercialisés aux USA et proposant des algorithmes de pré-recrutement automatique anticipent ce risque juridique (Raghavan *et al.*, 2019) en intégrant une procédure automatique d'atténuation du biais (*fair learning*). Il n'y a aucune obligation ni mention en France de cet indicateur statistique, seulement une incitation de la part de la Défenseure des Droits et de la CNIL (2012) envers les services de ressources humaines des entreprises. Il leur est suggéré de tenir des statistiques ethniques, autorisées dans ce cas sous réserve de confidentialité, sous la forme de tables de contingence dont il serait facile d'en déduire des estimations d'effet disproportionné.

La mise en évidence d'un biais systémique est implicitement citée lors de l'étape d'analyse préliminaire des données (art. 10, 2., (f)) mais sans plus de précision sur la façon dont il doit être pris en compte, alors que renforcer algorithmiquement ce biais serait ouvertement discriminatoire. De plus, serait-il politiquement opportun d'introduire une part de discrimination positive afin d'atténuer la discrimination sociale ? C'est évoqué dans le guide des experts (CE, 2019, ligne directrice 52) pour *améliorer le caractère équitable de la société* et techniquement l'objet d'une vaste littérature académique nommée apprentissage équitable (*fair learning*). Cette opportunité n'est pas reprise explicitement dans l'*AI Act*, mais nous verrons dans l'exemple numérique ci-dessous qu'elle ne peut être exclue et peut même être pleinement justifiée en prenant en considération les autres types de biais ci-après.

Erreurs conditionnelles

Les taux d'erreur de prévision et donc les risques d'erreur de décision sont-ils les mêmes pour chaque groupe (*overall error equality*) ? Autrement dit, l'erreur est-elle indépendante de la variable sensible ? Ceci peut se mesurer par l'estimation (intervalle de confiance) du rapport de probabilités (probabilité de se tromper pour le groupe sensible sur la probabilité de se tromper pour l'autre groupe) :

$$REC = \frac{\mathbb{P}(\widehat{Y} \neq Y|S = 0)}{\mathbb{P}(\widehat{Y} \neq Y|S = 1)}.$$

Ainsi, si un groupe est sous-représenté dans la base d'apprentissage, il est très probable que les décisions le concernant soient moins fiables. C'est une des premières critiques formulées à l'encontre des algorithmes de reconnaissance faciale et ce risque est également présent dans les applications en santé (Besse *et al.*, 2020) ou en ressources humaines (De-Arteaga *et al.*, 2019). L'identification, la prise en compte et la surveillance de ce risque sont présents (art. 13, 3., (b), ii et art. 15, 1. & 2.) dans le projet de règlement et doivent donc être explicitement détaillés dans la documentation (art. 11).

16. https://www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml

Rapports de cote conditionnels

Même si les deux critères précédents sont trouvés équitables, les erreurs peuvent être dissymétriques (plus de faux positifs, moins de faux négatifs) au détriment d'un groupe avec un impact d'autant plus discriminatoire que le taux d'erreur est important. Cet indicateur (comparaison des rapports de cote ou *odds ratio* d'indépendance conditionnelle, nommés aussi *equalli odds*) est au cœur de la controversehttps://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing concernant l'évaluation COMPAS du risque de récidive aux USA (Larson *et al.*, 2016). Il est également présent dans l'exemple numérique ci-après. Cet indicateur double est mesuré par l'estimation de deux rapports de probabilités : rapport du taux de faux positifs du groupe sensible sur le taux de faux positifs de l'autre groupe et rapport des taux de faux négatifs pour ces mêmes groupes :

$$RFP = \frac{\mathbb{P}(\hat{Y} = 1 | Y = 0, S = 0)}{\mathbb{P}(\hat{Y} = 1 | Y = 0, S = 1)} \quad \text{et} \quad RFN = \frac{\mathbb{P}(\widehat{Y} = 0 | Y = 1, S = 0)}{\mathbb{P}(\widehat{Y} = 0 | Y = 1, S = 1)}.$$

L'évaluation de ce type de biais n'est pas explicitement mentionnée dans le projet de règlement. Néanmoins, elle fait partie de la procédure classique d'évaluation des erreurs en classification à l'aide d'une matrice de confusion ou de courbes ROC par groupes, et ne peut être négligée.

Notons qu'il est d'autant plus difficile de faire abstraction du dernier type de biais que les trois sont interdépendants et même en interaction avec les autres risques : précision et explicabilité. Ceci est clairement mis en évidence dans l'exemple numérique suivant. Il y a donc une forme d'obligation déontologique ou de cohérence statistique à devoir appréhender ces différents niveaux d'analyse.

4. Exemple numérique

L'exemple « jouet » ou « bac à sable » de cette section permet d'illustrer concrètement toute la complexité des principes précédemment évoqués en soulignant leur interdépendance. Ce jeu de données est ancien, largement utilisé pour illustrer tous les travaux visant une atténuation optimale du biais. Le monde académique espère avoir rapidement accès à bien d'autres « bacs à sable » représentatifs dont la construction est l'objet de l'article 53 de l'*AI Act*.

4.1. Données

Les données publiques[17] utilisées imitent le contexte du calcul d'un score de crédit. Elles sont extraites (échantillon de 45 000 personnes) d'un recensement de 1994 aux USA et décrivent l'âge, le type d'emploi, le niveau d'éducation, le statut marital, l'origine ethnique, le nombre d'heures travaillées par semaine, la présence ou non d'un enfant, les revenus ou pertes financières, le genre et le niveau de revenu bas ou élevé. Elles servent de référence ou *bac à sable* pour tous les développements d'algorithmes d'apprentissage automatique équitable. Il s'agit de prévoir si le revenu annuel d'une personne est supérieur ou inférieur à 50k$, et donc de prévoir, d'une certaine façon, sa solvabilité connaissant ses autres caractéristiques socio-économiques. Ces questions de discrimination dans l'accès au crédit sont toujours d'actualité (Campisi, 2021[18] ; Hurlin *et al.*, 2021 ; Kozodoi *et al.*, 2021) même si le principe du *score de crédit* s'est généralisé dès les années 90 avec l'envol du *data mining* devenu depuis de l'IA.

L'étude complète et les codes de calcul sont disponibles dans un tutoriel[19], mais l'illustration est limitée à un résumé succinct de l'analyse de la discrimination selon le genre.

17. https://archive.ics.uci.edu/ml/datasets/Adult
18. https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/
19. Calepin *Jupyter* : https://github.com/wikistat/Fair-ML-4-Ethical-AI/blob/master/AdultCensus/AdultCensus-R-biasDetectionCourt.ipynb

4.2. Résultats

Une analyse exploratoire – nettoyage des données, description statistique – préalable doit être incluse dans la documentation. Elle est l'objet d'un autre tutoriel[20] dont seuls quelques résultats sont retenus par souci de concision. Ils mettent en évidence un biais systémique ou de société important : seulement $11,6\%$ des femmes ont un revenu élevé contre $31,5\%$ des hommes. Le rapport $DI = 0,38$ est donc très disproportionné et peut s'expliquer par quelques considérations sociologiques bien identifiées sur le premier plan factoriel (fig. 1) d'une analyse factorielle multiple des correspondances calculée après avoir recodé qualitatives toutes les variables. Les femmes travaillent en moyenne moins d'heures (`HW1`) par semaine (occupations ménagères et enfants ?) ; même si le niveau de diplôme ne semble pas lié au genre, elles occupent un poste avec moins de responsabilités (`Admin`) (effet plafond de verre ?). Un autre type de biais semble présent dans ces données, les femmes sont associées (co-occurrences plus fréquentes que l'indépendance) à la présence d'enfants sans pour autant être en situation de couple contrairement aux hommes. Cette enquête s'adresse-t-elle de façon privilégiée au chef ou à la cheffe de famille éventuellement monoparentale ?

Les données ont été aléatoirement réparties en un échantillon d'apprentissage (29 000), destiné à l'estimation des modèles ou entraînement des algorithmes, un échantillon de validation (8 000) afin d'optimiser certains hyper paramètres et un échantillon de test (8 000) pour évaluer les différents indicateurs de performance et biais. La taille relativement importante de l'échantillon initial permet de considérer un échantillon de validation représentatif, comme demandé dans le règlement, afin d'éviter des procédures plus lourdes de validation croisée. Les résultats de prévision sont regroupés dans la figure 2.

Le biais systémique (`dataBaseBias`) des données est comparé avec celui de la prévision de niveau de revenu par un modèle classique linéaire de régression logistique `linLogit` : $DI = 0,25$. Significativement moins élevé (intervalles de confiance disjoints), il montre que ce modèle renforce le biais et donc discrimine nettement les femmes dans sa prévision. La procédure naïve (`linLogit-w-s`) qui consiste à éliminer la variable dite sensible (genre) du modèle ne supprime en rien ($DI = 0,27$) le biais discriminatoire car le genre est de toute façon présent à travers les valeurs prises par les autres variables (effet *proxy*). Une autre conséquence de cette dépendance aux proxys est que le *testing* ou *counterfactual test* (changement de genre toutes choses égales par ailleurs) ne détecte plus ($DI = 0,90$) aucune discrimination !

Un algorithme non-linéaire élémentaire (`tree`, arbre binaire de décision) augmente le biais mais pas de façon statistiquement significative car les intervalles de confiance ne sont pas disjoints. Sa précision est meilleure que celle du modèle de régression logistique mais, si l'objectif est une interprétation utile, il est nécessaire de réduire la complexité de l'arbre en pénalisant le nombre de feuilles, d'une centaine à une dizaine. Dans ce cas, la précision se dégrade pour rejoindre celle de la régression logistique ; l'explicabilité a un coût.

Un algorithme non linéaire plus sophistiqué (`random forest`) est très fidèle au biais des données avec un indicateur ($DI = 0,36$) proche de celui du biais de société et fournit une meilleure précision : $0,86$ au lieu de $0,84$ pour la régression logistique. Cet algorithme ne discrimine pas plus, apporte une meilleure précision, mais c'est au prix de l'explicabilité du modèle. Opaque comme un réseau de neurones, il ne permet pas d'expliquer une décision à partir de ses paramètres comme cela est facile avec le modèle de régression ou un arbre binaire de décision de taille raisonnable.

Une question délicate concerne le choix politique de procéder ou non à une atténuation du biais systémique dans le cas d'un score de crédit. Contrairement à Hurlin *et al.* (2021), Goglin (2021)[21] l'aborde de façon très incomplète en ne considérant, de manière exclusive, que le biais des erreurs selon le genre. Cet auteur « justifie » de ne pas considérer le biais systémique car le corriger condui-

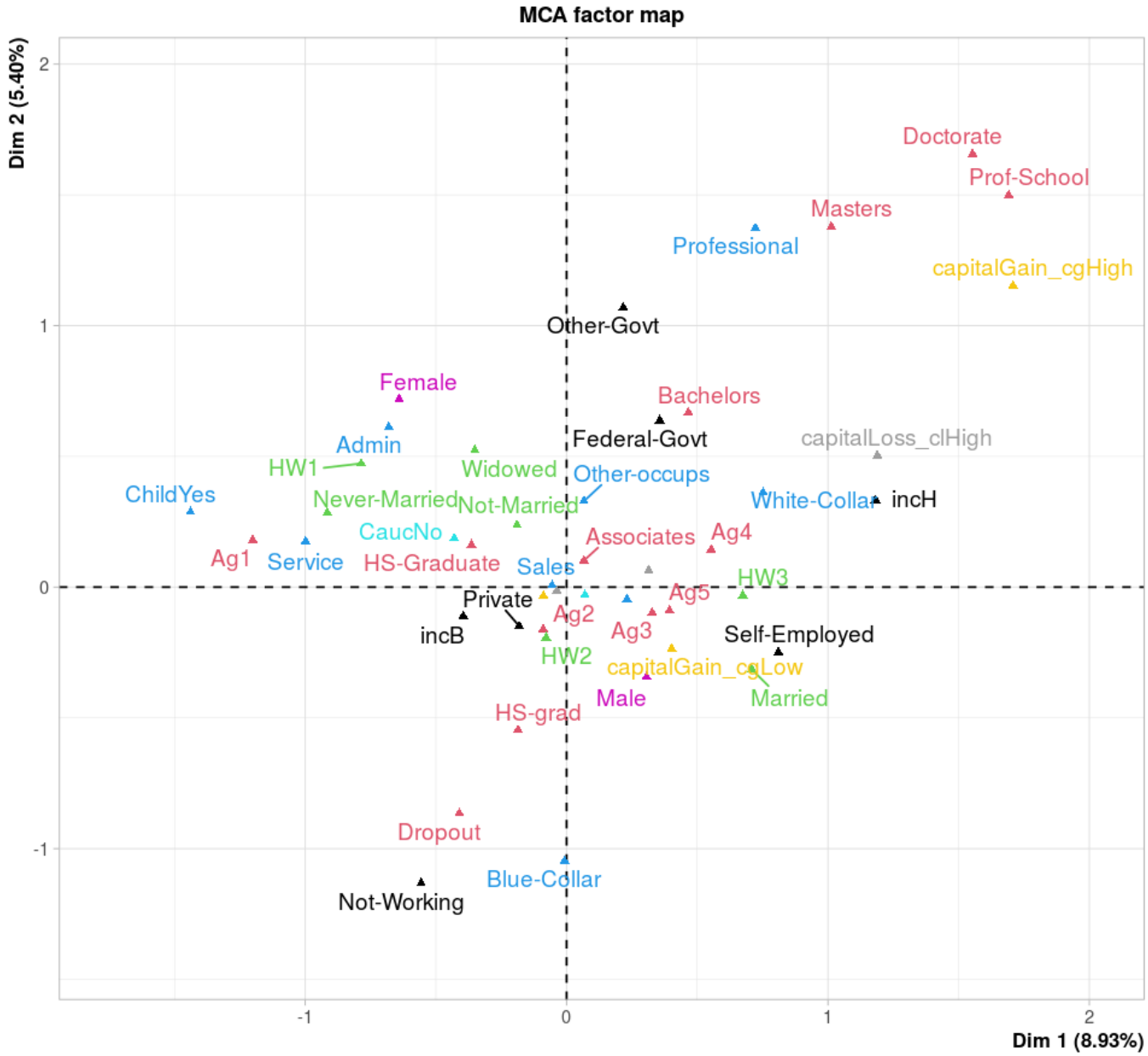


FIGURE 1 : Premier plan factoriel d'une analyse factorielle multiple des correspondances (librairie `FactoMineR`, Lê *et al.*, 2008)

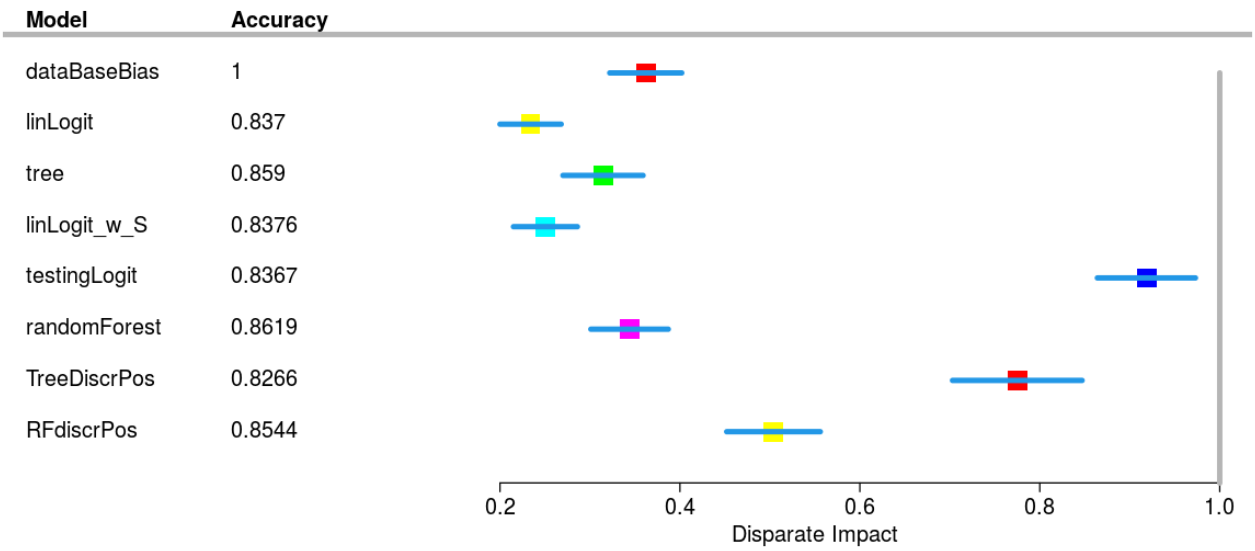


FIGURE 2 : Précision de la prévision (accuracy) et effet disproportionné (discrimination en fonction du genre) estimé par un intervalle de confiance sur un échantillon test (taille 9000) pour différents modèles ou algorithmes d'apprentissage.

rait des femmes à des situations de surendettement tandis que le 3ème type de biais est purement oublié. Une analyse plus fine montre, à travers cet exemple, toute l'importance de prendre en compte simultanément les trois types de biais afin d'éviter un positionnement quelque peu « paternaliste ».

En principe, la précision de la prévision pour un groupe dépend de sa représentativité. Si ce dernier est sous-représenté, l'erreur est plus importante ; c'est typiquement le cas en reconnaissance faciale mais pas dans l'exemple traité. Alors qu'elles sont deux fois moins nombreuses dans l'échantillon, le taux d'erreur de prévision est de l'ordre de $7,9\%$ pour les femmes et de 17% ($REC = 0,36$) pour les hommes (algorithme d'arbre binaire simplifié). Il est alors indispensable de considérer le troisième type de biais pour se rendre compte que c'est finalement au désavantage des femmes. Le taux de faux positifs est plus important pour les hommes $(0,081)$ que pour les femmes $(0,016)$ ($RFP = 0,20$). Ceci avantage les hommes qui bénéficient plus largement d'une décision favorable même à tort. En revanche, le taux de faux négatifs est plus important pour les femmes $(0,41)$, à leur désavantage, que pour les hommes $(0,38)$ ($Rfn = 1,08$) mais ces dernières différences ne sont pas significatives.

Dans une telle situation, en choisissant le seuil de décision par défaut à $0,5$, une banque prendrait peu de risque : faible taux de faux positifs et taux élevés de faux négatifs mais, conclusion importante, il apparaît une *rupture d'équité* au sens où la banque prend *plus de risques au bénéfice des hommes* alors que les taux d'erreur les concernant sont plus élevés.

Une atténuation du biais des rapports de cotes se justifie donc afin de rendre comparables les chances d'obtention d'un crédit selon le genre et ce même à tort. Plutôt que d'équilibrer ces chances en pénalisant celles des hommes, une part de discrimination positive est introduite au bénéfice des femmes pour plus d'équité en cherchant à rendre égaux les taux de faux positifs selon le genre et évalués sur l'échantillon de validation.

Les deux dernières lignes de la figure 2 proposent une façon simple (*post-processing*), parmi une littérature très volumineuse, de corriger le biais pour plus de *justice sociale*. Deux algorithmes sont entraînés, un par genre et le seuil de décision (revenu élevé ou pas, accord ou non de crédit...) est

abaissé pour les femmes : $0,3$ pour les forêts aléatoires, $0,2$ pour un arbre binaire, au lieu de celui par défaut de $0,5$ pour les hommes. Cette correction des faux positifs impacte également les taux d'erreur qui deviennent plus équilibrés selon le genre et provoque également une atténuation de l'effet disproportionné pour une *société plus équitable*. L'arbre binaire utilisé (`TreeDiscrPos`) est celui pénalisé (peu de feuilles) afin d'obtenir une interprétation facile au prix de la précision. Les seuils et le paramètre de pénalisation ont été déterminés sur l'échantillon de validation avant d'être appliqués indépendamment à l'échantillon test.

4.3. Discussion

Nous pouvons tirer quelques enseignements de cet exemple « jouet » imitant le calcul d'un score d'attribution de crédit bancaire :

- Sans précaution, si un biais est présent dans les données, il est appris et même renforcé par un modèle linéaire élémentaire.
- La suppression naïve de la variable sensible (genre) pour réduire le biais n'y change rien, d'où l'importance (art. 10, 5.) d'autoriser la prise d'un risque contrôlé de confidentialité pour intégrer des données personnelles sensibles afin de pouvoir détecter des biais.
- Un algorithme sophistiqué, non linéaire et impliquant les interactions entre les variables, ne fait que reproduire le biais mais, opaque, ne permet plus de justification des décisions si l'effet disproportionné est juridiquement attaquable comme aux USA ($DI < 0,8$). Dans le cas présent, un simple arbre binaire pénalisé pour contrôler le nombre de feuilles permet de concilier accroissement peu important du biais et explicabilité sans trop pénaliser la précision.
- En présence de proxys du genre comme c'est le cas dans cet exemple, une procédure de *testing* (*counterfactual test*) est complètement inadaptée à la détection *ex-post* d'une discrimination algorithmique. Seule une analyse rigoureuse d'une documentation loyale (art. 11) décrivant les données, la procédure d'apprentissage, les performances, peut donc s'avérer convaincante sur les capacités non discriminatoires d'un algorithme.
- Sur cet exemple, le choix d'un *post-processing* permettant d'atténuer le biais des rapports de cotes conditionnels (taux de faux positifs similaires) selon le genre impacte les trois types de biais pour en réduire simultanément l'importance. C'est une façon de légitimer l'introduction d'une dose de discrimination positive qui réduit le désavantage fait aux femmes sans pour autant nuire aux hommes.
- Finalement, dans cet exemple illustratif, un arbre pénalisé pour être suffisamment simple (nombre réduit de feuilles) et assorti d'une touche de discrimination positive fournit une aide à la décision explicable à un client et équitable en terme de risques de la banque vis-à-vis de son genre.
- Certes, dans le cas d'un score de crédit, cela aurait pour conséquence d'accroître le risque de la banque en réduisant la qualité de prévision et augmentant le taux de faux positifs pour les femmes, mais lui fournirait des arguments tangibles de communication pour une image « éthique » : des décisions inclusives donc plus équitables et plus explicables sans trop nuire à la précision.

5. Conclusion

Comme le rappelle Meneceur (2021b) dans une comparaison exhaustive des démarches institutionnelles, les très nombreuses approches éthiques visant à encadrer le développement et l'applica-

tion des systèmes d'IA ne sont pas des réponses suffisantes et convaincantes pour développer la confiance des usagers. Ceci motive la démarche de la CE aboutissant à la publication de ce projet de règlement alors que le *Conseil de l'Europe envisage également un mélange d'instruments juridiques contraignants et non contraignants pour prévenir les violations des droits de l'homme et des atteintes à la démocratie et à l'État de droit*; la nécessité de conformité se substitue à l'éthique.

L'analyse du projet de règlement européen montre des avancées significatives pour plus de transparence des systèmes d'IA :

- importance fondamentale des données et donc de leur analyse préalable fouillée et documentée,
- évaluation et documentation explicite des performances et donc des risques d'erreur ou de manquement : robustesse, résilience,
- documentation explicite sur les capacités d'interprétation d'un système, d'une décision, à la mesure des technologies et méthodes disponibles,
- prise en compte de certains types de biais : équité sociale dans les données, performances selon des groupes et suivi des risques possibles de discrimination associés,
- enregistrement de l'activité pour une traçabilité du fonctionnement,
- contrôle humain approprié pour réduire et anticiper les risques,
- obligation de fournir la documentation exhaustive à l'utilisateur (système d'IA de l'annexe III), qui est auditée *ex-ante* par un organisme notifié pour les systèmes d'IA de l'annexe II, pour l'obtention du marquage « CE ».

Néanmoins, ce projet de règlement, principalement motivé par une harmonisation des relations commerciales au sein de l'Union selon le principe de sécurité des produits ou de la responsabilité du fait des produits défectueux, ne prend pas en compte des dommages pouvant impacter les usagers. Les conséquences ou objectifs de la démarche adoptée par la CE rejoignent d'ailleurs les exigences de la FTC[22] (*Federal Trade Commission*) (Jillson, 2021) de loyauté et transparence vis-à-vis des performances d'un système d'IA commercialisé. Aussi, certains droits fondamentaux, bien que retenus comme *exigence essentielle* dans le livre blanc, se trouvent pour le moins négligés et ce d'autant plus que les systèmes d'IA à haut risque de l'annexe III ne sont pas concernés par la certification d'un organisme notifié indépendant.

- Plus largement que les seules applications de l'IA, une prise en compte d'une forme de frugalité numérique afin de réduire les impacts environnementaux ne semble pas, dans ce projet d'*AI Act*, une préoccupation majeure de la CE. Cela concerne la consommation énergétique pour le stockage massif et l'entraînement des algorithmes et la sur-exploitation des ressources minières nécessaires à la fabrication des équipements numériques.
- Il est certes conseillé de rechercher des biais potentiels dans les données (art. 10, 2., (f)) avec même la possibilité de prendre en compte des données personnelles sensibles (art.10, 5.) pour traquer des biais systémiques sources potentielles de discrimination. Néanmoins, l'absence de précisions sur la façon de mesurer ces biais, de les atténuer ou les supprimer dans les procédures d'entraînement laisse un vide potentiellement préjudiciable à l'usager. Alors qu'il est déjà fort complexe pour un usager d'apporter la preuve d'une présomption de discrimination, par exemple par *testing*, lors d'une décision humaine, l'exemple numérique ci-dessus montre que c'est mission impossible face à une décision algorithmique. Seule une procédure rigoureuse d'audit de la documentation décrivant les données, la procédure d'apprentissage et les dispositions mises en place pour gérer, atténuer les biais, peut garantir une protection *a minima* des

22. https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

usagers finaux contre ce type de discrimination. Cette mise en conformité agit comme un renversement de la charge de la preuve mais qui ne bénéficie, pour les systèmes d'IA de l'annexe III, qu'à l'information de l'utilisateur et pas, dans l'état actuel, à la protection de l'usager.

– Consciente de ces problèmes, la Défenseure des Droits a récemment publié un avis en collaboration avec le réseau européen EQUINET[23] dont les principales conclusions sont résumées dans un communiqué de presse[24]. Elle y *appelle à replacer le principe de non-discrimination (de l'usager) au cœur du projet d'AI Act*. Une des questions essentielles reste à savoir qui pourra, en dehors de l'utilisateur, avoir accès à la documentation d'un système d'IA à haut risque, et donc de pouvoir l'auditer dans de bonnes conditions. Ce sera sans doute à chaque État membre de légiférer sur ces questions.

– Notons que le Laboratoire National de Métrologie et d'Essai[25] (LNE) a pris les devants en proposant un référentiel de certification de processus pour l'IA[26] (LNE, 2021). Ce référentiel concerne le processus de conception d'un système d'IA et non la certification du produit final requérant la connaissance de normes encore à définir. Le LNE jouera le rôle d'organisme notifié pour les systèmes de transport de l'annexe II et sa filiale GMED[27] pour les dispositifs de santé sous la responsabilité de l'Agence Nationale de Sécurité des Médicaments comme autorité notifiante.

– Le Conseil d'État (2022) publie un rapport dont la recommandation finale vise spécifiquement à combler certaines des lacunes identifiées de l'*AI Act*.

– L'étude préconise enfin une transformation profonde de la CNIL en autorité de contrôle nationale responsable de la régulation des systèmes d'IA, notamment publics, pour incarner et internaliser le double enjeu de la protection des droits et libertés fondamentaux, d'une part, et de l'innovation et de la performance publique, d'autre part.

L'exemple numérique jouet a également pour mérite de montrer clairement l'*interdépendance* de toutes les contraintes : confidentialité, qualité, explicabilité, équité (types de biais), que devrait satisfaire un système d'IA pour gagner la confiance des usagers. Il montre aussi que le problème ne se réduit pas à un simple objectif de minimisation d'un risque quantifiable pour l'obtention d'un meilleur compromis. C'est plutôt la recherche d'une moins mauvaise solution imbriquant des choix techniques, économiques, juridiques, politiques qu'il sera nécessaire de clairement expliciter dans la documentation rendue obligatoire par l'adoption à venir d'un *AI Act* qui serait, de toute façon et malgré les limites actuelles du projet de texte, une avancée notable pour plus de transparence.

Références

Bachoc F., Gamboa F., Halford M., Loubes J.-M., and Risser L. (2020), « Entropic Variable Projection for Model Explainability and Intepretability », arXiv preprint : 1810.07924, https://arxiv.org/abs/1810.07924.

Barocas S. and Selbst A. (2016), « Big Data's Disparate Impact », *California Law Review*, 104, pp. 671-732, http://dx.doi.org/10.2139/ssrn.2477899.

Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., and Herrera F. (2020), « Explainable Artificial

23. https://www.defenseurdesdroits.fr/sites/default/files/atoms/files/avis_equinet_sur_lintelligence_artificielle_et_legalite.pdf
24. https://www.defenseurdesdroits.fr/sites/default/files/atoms/files/cp_-_defenseur_des_droits_-_intelligence_artificielle.pdf
25. https://www.lne.fr/fr
26. https://www.lne.fr/sites/default/files/bloc-telecharger/referentiel-certification-LNE-processus-IA.pdf
27. https://lne-gmed.com/fr

Intelligence (XAI) : Concepts, taxonomies, opportunities and challenges toward Responsable AI », arXiv, https://arxiv.org/abs/1910.10045.

Bénesse C., Gamboa F., Loubes J.-M., and Boissin T. (2021), « Fairness seen as Global Sensitivity Analysis », ArXiv, https://arxiv.org/pdf/2103.04613.pdf.

Besse P. (2021), « Médecine, police, justice: l'intelligence artificielle a de réelles limites », *The Conversation*, 01/12/2021.

Besse P., Besse-Patin A. et Castets-Renard C. (2020), « Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé », *Statistique et Société*, 8(3), pp. 21-53.

Besse P., Castets-Renard C., Garivier A. et Loubes J.-M. (2019), « L'IA du quotidien peut-elle être éthique ? Loyauté des algorithmes d'apprentissage automatique », *Statistique et Société*, 6(3), pp. 9-31.

Besse P., del Barrio E., Gordaliza P., Loubes J-M., and Risser L. (2021), « A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set », *The American Statistician*, DOI : 10.1080/00031305.2021.1952897, https://arxiv.org/pdf/2003.14263.pdf (version en accès libre).

Breiman L. (2001), « Random forests », *Machine Learning*, 45, pp. 5-32.

Campisi N. (2021), « From Inherent Racial Bias to Incorrect Data—The Problems With Current Credit Scoring Models », *Forbes Advisor*.

Castets-Renard C. et Besse P. (2022), « Responsabilité ex ante de l'AI Act : entre certification et normalisation, à la recherche des droits fondamentaux au pays de la conformité », *in* Castets-Renard C. et Eynard J. (éds.), *Un droit de l'intelligence artificielle : entre règles sectorielles et régime général. Perspectives de droit comparé* (à paraître), Bruylant.

CE (2019), « Lignes Directrices pour une IA digne de Confiance », rédigé par un groupe d'experts européens.

CE (2020), « Livre blanc sur l'intelligence artificielle: une approche européenne d'excellence et de confiance ».

CE (2021), « Règlement du parlement et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union ».

Conseil d'État (2022), « S'engager dans l'intelligence artificielle pour un meilleur service public », Rapport d'étude mis en ligne le 30/08/2022.

De-Arteaga M., Romanov A. *et al.* (2019), « Bias in Bios : A Case Study of Semantic Representation Bias in a High-Stakes Setting », Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 120-128, https://dl.acm.org/doi/pdf/10.1145/3287560.3287572.

Défenseure des Droits (2020), « Algorithmes: prévenir l'automatisation des discriminations », Rapport.

Défenseur des Droits, CNIL (2012), « Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi ».

Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E., Roth D. (2019), « Comparative study of fairness-enhancing interventions in machine learning », Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329-338, http://dl.acm.org/citation.cfm?doid=3287560.3287589.

Goglin C. (2021), « Discrimination et IA : comment limiter les risques en matière de crédit bancaire », *The Conversation*, 23/09/2021.

Hurlin C., Pérignon C., and Saurin S. (2021), « The fairness of credit score models », Preprint SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3785882.

Jillson E. (2021), « Aiming for truth, fairness, and equity in your company's use of AI », Blog (consulté le 29/05/2021).

Kozodoi N., Jacob J., and Lessman S. (2021), « Fairness in credit scoring : assessment, implementation and profit implications », arXiv preprint : 2103.01907, https://arxiv.org/abs/2103.01907.

Larson J., Mattu S., Kirchner L., and Angwin J. (2016), « How we analyzed the compas recidivism algorithm », *ProPublica* (en ligne, consulté le 28/04/2020), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lê S., Josse J., and Husson F. (2008), « FactoMineR : An R Package for Multivariate Analysis », *Journal of Statistical Software*, 25(1), pp. 1-18.

Liu X., Faes L., Kale A. U. *et al.* (2019), « A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging : a systematic review and meta-analysis », *The Lancet Digital Health*, 1, pp. e271–e297, https://doi.org/10.1016/S2589-7500(19)30123-2.

LNE (2021), « Référentiel de Certification du Processus IA », Laboratoire National de Métrologie et d'Essais.

Meneceur Y. (2021), « Analyse des principaux cadres supranationaux de régulation de l'intelligence artificielle: de l'éthique à la conformité », Projet d'étude, Institut des Hautes Études sur la Justice (IHEJ), Version d'étude du 27/05/2021.

Raghavan M., Barocas S., Kleinberg J., and Levy K. (2019), « Mitigating bias in Algorithmic Hiring : Evaluating Claims and Practices », Proceedings of the Conference on Fairness, Accountability, and Transparency, https://arxiv.org/abs/1906.09208.

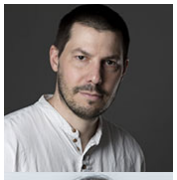
Rich J. (2014), « What Do Field Experiments of Discrimination in Markets Tell Us ? A Meta Analysis of Studies Conducted since 2000 », *IZA Discussion Paper*, 8584, http://ftp.iza.org/dp8584.pdf.

Roberts M., Driggs D., Thorpe M., Gilbey J., Yeung M., Ursprung S., Aviles-Rivero A. I., Etmann C., McCague C., Beer L., Weir-McCall J. R., Teng Z., Gkrania-Klotsas E., AIX-COVNET, Rudd J. H. F., Evis Sala, Schönlieb C.-B. (2021), « Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans », *Nature Machine Intelligence*, 3, pp. 199-217.

Verzelen N. (2012), « Minimax risks for sparse regressions : Ultra-high dimensional phenomenons », *Electron. J. Statist.*, 6, pp. 38-90, https://doi.org/10.1214/12-EJS666.

Zliobaité I. (2017), « Measuring discrimination in algorithmic decision making », *Data Mining and Knowledge Discovery*, 31(4), pp. 1060-1089, https://dl.acm.org/doi/10.1007/s10618-017-0506-1.

L'équité de l'apprentissage machine en assurance



Arthur CHARPENTIER[1]
Professeur, Université du Québec, Montréal

Laurence BARRY[2]
Chaire PARI, Fondation Institut Europlace de Finance

TITLE
Machine Learning, Fairness and Insurance

RÉSUMÉ
Les assureurs sont réputés utiliser des données pour classer et tarifer les risques. À ce titre, dès la fin du 19e siècle, ils ont été confrontés aux problèmes d'équité et de discrimination associées aux données. Pourtant, si cette question est récurrente, elle connait un regain d'importance avec l'accès à des données de plus en plus granulaires, massives et comportementales. Nous verrons ici comment les biais de l'apprentissage machine en assurance renouvellent ou transforment ce questionnement pour rendre compte des technologies et des préoccupations sociétales actuelles : paradoxalement, alors que la plupart de ces biais ne sont pas nouveaux, la recherche d'une équité pour les contrer, elle, se transforme.

Mots-clés **:** *assurance, classification, big data, algorithmes, discrimination, biais, équité.*

ABSTRACT
Insurers have been known to use data to classify and price risks. As such, they were confronted since the end of the nineteenth century with the problems of equity and discrimination associated with data. However, although this issue is recurrent, it is becoming more important with the access to increasingly granular and behavioral data. We will see here how machine learning biases in insurance renew or transform this questioning to account for current technologies and societal concerns: paradoxically, while most of these biases are not new, the search for fairness to counter them is being transformed.

Keywords**:** *insurance, classification, big data, algorithms, discrimination, biases, fairness.*

1. charpentier.arthur@uqam.ca
2. barry678@outlook.com

1. Introduction

Les assureurs quantifient le réel, fabriquant et utilisant des données pour classer et tarifer les risques. Dans leur analyse de l'impact des techniques actuarielles au cours des deux derniers siècles, Knights and Vurdubakis (1993) soutiennent que l'assurance *crée* le risque ou départage ce qui, dans l'incertitude, sera couvert par des mécanismes collectifs de ce qui restera du domaine de l'incertain. Le risque est la part quantifiée, modélisée, de l'incertitude ; c'est aussi sa part prise en charge par les institutions, État providence ou assureurs. Ainsi, « *the quantitative principles adopted by insurance (...) derive their particular rationality from the institution of* ***socially and historically specific modes of cognition and intervention*** » (Knights & Vurdubakis, 1993, p. 735, notre accentuation).

Cette pratique de quantification des risques est donc aussi un geste politique, dans la mesure où l'assurance joue dans les sociétés industrialisées un rôle prépondérant dans l'ouverture ou la fermeture d'opportunités de vie (Baker & Simon, 2002 ; Horan, 2021). À ce titre, les assureurs ont été confrontés dès la fin du 19e siècle aux questions d'équité associées aux données. Mettre en évidence un biais, c'est adopter une position critique par rapport à un calcul, et éclairer la dimension politique que cache sa prétendue objectivité.

On notera d'emblée que, dans la pratique actuarielle, la discrimination est une notion technique, d'ordre statistique, qui se veut neutre et objective. Pour Charpentier (2021), l'usage du terme en statistique remonte probablement aux premiers travaux de Ronald Fisher dans les années 20, dont le but était de différencier et classer (en anglais, *discriminate*) deux espèces de fleurs suivant les caractéristiques de leurs pétales. Il n'empêche qu'en assurance, pour les raisons évoquées plus haut, toute classification comme discrimination statistique est susceptible d'être perçue comme une injustice, rejoignant ainsi le langage courant de discrimination *sociale*.

Dès 1909, le régulateur du Kansas pose les contours d'une pratique éthique de la classification, dans le but de protéger les petits souscripteurs d'une assurance incendie qui payaient des primes beaucoup plus élevées que les gros industriels, pour un risque identique. Il définit ainsi une tarification comme « non inéquitablement discriminatoire » (*not unfairly discriminatory*), si elle traite de la même manière des risques semblables (Frezal & Barry, 2020 ; M. J. Miller, 2009). C'est sur la base de ce principe que l'usage de certains paramètres en tarification assurantielle a été contesté dans le courant du 20e siècle, aboutissant à une typologie assez précise des biais liés à un traitement classique des données.

L'émergence des données massives et des nouveaux algorithmes bouscule à première vue cette typologie, puisque l'accent n'est plus sur le choix des variables. On verra cependant ici que les biais induits par ces techniques et les diverses notions d'équité algorithmique remettent au goût du jour et renouvellent, avec des points de rupture et de continuité, des débats plus anciens liés à la discrimination en assurance.

La première partie propose une mise en perspective historique des biais en assurance et débouche sur une typologie de ces biais. La deuxième partie met à profit cette typologie pour étudier l'impact potentiel des données massives et de l'apprentissage machine sur les biais en assurance. La dernière partie discute enfin plus généralement des enjeux éthiques, pour l'assurance toujours, de cette transformation technologique.

2. Assurance, biais et équité : une approche historique

2.1 La pratique de tarification avant l'apprentissage machine

L'assurance consiste en la mise en commun de l'incertitude : la contribution de chacun permet la compensation des accidents survenus aux plus malchanceux. Dans sa forme la plus grossière, la prime de risque assurantielle est l'espérance mathématique des dommages de l'accident, calculée sur le groupe en question. Si la concurrence ne joue pas entre assureurs, on peut très bien se contenter d'un tarif unique, moyenne du risque pour l'ensemble de la population. Mais la concurrence fait craindre l'antisélection : en réduisant la prime des meilleurs risques, l'assureur A peut les attirer à lui, bonifiant ainsi son portefeuille aux dépens de ses concurrents qui eux feront des pertes. La segmentation, qui consiste à distinguer des groupes porteurs de risques voisins, devient donc très vite la règle du jeu.

Cette segmentation a consisté pendant très longtemps en la création de classes supposées homogènes, sur lesquelles le risque est estimé en moyenne (Charpentier *et al*., 2015). Le travail de l'actuaire était donc, avant tout calcul, celui du choix des variables, choix qui dictait une homogénéité projetée sur le monde, et ce de deux façons : dans le choix de ce qui est ignoré d'une part, puisque ce qui n'est pas collecté contient des différences qui ne seront pas vues ; dans la catégorisation de ce qui est collecté d'autre part, qui conduit là encore à écraser des différences potentielles.

Petit à petit, émerge l'idée d'une « tarification parfaite », dans laquelle la classe tarifaire ne comporterait que des risques parfaitement identiques. En reprenant la formalisation de Denuit et Charpentier (2004), et si l'on admet que θ est la variable qui caractériserait parfaitement le risque Y :

	Assuré	**Assureur**
Perte	$E[(Y\|\theta)]$	$Y - E[(Y\|\theta)]$
Perte moyenne	$E[Y]$	0
Variance	$Var[E[(Y\|\theta)]]$	$Var[Y - E[(Y\|\theta)]]$

La variance sur le portefeuille est ainsi distribuée entre les assurés qui paient des primes proportionnelles à leur risque (capté par θ) et l'assureur qui porte la variance résiduelle, inexpliquée par θ. Dans les années 80, et avec une notation similaire, De Wit et Van Eeghen (1984) estiment que les capacités croissantes de collecte de données et de calcul des ordinateurs, permettent d'envisager l'affinement de la part expliquée de la variance (et les primes segmentées), diminuant ainsi celle portée par l'assureur.

Cependant, le paramètre θ, supposé caractériser le risque de façon parfaite, n'est en réalité jamais connu. Cette incertitude est le fondement même de l'assurance : si on cherche, par exemple, à modéliser et valoriser des garanties en cas de décès, on peut estimer de manière plus fine la probabilité de décès (certains ayant 1 chance sur 10,000 et d'autres 1 chance sur 1,000 de mourir), mais il demeure impossible de prédire *qui* va décéder dans l'année (Charpentier, Barry, & Gallic, 2020). Cette incertitude résiduelle fondamentale reste irréductiblement à la charge de l'assureur, créant ce que De Wit et Van Eeghen (1984) appellent une solidarité purement probabiliste (couverte par la loi des grands nombres). La classification est alors repensée comme un moyen d'approcher θ : on ne cherche plus seulement à contrer l'antisélection avec des classes de plus en plus fines, mais on interprète ce travail comme une approximation de θ par les paramètres de tarification, comme un moyen de faire converger la variance non expliquée par le modèle vers la variance minimale du portefeuille. Ainsi la pratique actuarielle

ne change pas, même si son sens évolue.

C'est dans ce cadre d'ajustement qu'apparait la notion de biais : dans l'hypothèse où un calcul exact du risque est possible, il se produit si la classification est imparfaite et conduit à mal tarifer certains groupes, créant des transferts croisés entre assurés (De Pril & Dhaene, 1996 ; Walters, 1981).

2.2 La critique d'une tarification biaisée

À partir des années 60 aux États-Unis, la classification des risques est remise en cause, sur deux aspects spécifiques. Dans le contexte de lutte pour les droits civiques des noirs tout d'abord, c'est la pratique du « red lining » ou d'exclusion de certaines zones géographiques des portefeuilles assurés qui est montrée du doigt. Puis, à la fin des années 70, les mouvements féministes tentent de contrer l'usage du sexe dans la tarification (Horan, 2021). Ainsi, dans l'affaire Manhart en 1978 puis dans l'affaire Norris en 1983, la Cour suprême a jugé que l'utilisation du paramètre homme/femme dans la tarification d'un régime de retraite à prestations définies était illégale, car cela violait les principes d'égale opportunité et d'avancement individuel (Austin, 1983 ; Avraham, 2017 ; Horan, 2021).

Les débats autour des actions de groupe menées alors mettent en lumière les différents aspects de ce que l'on peut appeler une « tarification biaisée ». Nous proposons dans cette section une description théorique de ces débats, débouchant sur une typologie des biais assurantiels pré-machine learning. En réalité, le sens de la critique dépend fortement de l'hypothèse que l'on fait sur le monde et la nature du risque.

La classification peut être pensée tout d'abord comme une méthode de répartition *ex-ante* des coûts futurs, toujours plus ou moins arbitraire. Dans cette hypothèse, le travail de quantification (du statisticien ou de l'actuaire) est critiqué car il se nourrirait d'une vision du monde toujours subjective, dictée par un contexte historique et culturel spécifique (Desrosières, 2008). Pour l'exemple, une lecture des manuels de formation à la souscription des années 70 révèle une description des femmes comme peu fiables, instables dans leur travail, incapables de prendre des décisions financières de façon autonome et dépendantes de leur partenaire masculin pour vivre (Horan, 2021, p. 174), justifiant l'usage du paramètre homme/femme dans la tarification.

Glenn (2000) fait alors remarquer que, comme le dieu romain Janus, le processus de sélection des risques d'un assureur a en réalité deux visages. Il y a d'un côté le visage des chiffres, des tables actuarielles, et des statistiques, qui se posent comme objectives et rationnelles. Mais de l'autre, il y a le visage des récits, du caractère et du jugement subjectif. Pour Glenn, l'actuaire crée un mythe dans lequel les décisions apparaissent comme objectives alors qu'elles reposent sur beaucoup de subjectivité, de préjugés et de stéréotypes. Ces derniers sont visibles en amont des tables actuarielles, dans les histoires que se racontent les techniciens de l'assurance (actuaires et souscripteurs), et qui les amènent à privilégier telle variable plutôt que telle autre. En effet, comme la collecte des données se fait encore sous forme de questionnaires, elle est à la fois coûteuse et nécessairement contrainte en volume. Elle est aussi contingente à ce que l'on peut techniquement et/ou historiquement mesurer, ce qui induit une certaine instabilité dans la classification. Comme le dit Baker (2002) : « *While some 'low risk' individuals may believe that they are benefited by risk classification, any particular individual is only one technological innovation away from losing his or her privileged status* » (voir aussi Frezal & Barry, 2020).

De plus, dans les controverses autour de la classification, Horan (2021, pp. 170–71) montre que les paramètres de tarification évoluent aussi pour répondre aux contraintes réglementaires, politiques ou sociales : « *the categories insurance companies used to create risk classifications throughout the twentieth century reflected changing political trends and social values, and not simply*

objective realities ». L'histoire des biais en assurance est de fait aussi l'histoire de ce qui est perçu comme acceptable ou inacceptable dans une société donnée. Dans cette perspective, diverses classifications peuvent avoir une efficacité équivalente. Le choix de l'une d'entre elles est livré à l'arbitraire des décisions des praticiens, eux-mêmes guidés ou contraints par le contexte dans lequel ils évoluent :

> « *Insurers can rate risks in many different ways depending on the stories they tell about which characteristics are important and which are not (…) The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums* » (Glenn, 2003, p. 135).

Dans l'affaire Manhart, l'un des juges met ainsi en avant la fluidité culturelle et historique de ce qui est perçu comme légitime, à la fois comme classification ex-ante mais aussi comme explication ex-post du modèle :

> « *Habit, rather than analysis, makes it seem acceptable and natural to distinguish between male and female, alien and citizen, legitimate and illegitimate; for too much of our history there was the same inertia in distinguishing between black and white. But* ***that sort of stereotyped reaction may have no rational relationship—other than pure prejudicial discrimination—to the stated purpose for which the classification is being made*** » (cité dans Simon, 1988, p. 796, notre accentuation).

Pour Schauer (2003), il conviendrait de distinguer deux types de stéréotypes. Certaines généralisations sont totalement infondées : des généralisations sur la base du signe astrologique de la personne, par exemple, relèvent de purs préjugés. Mais d'autres ont un fondement statistique, lorsque la probabilité d'avoir un caractère y sachant x est significativement différente du cas où l'on ne sait rien. Dans cette perspective, l'usage du paramètre homme/femme reste légitime car statistiquement fondé pour estimer une probabilité de décès ou d'accident automobile. Serait alors légitime toute classification sur la base de variables effectivement corrélées au risque que l'on cherche à modéliser.

Works (1977) met cependant en garde contre les « variables de procuration », par opposition aux « vraies variables » du risque. Ces dernières étant plus difficiles à obtenir, elles sont remplacées par de simples corrélations. L'hypothèse sous-jacente n'est plus que la classification est nécessairement arbitraire, mais qu'au contraire il existerait de « vraies variables » du risque, qui expliqueraient les accidents de façon causale, toutes les autres étant invalides. L'usage de variables de procuration ouvrirait alors la porte aux biais dans la tarification et la souscription :

> « *Although the core concern of the underwriter is the human characteristics of the risk,* ***cheap screening indicators are adopted as surrogates for solid information*** *about the attitudes and values of the prospective insured (…) The invitations to underwriters* ***to introduce prejudgments and biases and to indulge amateur psychological stereotypes are apparent****. Even generalized underwriting texts include occupational, ethnic, racial, geographic, and cultural characterizations certain to give offense if publicly stated* » (Works, 1977, p. 471, notre accentuation).

Dans les actions de groupe menées aux États-Unis contre l'usage du paramètre homme/femme, c'est cette approche qui est adoptée par les plaignantes. Leur argument principal est en effet que la corrélation observée entre coût des sinistres en assurance automobile et sexe du conducteur est due au moindre kilométrage parcouru par les femmes ; c'est le kilométrage qui est la variable causale, donc légitime, et non le sexe qui n'est qu'une approximation biaisée de cette dernière (Horan, 2021 ; Krippner & Hirschman, 2022).

Le problème de ce type d'argument vient de la difficulté à établir l'existence d'une causalité directe, et du fait que, par conséquent, cette détermination relève le plus souvent d'un

jugement plutôt que d'une vraie preuve scientifique : la causalité ne serait qu'un narratif accepté comme scientifiquement et/ou politiquement valide. Pour l'exemple, Hoffman à la fin du 19e siècle s'appuie sur une corrélation entre la durée de vie et la couleur de la peau pour affirmer l'existence d'une causalité innée liée à la race noire et qui la rend plus risquée, là où d'autres auraient cherché les causes environnementales et sociales expliquant la plus grande mortalité des noirs (Heen, 2009, p. 377).

Simon (1988, pp. 795-796) soutient que causalité ou corrélation finalement importent peu lorsqu'il s'agit de lutter contre une discrimination sociale flagrante : sur cette base, l'usage du paramètre, causal ou pas, contribue à naturaliser la différence de traitement (social) et donc à ancrer dans la réalité la discrimination. La solution consiste alors à « protéger la variable », c'est-à-dire à l'éliminer des variables autorisées dans le traitement statistique (voir section suivante).

Les critiques de la classification classique déterminent ainsi une typologie des différents biais possibles, que l'on retrouvera de façon modifiée dans les méthodes d'apprentissage machine :

- Les biais de type 1 sont liés à des classes qui ne reflèteraient pas la réalité du risque, mais seraient motivés par de purs préjugés (critique qui ne remet pas en question le principe du bien-fondé de la classification). Une classification sur la base des signes du zodiaque se révèlerait à l'usage comme « biaisée », au sens trivial où le modèle est faux ;
- Les biais de type 2 sont liés à des classes qui reflètent une réalité statistique avérée (une corrélation avec le risque, donc un modèle exact) mais non causale, ce qui rend leur usage suspect d'un parti-pris et d'un choix arbitraire. C'est le cas par exemple du paramètre homme/femme ;
- Les biais de type 3 sont liés à des classes qui reflètent une réalité statistique et causale, mais qui est elle-même le fait de discriminations sociales en amont. Dans ce cas, le modèle est exact mais la classification est intrinsèquement nuisible car elle reproduit et ancre dans la réalité une situation contre laquelle il faut lutter.

Il est intéressant de noter ici que cette typologie ne décrit pas intrinsèquement telle ou telle variable, mais la façon dont on se représente les biais. On verra dans la section qui suit comment certains paramètres peuvent être reconnus comme des variables causales et acceptables socialement à un moment donné de leur histoire, pour basculer ensuite dans la catégorie des variables corrélées à une cause plus profonde et/ou dans celle des variables protégées.

2.3 Les variables protégées

Pour répondre au troisième type de biais et prévenir ou remédier à une discrimination sociale, on peut choisir d'interdire l'usage de certaines variables, dites protégées ou sensibles. Cette section décrit de manière non exhaustive les controverses associées à l'usage historique de quelques paramètres controversés, réputés créer des biais d'un type ou d'un autre ; nous verrons notamment que la sensibilité d'une variable est contingente au contexte culturel. En Europe, les données protégées concernent aujourd'hui notamment les croyances religieuses, le sexe, l'orientation sexuelle, l'engagement syndical, l'appartenance ethnique, la situation médicale, les condamnations et infractions pénales, les données biométriques, les informations génétiques.

2.3.1 L'origine ethnique

Alors qu'en France la collecte et l'usage statistique de l'origine ethnique des individus reste un sujet polémique, ils sont assez répandus aux États-Unis. En assurance-vie, Bouk (2015) décrit comment, à la fin du 19e siècle, les assureurs faisaient payer la même prime à tout le monde mais réglaient les sinistres de façon différenciée suivant la couleur de peau (voir aussi Heen, 2009). Plusieurs États adoptent alors des lois anti-discrimination. Ainsi, au cours de l'été 1884, l'État du Massachusetts promulgue une loi interdisant de faire « *any distinction or discrimination*

between white persons and colored persons wholly or partially of African descent, as to the premiums or rates charged for policies upon the lives of such persons » (cité par Wiggins, 2013, p. 68). Pour contrer la loi, Frederick L. Hoffman, soutenu par Prudential Life Insurance, publie en 1896 un ouvrage démontrant statistiquement la mortalité plus élevée des Noirs américains (Bouk, 2015, pp. 49-52 ; Heen, 2009, p. 377). Les assurer au même tarif que les Blancs serait statistiquement inéquitable, soutenait-il ; ne pas les assurer était donc la seule manière de se conformer à la loi, qui rendait de fait les Noirs américains non-assurables.

Le sujet reste d'actualité pendant la majeure partie du 20e siècle[3], même si la couleur de peau disparait des tables actuarielles après la seconde guerre mondiale. Pour Heen (2009, p. 364), c'est moins la législation – qui interdisait l'usage de l'origine raciale depuis la fin du 19e siècle – que les leçons de la guerre et du nazisme qui conduisent les assureurs à bannir le paramètre : « *change came from a form of collective action by life insurance industry professional groups, which was achieved only after a fundamental rethinking of race, a 'change in the habit of the public mind' that led to reconsideration of long-established commercial practice* » (Heen, 2009, p. 399).

L'efficacité de ce ban est cependant discutable. Très vite, en effet, la zone géographique, comme variable de procuration de l'origine ethnique, est utilisée dans les tarifs. Une enquête commissionnée par l'État fédéral dans les années 60 met ainsi en évidence la pratique systématique de « *red-lining* » (Austin, 1983 ; Horan, 2021) : de nombreuses institutions financières, dont des compagnies d'assurance, refusent de desservir des zones géographiques à prédominance afro-américaine, conduisant à une détérioration des services et des infrastructures dans certaines villes. Une variable corrélée au risque mais devenue inacceptable, est remplacée par une autre variable corrélée, mais considérée comme neutre : on choisit volontairement un biais de type 2 pour contourner la législation qui visait à éviter un biais de type 3 – sans fondamentalement changer la réalité de la discrimination sociale.

Des études récentes en assurance automobile révèlent, en effet, que les quartiers à prédominance afro-américaine continuent d'être tarifés plus cher que les autres, la surprime étant estimée à 70% pour Heller (2015) et 10% pour Larson *et al*. (2017). En réponse, l'association étatsunienne des assureurs *Property Casualty (Property Casualty Insurers Association of America)* soutient que « *insurance rates are color-blind and solely based on risk* » (cité dans Larson *et al*., 2017).

2.3.2 Discrimination Homme/Femme : aléa subi ou volontaire ?

Comme évoqué plus haut, c'est sur l'usage du paramètre homme/femme que les premières controverses autour de la classification actuarielle se sont faites jour. En Europe, une directive de 2004 visait à réduire les écarts entre les sexes dans l'accès à tous les biens et services, mais une dérogation permettait aux assureurs de fixer des prix fondés sur le paramètre homme/femme, à condition qu'ils fournissent des données actuarielles et statistiques permettant d'établir qu'il constitue un facteur objectif d'évaluation du risque. En 2011, soit trente ans après les controverses étatsuniennes, la Cour de justice des Communautés européennes a annulé cette exception, rendant l'usage du paramètre homme/femme caduque pour toutes les classifications (Rebert & Van Hoyweghen, 2015 ; Schmeiser *et al*., 2014), au motif qu'il ne serait que corrélé avec la cause réelle de l'accident (donc biais de type 2).

Dans sa décision, la juge fait par ailleurs la distinction entre deux types de variables, pointant ce qui pourrait être considéré comme une classification équitable, une fois éliminées les variables non significatives et les variables non causales : « *À l'instar de la race et de l'origine ethnique, le sexe est lui aussi une caractéristique inséparable de la personne de l'assuré sur laquelle* ***celui-ci n'a***

3. Heen (2009) soutient qu'il est possible que dans certains États du Sud des États-Unis, d'anciennes polices d'assurance-vie issues de la période Jim Crow (i.e., faisant usage de la race comme paramètre de tarification) soient encore en vigueur aujourd'hui.

pas la moindre influence » (CURIA, 2010, notre accentuation). Cette distinction renvoie à ce que Dworkin (1981) appelle « *brute and option luck* » : les aléas que l'on dira volontaires sont liés à des choix personnels (*option luck*) et peuvent être imputés à l'individu ; les aléas subis, causés par des éléments sur lesquels l'individu n'a aucune prise (*brute luck*), doivent eux être pris en charge par la collectivité (et donc protégés et éliminés de la tarification)[4].

2.3.3 Discrimination par l'âge

À première vue, l'âge comme le sexe ou l'appartenance ethnique est une donnée personnelle sur laquelle l'individu n'a pas prise et devrait donc être, au regard du texte précédent, proscrit des tables actuarielles. Il y a cependant une différence majeure qui en fait un paramètre acceptable. En effet, toujours dans les conclusions de la juge on trouve :

> « *S'il est vrai que l'âge est, lui aussi, une caractéristique indissociablement liée à la personne,* ***tout homme traverse différentes tranches d'âge au cours de son existence****. C'est ainsi que, si les primes et prestations d'assurance sont calculées différemment en fonction de l'âge, cela ne permet pas de craindre, en soi, que l'assuré s'en trouve lésé* ***en tant que personne****. Quiconque peut, au cours de sa vie, bénéficier, en fonction de son âge, de produits d'assurance plus ou moins avantageux pour lui* » (CURIA, 2010, notre accentuation).

Dans une perspective temporelle longue, le traitement différentiel en fonction de l'âge ne génère pas nécessairement des inégalités entre les personnes : « *une société qui discrimine sans relâche les gens en raison de leur âge peut encore les traiter de manière égale tout au long de leur vie (...) Le tour de chacun <d'être discriminé> viendra* » (Gosseries, 2014). Cet argument n'est pourtant valide que si la discrimination reste fixe au cours du temps. Mais les normes sociales et les mécanismes de solidarité évoluent. Ainsi la retraite, financée par répartition en France, fonctionne grâce une solidarité intergénérationnelle qui fait que le poids des retraites pèse sur les actifs. Or cet équilibre dépend de la pyramide des âges, qui est dynamique et fait qu'au cours du temps, certaines générations se trouvent pénalisées par rapport à d'autres. Après-guerre, lors de la mise en place du régime, du fait de l'espérance de vie et de l'âge légal de départ à la retraite, beaucoup de cotisants ne bénéficièrent jamais de leur retraite, par exemple. Plus tard, l'allongement de la durée de cotisation et la baisse du niveau des retraites montrent bien que cette notion de compensation au cours de la vie ne fonctionne pas toujours.

Par ailleurs, en suivant la distinction (dans le biais de type 2) entre variable causale et variable simplement corrélée, il n'est pas évident que l'âge soit la cause de la mortalité. L'âge permet d'inférer assez précisément l'état de santé de la personne, cause réelle du décès, mais variable protégée. L'usage de l'âge pourrait donc introduire des biais dans les modèles. Analysant ainsi un arrêt de la cour d'appel de 2008, Mercat-Bruns (2020) conclut que « *le législateur a pris soin d'opérer une distinction entre l'âge et l'état de santé. Il ne peut dès lors être procédé à un amalgame entre ces deux motifs en considérant que l'âge avancé induit nécessairement une santé défaillante* ».

2.3.4 Discrimination des fumeurs

La responsabilité du tabagisme dans la genèse des cancers (en particulier du poumon) a été longue à établir. Le rôle cancérigène du tabac a été suspecté au lendemain de la Première Guerre mondiale, et le lien entre certains cancers et le tabagisme est établi par les assureurs dès 1930 (Patterson, 1989). Hoffman – le statisticien de Prutential responsable des tables de mortalité raciales – collecte notamment des statistiques à partir de 1915 et conclut : « *smoking habits unquestionably increase the liability to cancer of the mouth, the throat, the oesophagus, the larynx and the lungs* » (Hoffman, 1931, p. 67).

4. Cette distinction n'est toutefois pas toujours simple à établir : voir Charpentier, Barry, & James (2020) pour une discussion dans le cas des catastrophes naturelles.

Les premières quantifications interviennent après-guerre, avec notamment les travaux de Johnston (1945) qui présentent des tables de mortalité comparant non-fumeurs et fumeurs. Des études de grande envergure ont lieu dans les années 1950 et 1960 : Doll et Hill (1964) confirment ainsi le lien entre tabagisme et cancer. Dans un contexte purement actuariel, il faut attendre les années 80, toujours aux États-Unis, pour que les tables de mortalité homologuées tiennent compte de cette variable. Une « task force » est ainsi créée par la Société des Actuaires en 1982 pour proposer une correction aux tables de mortalité en vigueur, grâce à un facteur fumeur/non-fumeur (G. H. Miller & Gerstein, 1983 ; Society of Actuaries, 1982). Dans les années 80, des travaux similaires seront menés en Europe (Benjamin & Michaelson, 1988). En France, le paramètre est rarement utilisé jusqu'à aujourd'hui, même si l'impact sur la mortalité est avéré.

Le facteur a en réalité longtemps fait polémique : ainsi Fisher (1958) met en garde contre l'amalgame entre corrélation et causalité. Pour lui, les études montrent toutes l'existence d'une corrélation avec le cancer du poumon, mais ne prouvent pas que le tabagisme en est la cause : « *it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others* » (Fisher, 1958). Il s'évertue à montrer qu'en réalité l'inclination à fumer est génétique et que c'est aussi cette configuration génétique qui est à l'origine du surplus de cancers dans la population des fumeurs. Dans la perspective de ce papier, le débat autour du tabagisme proposé par Fisher met en avant deux types de biais potentiel : le fait que le facteur fumeur/non-fumeur ne serait pas un facteur causal (biais de type 2) ; le fait que si la cause est génétique, alors elle tombe dans la catégorie des variables sur lesquelles l'individu n'a pas prise et devrait donc être bannie des tarifs pour des raisons d'équité (biais de type 3).

2.3.5 Les scores de crédit

En Amérique du Nord, diverses entreprises telles qu'Experian, Equifax et TransUnion, tiennent des registres des activités d'emprunt et de remboursement d'une personne. La société Fair Isaac Corporation (FICO) a mis au point une formule (tenue secrète) calculant, sur la base de ces registres, un score, fonction de la dette et du crédit disponible (Guseva & Rona-Tas, 2001). Ce score est utilisé pour l'octroi de crédit, à l'embauche (Bartik & Nelson, 2019) et dans la tarification assurantielle (Kiviat, 2019 ; M. J. Miller & Smith, 2003).

Ces usages font cependant aujourd'hui débat car ils créent un cercle vicieux d'appauvrissement des plus pauvres (O'Neil, 2016). François (2021) met également en avant l'aspect auto-réalisateur de la pratique puisqu'un mauvais score augmente le coût du crédit et par conséquent les chances de ne pouvoir le rembourser. En assurance, le régulateur américain s'est récemment penché sur l'équité de la pratique (Kiviat, 2019). Il a notamment cherché à expliquer la corrélation avérée entre mauvais score de crédit et sinistralité. S'il est clair que le score de crédit fonctionne comme procuration de la variable causale, quelle est-elle ? Si, comme le soutiennent les assureurs, le score est une indication de la prudence du conducteur il est un paramètre légitime ; mais s'il est un indicateur du statut socio-économique de l'assuré, et qu'il ne prédit pas l'occurrence d'un accident mais sa demande d'indemnisation, alors son usage dans la tarification constitue un biais de type 3 qui renforce des discriminations sociales existantes (Kiviat, 2019).

3. Les enjeux de l'apprentissage machine pour les biais en assurance

Les techniques de segmentation décrites dans la partie précédente, mises en place dans le courant du 20e siècle, impliquaient toujours l'intervention lourde de l'actuaire ou du statisticien, de ce fait responsable des biais de ses modèles. À partir des années 2000, avec l'émergence des données massives, on a de plus en plus recours à des techniques d'apprentissage machine, qui permettraient de remplacer l'humain par la machine dans un certain nombre de tâches : peut-

on en déduire pour autant que les biais seront réduits ? Rien n'est moins sûr. Et qu'en est-il, plus précisément en assurance ?

3.1 L'apprentissage machine en assurance : qu'est-ce qui change ?

Mesurer l'impact des données massives en assurance est peut-être plus difficile que dans d'autres domaines. D'un côté, comme toute autre organisation, les assureurs sont amenés à modifier leurs pratiques pour intégrer les nouvelles sources de données devenues accessibles, les capacités de calcul accrues et les nouveaux algorithmes. De l'autre pourtant, ces techniques apparaissent souvent comme la continuation d'une pratique de segmentation presque séculaire (Swedloff, 2014). De plus, certaines études montrent qu'à ce jour les modèles de tarification n'ont pas profondément changé, ni que de nouveaux produits n'ont émergé, suite par exemple à l'apparition des boitiers télématiques (Barry & Charpentier, 2020 ; François & Voldoire, 2022). L'étude ci-dessous tient donc plus d'une analyse de ce que les nouveaux modèles *rendent possibles*, même si le basculement, en assurance, n'a pas (encore ?) été observé en pratique.

La première modification qui vient nourrir l'apprentissage machine est l'apparition des données massives. À la différence de l'ère précédente, ces données ne sont plus obtenues via des questionnaires qui impliquaient un travail en amont dans le choix de ce que l'on voulait collecter et suivant quelle codification (Desrosières, 2008). Aujourd'hui, ces données sont obtenues via des senseurs, des objets connectés, ou sont nativement numériques car procédant d'actions en ligne – autant de sources qui ne demandent pas *a priori* d'intervention humaine. À la grande différence des données issues de questionnaires, ces données sont par ailleurs le plus souvent des données comportementales : pour les senseurs, et en se limitant à l'assurance des particuliers, on peut citer les boitiers télématiques qui collectent en continu la position, la vitesse et l'accélération du véhicule (Barry & Charpentier, 2020), ou les bracelets connectés mesurant des données biométriques de leurs porteurs (Lupton, 2014, 2016).

La deuxième modification majeure tient aux capacités de calcul des ordinateurs, sans commune mesure avec la génération précédente. Ainsi, lorsque De Wit et Van Eeghen (1984) évoquent la possibilité d'affiner la segmentation, ils s'appuient sur l'idée que « ***with the help of computers*** *it has become possible to make thorough risk analyses, and consequently to arrive at further premium differentiation* » (De Wit & Van Eeghen, 1984, p. 155, notre accentuation) : c'est l'existence même des ordinateurs qui, dans les années 80, changent la donne par rapport à une époque antérieure où les calculs étaient pratiquement manuels (Barry, 2020). Aujourd'hui, ce sont les capacités de calcul qui permettent le traitement de bases de données beaucoup plus importantes.

Enfin, l'apprentissage machine, quant à lui, permet d'automatiser une partie des tâches, notamment celle du choix des variables significatives, ce qui démultiplie le nombre de variables dont on peut tenir compte. Les modèles deviennent ainsi plus complexes, sans nécessairement changer de nature. C'est le cas par exemple avec les modèles de « *price optimization* », qui permettent de tenir compte dans la tarification non seulement du risque de l'assuré, mais aussi de sa sensibilité au prix de l'assurance et de sa propension à résilier son contrat. Ces modèles posent des problèmes nouveaux en termes d'équité, puisque ce serait les clients les plus loyaux qui se trouveraient pénalisés au profit d'assurés dans la même classe de risque, mais eux plus sensibles au prix de leur assurance (Frees & Huang, 2021).

Un saut conceptuel a lieu en revanche avec les algorithmes d'apprentissage profond (ou *deep learning*, pris ici comme une catégorie d'apprentissage machine). LeCun, Bengio et Hinton (2015) caractérisent en effet l'apprentissage profond par sa capacité à inférer seul les relations potentielles entre variables, antérieurement imposées aux données par l'analyste : « *the key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure* ».

Mis en perspective avec la partie précédente, l'apprentissage machine semble donc *a priori* lever les biais de type 1 et 2 qui résultaient des préjugés et stéréotypes de l'actuaire dans son choix et sa codification des variables. L'accès récent aux données comportementales semblent par ailleurs répondre au besoin de distinguer entre variables décrivant un choix conscient de l'assuré (son comportement) et celles relevant de caractéristiques intrinsèques auxquelles il ne peut rien.

Dans les conclusions de l'affaire Test-Achats, la juge critique l'imprécision des statistiques lorsque le risque serait en réalité individuel (et comportemental). Elle va ainsi dans le sens du mouvement général qui conçoit le risque comme associé au mode de vie, donc comme individuel, et non plus comme déterminé sur la base de classes statistiques (Rebert & Van Hoyweghen, 2015). C'est aussi l'objet du projet de loi étatsunien PAID (*Prohibit Auto Insurance Discrimination Act*), qui stipule que tout paramètre *non directement lié à la conduite* devrait être interdit dans la tarification du risque automobile (Metz, 2020). On retrouve ici, transposée à l'assurance, l'utopie que les algorithmes actuels seraient capables de personnaliser les décisions au niveau individuel, là où leurs ancêtres se contentaient de travailler sur des moyennes sur des sous-groupes (Lury & Day, 2019 ; Moor & Lury, 2018).

3.2 Les biais de l'apprentissage machine en assurance

Paradoxalement pourtant, la bascule de classes statistiques aux données massives (et comportementales) dans un but de personnalisation et d'ajustement du risque ne fait qu'exacerber les biais évoqués en première partie, tout en modifiant à la marge leur nature.

3.2.1 Refléter la réalité des risques

Le biais de type 1 consistait à se servir de données sans rapport avec le risque. Aujourd'hui, les compagnies d'assurances s'appuient de plus en plus sur des données de sources externes, pour essayer de mieux saisir la réalité (Charpentier, 2021). Dans l'utopie un peu mythologique du *big data*, ces données seraient enfin devenues exhaustives, permettant de rendre compte de la réalité de façon plus riche : sans compromis lié à l'échantillonnage, sans contraintes de volume de données (Mayer-Schönberger & Cukier, 2014), et sans la réduction du réel due au travail de quantification (Desrosières, 2008).

Mais l'un des problèmes essentiels liés à ces données tient au fait qu'elles résultent de l'observation et non d'expériences construites ad hoc (Charpentier, 2021 ; Rosenbaum, 2017) – d'où un biais d'échantillon. C'était déjà le point mis en avant par Ronald Fisher dans la polémique autour du tabagisme : pour lui, sans expérience randomisée qui permettrait de comparer des populations identiques de fumeurs et de non-fumeurs, l'observation d'une corrélation entre tabagisme et cancer ne prouve rien (Fisher, 1958). La réalité présentée par les données massives est elle-aussi filtrée, même si cela n'est plus le fait du statisticien qui construit sa base (Boyd & Crawford, 2012). Pour Barocas et Selbst (2016), les populations en marge de l'économie formelle et des activités en ligne sont nécessairement sous-représentées dans ces données, créant des risques de discrimination à leur égard.

De plus, comme l'écrivait déjà Desrosières (1993) à propos des statistiques classiques, « *les indicateurs quantitatifs rétroagissent sur les acteurs quantifiés* ». Le biais de rétroaction intervient lorsque les acteurs intègrent le fait qu'un paramètre fait l'objet d'une mesure pour la tarification : ils modifient alors leur comportement afin d'agir en retour sur le paramètre mesuré. Ce biais est magnifié lorsqu'il s'agit de variables comportementales. L'actuaire, qui n'a pas lui-même construit les bases de données auxquelles il a à présent accès, conçoit parfois mal ces limites.

Un autre biais d'échantillon est lié à l'auto-sélection induite par le RGPD (Charpentier, 2021).

En effet, jusque très récemment, les données en ligne étaient stockées automatiquement. Paradoxalement, le RGPD – dont le but essentiel est la protection des données personnelles –, conduit à un biais lié au non-consentement de certains : ceux qui en font la demande peuvent supprimer leurs données des bases collectées. Ce concept de *opting-out* peut fortement biaiser les données conservées.

3.2.2 Corrélation vs. causalité : l'efficacité opaque des nouveaux algorithmes

Dans les débats autour de l'intelligence artificielle, l'opacité des nouveaux algorithmes, décriée, est mise en balance avec leur précision accrue (Breiman, 2001). En 2017, lors de l'un des premiers de ces débats[5], l'un des participants avançait ainsi que « *if we wish to make AI systems deployed on self-driving cars safe, straightforward black-box models will not suffice, as we need methods of understanding their rare but costly mistakes* ». En réponse, Yann LeCun souligne que lorsqu'on présentait aux gens deux modèles (l'un parfaitement interprétable et précis à 90%, l'autre une boîte noire ayant une précision supérieure de 99%), ils choisissaient toujours le modèle plus précis. LeCun en conclut que « *people don't really care about interpretability but just want some sort of reassurance from the working model* ». Autrement dit, l'interprétabilité n'est pas importante si l'on est convaincu que le modèle fonctionne bien dans les conditions dans lesquelles il est censé fonctionner.

Pour Napoletani, Panza et Struppa (2011, p. 3), il s'agit d'un nouveau paradigme scientifique, ouvrant la voie vers une science devenue « *agnostique* ». À ce titre, dans un article désormais célèbre, Anderson (2008) parle de la « *fin des théories* » pour caractériser la nouvelle approche :

> « *Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise. But* ***faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete*** » (Anderson, 2008, notre accentuation).

Dans la perspective de ce papier, cela revient à renoncer au biais de type de 2 et à l'explicitation des relations entre variables, causales ou corrélées : la machine produit un score, suffisamment précis pour justifier l'abandon d'une interprétation par les données en entrée. Quelques exemples, encore assez rares, de cette approche boite noire existent en assurance. On peut penser à des applications en gestion de la fraude (quand envoyer un expert ?) ou en marketing (qui solliciter, ou quel produit proposer ?) : dans ces domaines, l'interprétation est largement négligée au profit de la rentabilité. Pour la tarification, certains algorithmes de reconnaissance d'images peuvent aujourd'hui inférer des facteurs de risque. Par exemple, Kita-Wojciechowska et Kidziński (2019) proposent de prédire la fréquence d'accident automobile à partir d'images-satellite du lieu d'habitation du conducteur ; ou encore Shikhare (2021) calcule un score de santé sur la base d'une photo-portrait de la personne.

3.2.3 Corriger les discriminations sociales : le paradoxe des variables protégées dans un environnement de données massives

Dans les modèles boite noire, les biais dus aux préjugés et stéréotypes du statisticien que Works (1977) essayait d'éviter en recommandant l'usage de variables causales, seraient écartés puisque c'est l'algorithme qui établit des liens (inconnus) entre les variables devenues pléthoriques. En réalité, on sait aujourd'hui qu'au contraire les préjugés, stéréotypes et autres discriminations se retrouvent dans les données elles-mêmes, donc bien en amont du jugement des statisticiens : au-delà des biais d'échantillon évoqués plus haut, c'est vraiment la nature des données qui est

5. Appelé « *The Great AI Debate: Interpretability is necessary for machine learning* », opposant Rich Caruana et Patrice Simard (pour) à Kilian Weinberger et Yann LeCun (contre) : https://youtu.be/93Xv8vJ2acI.

en cause (Caliskan *et al*., 2017).

De plus, alors que dans les modèles classiques on pouvait espérer corriger les biais en interdisant l'usage de certaines variables dites protégées, la colinéarité de ces variables avec d'autres, facialement neutres, dans les données massives rend cette « protection » illusoire : « *thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait* ***whether or not that information is an input*** » (Barocas et Selbst, 2016, notre accentuation).

Pour Prince et Schwarcz (2019), la discrimination par procuration, évoquée déjà pour les modèles classiques, est magnifiée par les nouveaux algorithmes. Alors qu'elle était intentionnelle par le passé (puisqu'une décision humaine présidait au choix des variables – par exemple le *red-lining*), la discrimination par procuration devient non-intentionnelle. Suivant la distinction établie par Barocas et Selbst (2016), la discrimination par procuration ne résulte plus d'un traitement consciemment différencié des segments protégés (*disparate treatment*), mais elle fait partie des discriminations transparentes dont on ne perçoit que les effets *ex-post* (*disparate impact*). Ce phénomène est inévitable, en particulier lorsqu'une variable directement liée au phénomène (une variable causale) est absente des données. C'est ce qui menace de se produire lorsque l'on exclut des variables causales mais protégées car reflétant un aléa subi – par exemple les données génétiques en assurance santé (Prince & Schwarcz, 2019, p. 1264) :

> « *An AI deprived of information about a person's genetic test results or obvious proxies for this information (like family history) will use other information—ranging from TV viewing habits to spending habits to geolocational data—to proxy for the directly predictive information contained within the genetic test results* » (Prince & Schwarcz, 2019, p. 1274).

On risque alors de créer des algorithmes qui associent le visionnage de certains programmes télévisés à un facteur de risque en santé !

Pour lutter contre ce phénomène, Williams, Brooks et Shmargad (2018) montrent que paradoxalement, il ne faut pas interdire la collecte et l'usage des variables protégées, mais au contraire s'en servir comme moyen de piloter la non-discrimination. C'est de cette manière par exemple que les Britanniques appréhendent les données ethniques, par opposition à la France (Ducourtieux, 2021).

4. Apprentissage machine et équité assurantielle : collective ou individuelle ?

L'équité est l'autre face de la médaille de ce qui est perçu comme biais ou discrimination dans un contexte culturel et historique donné. Dans quelle mesure les technologies disponibles influent-elles sur cette conception de la justice ? Est-ce que, notamment, l'environnement des données massives et des nouveaux algorithmes modifie la conception de l'équité assurantielle ?

Pour Thiery et Schoubroeck (2006), les juristes et les actuaires ont des conceptions fondamentalement différentes de l'équité. L'équité assurantielle et la segmentation reposeraient sur une vision collective de l'équité, alors que l'équité juridique met en avant les droits individuels. Juridiquement, le droit à l'égalité de traitement est octroyé à une personne en sa qualité d'individu, qui ne peut être traité différemment en raison de son appartenance à tel groupe ou tel groupe. Mais cette vision s'oppose fondamentalement à l'approche actuarielle qui, historiquement, analyse les risques et calcule les primes en termes collectifs (Ewald, 2011).

Ainsi, dans sa décision sur l'affaire Norris, le juge maintient qu'une classification statistiquement valide (qu'il s'agisse d'un lien causal ou d'une corrélation) n'en fait pas une classification

légitime. En réalité, aucune classification ne peut l'être puisque « *even a true generalization about class cannot justify class-based treatment. An individual woman may not be paid lower monthly benefits simply because* women as a class *live longer than men* » (cité dans Horan, 2021, p. 187). Pour l'individu auquel on l'impose, la classification constitue toujours une discrimination, dite statistique (Binns, 2018), ou une généralisation arbitraire de l'individu à un groupe.

Pour Simon (1988) et Horan (2021), l'adoption de ce point de vue individuel par le juge a contribué à renforcer l'effacement du principe de solidarité pourtant au cœur de la pratique assurantielle. L'assurance repose en effet sur la mise en commun de l'incertitude et s'appuie sur le voile d'ignorance qui met les uns et les autres à égalité devant l'aléa (Ewald, 1986). Mais une fois posée l'existence d'un risque individuel qu'il faudrait approcher, par la classification puis par les nouveaux algorithmes, la tarification devient un exercice mathématique d'optimisation et de minimisation de la variance portée par l'assureur. La « personnalisation » associée aux nouveaux algorithmes devient en assurance « l'individualisation » du risque (Barry & Charpentier, 2020). Avec cette dernière, la distinction entre équité assurantielle-collective et équité juridique-individuelle tend à disparaître (Barry, 2020). Car même si les actuaires n'ont pas fondamentalement bouleversé leur pratique, la notion d'équité assurantielle semble évoluer avec les nouvelles technologies. Dans les produits télématiques, le risque n'est plus présenté comme une incertitude mise en commun, mais comme un choix individuel. Chacun devrait payer en fonction de son comportement de chacun, et non plus de données démographiques agrégées. L'équité dans ce cas consiste à ajuster la prime au comportement individuel, pour que chacun paie suivant « son » risque (Meyers & Van Hoyweghen, 2018). Dans cette perspective, le biais statistique évoqué par le juge dans l'affaire Norris prend une importance renouvelée.

Mais cette individualisation, si elle a lieu, est problématique à plus d'un titre ; il n'est pas évident tout d'abord que le résultat soit équitable pour tous les assurés concernés. Elle conduirait en effet à des tarifications plus disparates, avec des primes pour les individus perçus comme les plus risqués qui pourraient devenir inabordables, les excluant de fait de la communauté assurée (Charpentier, Barry & Gallic, 2020). Il n'est pas évident non plus que le machine learning puisse résoudre cette tension entre équité individuelle et collective.

Face à l'opacité des modèles et aux biais sociaux embarqués dans les données massives, l'équité algorithmique émerge ainsi comme une nouvelle discipline (Kusner & Loftus, 2020). On retrouve dans cette littérature la tension entre point de vue individuel ou collectif au cœur des questionnements actuels sur l'individualisation du risque en assurance. L'exactitude (mathématique) d'un algorithme se mesure en général à partir d'une matrice de confusion, qui permet d'observer les erreurs par type – faux négatifs et faux positifs. Mais la minimisation simultanée de ces erreurs n'est pas toujours possible, voire souhaitable, pour plusieurs raisons. En effet, faux positifs et faux négatifs ne sont pas comparables d'un point de vue éthique : la condamnation d'un innocent n'a pas la même « valeur » que la libération d'un coupable. Ainsi, suivant le contexte, il faudra choisir de minimiser l'une ou l'autre forme d'erreur.

Les choses se compliquent encore lorsque l'on tient compte des variables protégées. Pessach et Shmueli (2020) distinguent alors entre des indicateurs d'équité collectifs ou individuels. Les indicateurs collectifs visent à assurer la parité entre groupes, protégés ou non. On peut ainsi tenter de s'assurer que les fréquences de prédiction (exacte ou positive) soient égales sur les deux groupes ; ou mesurer les taux de faux positifs et faux négatifs séparément sur les deux groupes et vérifier qu'ils sont voisins. Ce n'était pas le cas, par exemple, pour l'algorithme étatsunien COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions* ou algorithme prédictif d'aide à la décision des juges sur le risque de récidive). Il présentait en effet un taux de faux positifs (faussement classés récidivistes) beaucoup plus élevé pour les noirs, et un taux de faux négatifs plus élevé pour les blancs, avec des précisions égales sur les deux groupes (Kleinberg *et al*., 2016). Les indicateurs individuels, eux, visent à s'assurer

que des individus similaires (hors variable protégée) obtiennent un score similaire. Kusner et Loftus (2020) définissent de la sorte l'équité « contrefactuelle », qui consiste à comparer les scores de deux observations identiques sur lesquelles seule la variable protégée prend une valeur différente. Cette technique permet de répondre aussi précisément que possible à la question « que se serait-il passé si seul l'attribut protégé avait été différent ? ». Tous les auteurs s'accordent sur le fait que ces différents indicateurs ne peuvent pas être optimisés simultanément, conduisant à de nécessaires compromis en fonction du contexte (Kleinberg *et al*., 2016 ; Pessach & Shmueli, 2020).

Le ban du paramètre homme/femme par la directive européenne exemplifie ces dilemmes en assurance : soit on ignore la variable, mais alors si une différence statistique existe elle réapparaîtra au travers d'autres variables, colinéaires au paramètre interdit et par conséquent la moyenne sur les hommes et les femmes restera différente ; soit au contraire on utilise cette variable pour maintenir des moyennes identiques mais alors toutes choses égales par ailleurs, le tarif variera en fonction du sexe de la personne. On ne pourra jamais maintenir la parité entre les groupes et assurer l'équité contrefactuelle.

Plusieurs approches statistiques sont en train de se faire jour pour corriger le biais une fois identifié. En amont du modèle (*pre-processing*), il est possible de modifier les bases d'apprentissage pour que les algorithmes entraînés ensuite passent les tests, par exemple en utilisant des poids. La correction peut aussi se faire pendant la modélisation, en ajoutant une contrainte dans la fonction optimisée. Cette méthode, classique pour éviter le surapprentissage, est ici appliquée à l'équité de l'algorithme. Elle consiste à pénaliser la fonction de coût avec un terme rendant compte du niveau d'équité suivant un indicateur de parité défini à l'avance (Bechavod & Ligett, 2018). Enfin, des corrections en aval (*post-processing*) sont possibles, en ajustant les décisions. Par exemple dans une classification basée sur un score, il sera utilisé avec des seuils différents en fonction de la variable sensible. Pour l'attribution d'un crédit, par exemple, on l'autorisera pour la catégorie favorisée si leur score dépasse 60%, mais pour la catégorie défavorisée on pourra l'autoriser s'il dépasse 55% (Charpentier, 2022).

Dans tous les cas, la variable sensible est nécessaire à l'identification et la correction du biais. Pour Charpentier (2021, p. 148), interdire l'usage de la variable protégée est contre-productif car « *dans la plupart des cas réalistes, non seulement la suppression de la variable sensible ne rend pas les modèles de régression équitables, mais au contraire, une telle stratégie est susceptible d'amplifier la discrimination* ».

5. Conclusion

L'équité assurantielle est une notion dynamique, dont on a vu ici qu'elle dépendait de contextes historiques, culturels et techniques divers. En pleine ère industrielle, on s'appuyait sur le voile d'ignorance pour justifier des couvertures très larges en termes de solidarité, et sur l'idée de l'égalité du plus grand nombre face à une adversité mal connue. Cette équité était critiquée par les libéraux qui y voyaient une incitation à la licence. Dans le courant du 20^{e} siècle, avec les capacités croissantes de collecte et de calcul se mettent en place des modèles segmentés, qui assoient l'assurance sur la classification des risques, perçus comme des groupes homogènes de personnes qui se ressemblent. A partir des années 80, des controverses se font jour autour de l'usage de telle ou telle variable, controverses qui constituent le lit des critiques actuelles concernant les biais et les discriminations associées à l'apprentissage machine.

L'examen de cette histoire permet d'identifier quelques familles principales de biais, dans les pratiques traditionnelles de classification puis leur déclinaison dans les algorithmes de machine-learning. On distingue ainsi avant tout les critiques qui admettent la classification dans son principe, mais contestent l'usage de telle ou telle variable. Ces critiques sont de deux ordres :

- On critique tout d'abord des variables reflétant les préjugés du statisticien qui choisit de les collecter pour créer son modèle, même lorsqu'elles n'ont aucun lien avec le phénomène à étudier. C'est le cas de la couleur de peau aux États-Unis dans les produits d'assurance-vie à la fin du 19e siècle. Ce type de biais disparait en principe avec les données massives : étant nativement numériques, elles court-circuitent le travail de quantification de la période précédente. Mais on s'est rendu compte, au cours de ces vingt dernières années, que les préjugés ont la vie dure et que les discriminations sociales se retrouvent dans les données. Un usage aveugle de l'apprentissage machine conduirait alors à reproduire ces biais dans les modèles.
- Une autre forme de discrimination mise au jour dans les années 60-80 et que l'on retrouve magnifiée avec les nouveaux algorithmes tient à l'usage de variables corrélées sans être causales : ainsi l'usage des paramètres homme/femme, le score de crédit ou le critère fumeur/non-fumeur ont provoqué des controverses dont certaines se poursuivent jusqu'à aujourd'hui. La solution préconisée par les critiques, sûrement inopérable en pratique, serait de se limiter à des variables purement causales. Cette exigence de causalité avérée est totalement abandonnée dans les algorithmes de machine-learning, dont certains disent qu'ils signent l'avènement d'un nouvel *épistémê* : ils se contentent en effet de mettre en évidence des corrélations entre les données en entrée, sans même expliciter ces liens. Ceci conduit à un biais nouveau lié à ces techniques, celui de leur opacité, même s'il est la contrepartie d'une plus grande précision.

Une autre grande famille de critiques rejette peu ou prou la classification :

- Dans les modèles classiques, l'équité de l'assurance exigeait que certaines variables causales soient exclues de l'analyse parce qu'elles reflètent un aléa subi et non choisi par la personne : l'usage de données génétiques en assurance santé par exemple est interdit dans la plupart des pays. Dans ce cas, l'assurance est perçue comme un moyen non plus de refléter le risque mais, en éliminant la variable des modèles, de le faire porter par l'ensemble de la population assurée. La solution de l'élimination des variables protégées, si elle est effective dans les modèles traditionnels, est beaucoup plus difficile à mettre en œuvre avec les données massives et l'apprentissage machine, respectivement parce que les variables protégées sont captées via leur colinéarité avec d'autres, et que l'opacité des algorithmes rend la mise en évidence de ces discriminations plus complexe.
- Plus fondamentalement, une critique légaliste opposait traditionnellement les droits de l'individu à la classification, soit encore une approche individuelle de l'équité à celle collective portée par l'assurance, mettant en avant le biais statistique induit par la réduction, nécessairement arbitraire, d'un individu aux données d'une classe. Avec les données massives dont certaines sont comportementales, l'utopie est de résoudre ce biais, en personnalisant et individualisant les modèles. Mais là aussi, la promesse n'est pas tenue : les théoriciens de l'équité algorithmique mettent en avant l'impossibilité d'optimiser les algorithmes sur divers critères simultanés, dont aucun ne peut être *a priori* préféré à un autre. Dans le contexte assurantiel, l'équité individuelle menace cependant de conduire à des tarifs de plus en plus différenciés, donc inabordables pour certaines personnes classées très risquées.

Faut-il alors en rester aux bonnes vieilles tables de tarification, pour lesquelles tous les paramètres sont explicites, connus à l'avance et par là-même ouverts à la contestation ? Un peu comme toute théorie scientifique se doit d'être falsifiable, une tarification se devrait d'être transparente afin d'être contestable. C'est à cette exigence de contestabilité que doivent répondre aujourd'hui les nouveaux algorithmes.

Références

Anderson C. (2008), « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*, https://www.wired.com/2008/06/pb-theory/.

Austin R. (1983), « The Insurance Classification Controversy », *University of Pennsylvania Law Review*, 131(3), pp. 517-582, https://doi.org/10.2307/3311844.

Avraham R. (2017), « Discrimination and Insurance », SSRN Scholarly Paper ID 3089946, Social Science Research Network, https://papers.ssrn.com/abstract=3089946.

Baker T. (2002), *Risk, Insurance, and (the Social Construction of) Responsibility*, University of Connecticut School of Law Articles and Working Papers, http://lsr.nellco.org/uconn_wps/8.

Baker T. and Simon J. (2002), « Embracing Risk », *in* Baker T. and Simon J. (eds.), *Embracing Risk: The Changing Culture of Insurance and Responsibility*, Chicago, University of Chicago Press, pp. 1-25.

Barocas S. and Selbst A. D. (2016), « Big Data's Disparate Impact Essay », *California Law Review*, 104, pp. 671-732.

Barry L. (2020), « Insurance, Big Data and Changing Conceptions of Fairness », *European Journal of Sociology / Archives Européennes de Sociologie*, 61(2), pp. 159-184, https://doi.org/10.1017/S0003975620000089.

Barry L. and Charpentier A. (2020), « Personalization as a promise: Can Big Data change the practice of insurance? », *Big Data & Society*, January-June, pp. 1-12, https://journals.sagepub.com/doi/full/10.1177/2053951720935143.

Bartik A. and Nelson S. (2019), « Deleting a Signal: Evidence from Pre-Employment Credit Checks », Working Paper N° 2019-137, Chicago, Becker Friedman Institute - University of Chicago, https://www.ssrn.com/abstract=3498458.

Bechavod Y. and Ligett K. (2018), « Penalizing Unfairness in Binary Classification », arXiv:1707.00044, https://doi.org/10.48550/arXiv.1707.00044.

Benjamin B. and Michaelson R. (1988), « Mortality differences between smokers and non-smokers », *Journal of the Institute of Actuaries*, 115(3), pp. 519-525, https://doi.org/10.1017/S0020268100042797.

Binns R. (2018), « Fairness in Machine Learning: Lessons from Political Philosophy », *Conference on Fairness, Accountability and Transparency*, pp. 149-159, http://proceedings.mlr.press/v81/binns18a.html.

Bouk D. (2015), *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*, Chicago, University Of Chicago Press.

Boyd D. and Crawford K. (2012), « Critical Questions for Big Data », *Information, Communication and Society*, 15(5), pp. 662-679, https://doi.org/10.1080/1369118X.2012.678878.

Breiman L. (2001), « Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author) », *Statistical Science*, 16(3), pp. 199-231, https://doi.org/10.1214/ss/1009213726.

Caliskan A., Bryson J. J., and Narayanan A. (2017), « Semantics derived automatically from language corpora contain human-like biases », *Science*, 356(6334), pp. 183-186, https://doi.org/10.1126/science.aal4230.

Charpentier A. (2021), *Assurance : Bais, Discrimination & Équité*, unpublished manuscript.

Charpentier A. (2022), « Assurance : Discrimination, biais et équité », *Opinions & Débats*, 25, Institut Louis Bachelier.

Charpentier A., Barry L. et Gallic E. (2020), « Quel avenir pour les probabilités prédictives en assurance ? », *Annales des Mines – Réalités industrielles*, 2020(1), pp. 74-77.

Charpentier A., Barry L., and James M. (2020), « Insurance against Natural Catastrophes: Balancing Actuarial Fairness and Social Solidarity », Working Paper no 22, Paris, Chaire PARI.

Charpentier A., Denuit M. M. et Elie R. (2015), « Segmentation et Mutualisation, les deux faces d'une même pièce », *Risques*, 103, pp. 19-23.

CURIA (2010), Test-Achats – Conclusions de l'Avocat General, https://curia.europa.eu/juris/document/document_print.

De Pril N. and Dhaene J. (1996), *Segmentering in verzekeringen*, KUL, Departement toegepaste economische wetenschappen.

De Wit G. W. and Van Eeghen J. (1984), « Rate Making and Society's Sense of Fairness », *ASTIN Bulletin*, 14(2), pp. 151-164.

Denuit M. et Charpentier A. (2004), *Mathématiques de l'assurance non-vie : Principes fondamentaux de théorie du risque*, Paris, Economica.

Desrosières A. (1993), *La Politique des Grands Nombres. Histoire de la Raison Statistique*, Paris, La découverte.

Desrosières A. (2008), *L'argument statistique. I, Pour une sociologie historique de la quantification*, Paris, Presses de l'école des Mines.

Doll R. and Hill A. B. (1964), « Mortality in Relation to Smoking: Ten Years' Observations of British Doctors », *British Medical Journal*, 1(5396), pp. 1460-1467.

Ducourtieux C. (2021), « Les statistiques ethniques au Royaume-Uni, un outil essentiel pour lutter contre les inégalités », *Le Monde.fr*, avril 2021, https://www.lemonde.fr/economie/article/2021/04/22/les-statistiques-ethniques-au-royaume-uni-un-outil-essentiel-pour-lutter-contre-les-inegalites_6077646_3234.html?xtor&&M_BT=36351134033493#x3D;EPR-33281062-[la-lettre-eco]-20210422-.

Dworkin R. (1981), « What is Equality? Part 2: Equality of Resources », *Philosophy & Public Affair*s, 10(4), pp. 283-345.

Ewald F. (1986), *L'État Providence*, Paris, Grasset.

Ewald F. (2011), « Omnes et Singulatim. After Risk », *Carceral Notebooks*, 7, pp. 77-107.

Fisher R. A. (1958), « Cancer and Smoking », *Nature*, 182(4635), pp. 596-596, https://doi.org/10.1038/182596a0.

François P. (2021), « Catégorisation, individualisation. Retour sur les scores de crédit », Working Paper no 24, Paris, Chaire PARI, https://www.chaire-pari.fr/wp-content/uploads/2021/10/WP-24-categorisation-individualisation.pdf.

François P. and Voldoire T. (2022), « The revolution that did not happen. Telematics and car insurance in the 2010s », Working Paper no 26, Paris, Chaire PARI.

Frees E. W. (Jed) and Huang F. (2021), « The Discriminating (Pricing) Actuary », *North American Actuarial Journal*, 0(0), pp. 1-23, https://doi.org/10.1080/10920277.2021.1951296.

Frezal S. and Barry L. (2020), « Fairness in Uncertainty: Some Limits and Misinterpretations of Actuarial Fairness », *Journal of Business Ethics*, 167(1), pp. 127-136, https://doi.org/10.1007/s10551-019-04171-2.

Glenn B. J. (2000), « The Shifting Rhetoric of Insurance Denial », *Law & Society Review*, 34(3), pp. 779-808, https://doi.org/10.2307/3115143.

Glenn B. J. (2003), « Postmodernism: The Basis of Insurance », *Risk Management & Insurance Review*, 6(2), pp. 131-143, https://doi.org/10.1046/J.1098-1616.2003.028.x.

Gosseries A. (2014), « What Makes Age Discrimination Special? A Philosophical Look at the ECJ Case Law », *Netherlands Journal of Legal Philosophy*, 43(1), pp. 59-80, https://doi.org/10.5553/NJLP/221307132014043001005.

Guseva A. and Rona-Tas A. (2001), « Uncertainty, Risk, and Trust : Russian and American Credit Card Markets Compared », *American Sociological Review*, 66(5), pp. 623-646, https://doi.org/10.2307/3088951.

Heen M. (2009), « Ending Jim Crow Life Insurance Rates », *Northwestern Journal of Law & Social Policy*, 4(2), pp. 360-399.

Heller D. (2015), « High Price of Mandatory Auto Insurance in Predominantly African American Communities », Consumer Federation of America, https://consumerfed.org/reports/high-price-of-mandatory-auto-insurance-in-predominantly-african-american-communities/.

Hoffman F. L. (1931), « Cancer and Smoking Habits », *Annals of Surgery*, 93(1), pp. 50-67.

Horan C. D. (2021), *Insurance Era: Risk, Governance, and the Privatization of Security in Postwar America*, Chicago, University of Chicago Press.

Johnston L. (1945), « Effects of Tobacco Smoking on Health », *British Medical Journal*, 2(4411), pp. 98.

Kita-Wojciechowska K. and Kidziński L. (2019), « Google Street View image predicts car accident risk », *Central European Economic Journal*, 6(53), pp. 152-163.

Kiviat B. (2019), « The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores », *American Sociological Review*, 84(6), pp. 1134-1158, https://doi.org/10.1177/0003122419884917.

Kleinberg J., Mullainathan S., and Raghavan M. (2016), « Inherent Trade-Offs in the Fair Determination of Risk Scores », arXiv:1609.05807, https://arxiv.org/abs/1609.05807v2.

Knights D. and Vurdubakis T. (1993), « Calculations of risk: Towards an understanding of insurance as a moral and political technology », *Accounting, Organizations and Society*, 18(7), pp. 729-764, https://doi.org/10.1016/0361-3682(93)90050-G.

Kranzberg M. (1986), « Technology and History: "Kranzberg's Laws" », *Technology and Culture*, 27(3), pp. 544-560, https://doi.org/10.2307/3105385.

Krippner G. R. and Hirschman D. (2022), « The person of the category: The pricing of risk and the politics of classification in insurance and credit », *Theory and Society*, pp. 1-43, https://doi.org/10.1007/s11186-022-09500-5.

Kusner M. J. and Loftus J. R. (2020), « The long road to fairer algorithms », *Nature*, 578(7793), pp. 34-36, https://doi.org/10.1038/d41586-020-00274-3.

Larson J., Angwin J., Kirchner L., and Mattu S. (2017), « How We Examined Racial Discrimination in Auto Insurance Prices », *ProPublica*, https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-methodology?token=oXaDaCvsdX3ZY7-YJd8F3L-6fSTJ6BUj.

LeCun Y., Bengio Y., and Hinton G. (2015), « Deep learning », *Nature*, 521(7553), pp. 436-444, https://doi.org/10.1038/nature14539.

Lupton D. (2014), « Self-Tracking Modes: Reflexive Self-Monitoring and Data Practices », SSRN Scholarly Paper ID 2483549, *Social Science Research Network*, https://papers.ssrn.com/abstract=2483549.

Lupton D. (2016), « The diverse domains of quantified selves: Self-tracking modes and dataveillance », *Economy and Society*, 45(1), pp. 101-122, https://doi.org/10.1080/03085147.2016.1143726.

Lury C. and Day S. (2019), « Algorithmic Personalization as a Mode of Individuation », *Theory, Culture & Society*, 36(2), pp. 17-37, https://doi.org/10.1177/0263276418818888.

Mayer-Schönberger V. and Cukier K. (2014), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston, Eamon Dolan/Mariner Books.

Mercat-Bruns M. (2020), « Les rapports entre vieillissement et discrimination en droit : Une fertilisation croisée utile sur le plan individuel et collectif », *La Revue des droits de l'homme. Revue du Centre de recherches et d'études sur les droits fondamentaux*, 17, Article 17, https://doi.org/10.4000/revdh.8641.

Metz J. (2020), « Sen. Booker's PAID Act Looks To Eliminate Discriminatory Non-Driving Factors In Auto Insurance Pricing », Forbes Advisor, 5/10/2020, https://www.forbes.com/advisor/car-insurance/paid-act/.

Meyers G. and Van Hoyweghen I. (2018), « Enacting Actuarial Fairness in Insurance : From Fair Discrimination to Behaviour-based Fairness », *Science as Culture*, 27(4), pp. 413-438, https://doi.org/10.1080/09505431.2017.1398223.

Miller G. H. and Gerstein D. R. (1983), « The life expectancy of nonsmoking men and women », *Public Health Reports (Washington, D.C.: 1974)*, 98(4), pp. 343-349.

Miller M. J. (2009), « Disparate Impact and Unfairly Discriminatory Insurance Rates », *Casualty Actuarial Society E-Forum, Winter 2009*, https://www.casact.org/pubs/forum/09wforum/.

Miller M. J. and Smith R. A. (2003), « The Relationship of Credit-Based Insurance Scores to Private Passenger Automobile Insurance Loss Propensity », EPIC Actuaries, LLC, https://www.progressive.com/content/PDF/shop/EPIC-CreditScores.pdf.

Moor L. and Lury C. (2018), « Price and the person: Markets, discrimination, and personhood », *Journal of Cultural Economy*, 11(6), pp. 501-513, https://doi.org/10.1080/17530350.2018.1481878.

Napoletani D., Panza M. and Struppa D. C. (2011), « Agnostic Science. Towards a Philosophy of Data Analysis », *Foundations of Science*, 16(1), pp. 1-20, https://doi.org/10.1007/s10699-010-9186-7.

O'Neil C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Crown.

Patterson J. T. (1989), *The Dread Disease: Cancer and Modern American Culture*, Harvard, Harvard University Press.

Pessach D. and Shmueli E. (2020), « Algorithmic Fairness », arXiv:2001.09784, https://arxiv.org/abs/2001.09784v1.

Prince A. E. R. and Schwarcz D. (2019), « Proxy Discrimination in the Age of Artificial Intelligence and Big Data », *Iowa Law Review*, 105, pp. 1257-1318.

Rebert L. and Van Hoyweghen I. (2015), « The right to underwrite gender. The Goods & Services Directive and the politics of insurance pricing », *Tijdschrift Voor Genderstudies*, 18(4), pp. 413-431.

Rosenbaum P. (2017), *Observation and Experiment: An Introduction to Causal Inference*, Harvard, Harvard University Press, .

Rudin C. and Radin J. (2019), « Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition », *Harvard Data Science Review*, 1(2), https://doi.org/10.1162/99608f92.5a8a3a3d.

Schauer F. (2003), *Profiles, Probabilities, and Stereotypes*, Harvard, Harvard University Press, https://doi.org/10.2307/j.ctvjz82xm.

Schmeiser H., Störmer T. and Wagner J. (2014), « Unisex Insurance Pricing: Consumers' Perception and Market Implications », *The Geneva Papers on Risk and Insurance – Issues and Practice*, 39(2), pp. 322-350, https://doi.org/10.1057/gpp.2013.24.

Shikhare S. (2021), « AI Enabled Next Generation LTC and Life Insurance Underwiring Using Facial Score Model », *Insurance Data Science Conference 2021*, 19, https://insurancedatascience.org/downloads/London2021/Session_4b/Shrinivas_Shikhare.pdf.

Simon J. (1988), « The Ideological Effects of Actuarial Practices », *Law Social Review*, 22, pp. 771-800.

Society of Actuaries (SOA) (1982), *Report of the Task Force on Smoker/Non Smoker Mortality*, Transactions of Society of Actuaries, https://www.soa.org/globalassets/assets/library/research/transactions-reports-of-mortality-moribidity-and-experience/1980-89/1982/january/TSR8210.pdf.

Swedloff R. (2014), « Risk Classification Big Data (R)Evolution », *Connecticut Insurance Law Journal*, 21(1), pp. 339-373.

Thiery Y. and Schoubroeck C. V. (2006), « Fairness and Equality in Insurance Classification », *The Geneva Papers on Risk and Insurance – Issues and Practice*, 31(2), pp. 190-211, https://doi.org/10.1057/palgrave.gpp.2510078.

Walters M. A. (1981), « Risk Classification Standards », *Proceedings of the Casualty Actuarial Society*, 68, pp. 1-23.

Wiggins B. A. (2013), *Managing risk, managing race: Racialized actuarial science in the United States, 1881-1948* [Minnesota], http://conservancy.umn.edu/handle/11299/159587.

Williams B. A., Brooks C. F. and Shmargad Y. (2018), « How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications », *Journal of Information Policy*, 8, pp. 78-115, https://doi.org/10.5325/jinfopoli.8.2018.0078.

Works R. (1977), « Whatever's FAIR—Adequacy, Equity, and the Underwriting Prerogative in Property Insurance Markets », *Nebraska Law Review*, 56(3), pp. 445-464.

L'équité dans la machine ou comment le *machine learning* devient scientifique en tournant le dos au réalisme métrologique

Bilel BENBOUZID[1]
Maître de conférences en sociologie, Université Paris Est, Marne-la-Vallée, Laboratoire Interdisciplinaire, Science, Innovation et Société (LISIS)

TITLE
Fairness in the machine or how machine learning becomes scientific by turning its back on metrological realism

RÉSUMÉ
Nous soutenons dans cet article que la prise en compte de la *fairness* dans le *machine learning* (FairML) est un bon observatoire de la politique des statistiques et de leur transformation actuelle. Si les statisticiens classiques ont longtemps veillé à ce que leurs outils de mesure ne contiennent aucune trace politique, les *data scientists*, développeurs de machines prédictives, sont désormais contraints, par les problèmes d'équité qu'ils doivent traiter, de faire converger deux postures souvent distinctes : la recherche exigeante de la fiabilité des procédures de calcul et le souci de transparence du caractère construit et politiquement situé des opérations de quantification. Après avoir localisé socialement la formation du domaine FairML et décrit le cadre épistémologique particulier dans lequel il s'inscrit, nous verrons dans un second temps comment, concrètement, les chercheurs parviennent à penser à la fois construction mathématique et construction sociale des approches, à l'issue de controverses sur les métriques d'équité et leur statut dans l'apprentissage machine. Nous montrerons enfin que les approches du FairML tendent vers une forme d'objectivité spécifique, celle du « jugement exercé », reposant sur une perspective partielle et une justification raisonnablement partiale du concepteur de la machine – cette dernière devenant dès lors elle-même située politiquement.

Mots-clés **:** *analyse des controverses, discrimination algorithmique, équité dans le machine learning, métriques, objectivité, sociologie de la quantification, vertus épistémiques.*

ABSTRACT
In this paper we argue that Fairness in Machine Learning (FairML) is a good observatory for the politics of statistics and its current transformation. While classical statisticians have long been careful to ensure that their measurement tools do not contain any political traces, data scientists, as developers of predictive machines, are now forced by the fairness problems they have to deal with to converge two often distinct stances: the demanding search for the reliability of computational procedures and the concern for transparency of the constructed and politically situated character of quantification operations. Having socially situated the emergence of the FairML field and described the particular epistemological framework in which it is embedded, we will then see how researchers concretely deals with both the mathematical construction and the social construction of approaches. We describe how FairML approaches emerge from controversies about equity metrics and their status in machine learning. Finally, we will show that these approaches tend towards a specific form of objectivity, that of 'exercised judgement', based on a partial perspective and a reasonably biased justification of the machine designer - the machine thus becoming politically situated itself.

Keywords**:** *controversy analysis, algorithmic discrimination, fairness in machine learning, metrics, objectivity, sociology of quantification, epistemic virtues.*

1. bilel.benbouzid198@gmail.com

L'équité (*fairness*) des algorithmes est un des enjeux majeurs de la régulation de l'intelligence artificielle. Si de nombreuses études ont montré que dans des situations de « jugement », les évaluations des machines, reposant sur des procédures de calcul, sont moins biaisées que celles des humains engagés dans des processus sociaux, affectifs et comportementaux (Kleinberg *et al*., 2018), d'autres ont alerté des risques de discrimination systématique par l'automatisation accrue des décisions algorithmiques (Eubanks, 2018 ; Noble, 2018 ; O'Neil, 2016 ; Pasquale, 2015). Depuis une dizaine d'années, l'avalanche de controverses sur le caractère raciste et sexiste des algorithmes (Crawford, 2021) a soulevé de nombreuses questions : comment peut-on faire confiance à des techniques de calcul pour prendre des décisions à l'égard d'une personne, par exemple, lors de la sélection à un entretien d'embauche, l'admission à l'université, la mise en liberté conditionnelle ou l'octroi d'un crédit ? Comment s'assurer que les systèmes d'intelligence artificielle (reconnaissance faciale, traduction automatique, etc.) reposant sur des procédures d'apprentissage statistique ne reproduisent pas les mécanismes sociaux indésirables qui se logent dans la production des données d'entrainement ? Comment les personnes peuvent-elles savoir si elles sont traitées équitablement par ces algorithmes par rapport à toutes les autres personnes qui les entourent ? Sur quels critères à la fois statistiques et juridiques peuvent-elles s'appuyer pour faire valoir leur droit à un traitement algorithmique équitable ? Et quelles sont les normes techniques et juridiques à respecter par les fabricants de machines pour garantir le respect des lois antidiscriminatoires ?

Un domaine nouveau dit FairML (*Fairness in Machine Learning*) tente désormais de répondre à ces questions. Il se manifeste notamment par une conférence scientifique annuelle, l'*ACM Fairness, Accountability and Transparency Conference* (*FAccT*), autour de laquelle se constitue un réseau de chercheurs, à la croisée des sciences computationnelles, du droit, de la philosophie politique et des sciences sociales (Laufer *et al*., 2022a). Depuis une dizaine d'années, cette communauté de recherche se concentre sur l'identification des biais de discrimination, les définitions des métriques d'équité comme référence pour atténuer les biais, les problèmes d'intelligibilité des algorithmes et les rapports étroits entre système algorithmique et politique.

Le FairML est un bon observatoire de la politique des statistiques et de leur transformation actuelle[2]. Si les statisticiens classiques ont longtemps veillé à ce que leurs outils de mesure ne contiennent aucune trace politique (Porter, 1996), les *data scientists*, développeurs de machines prédictives, sont désormais contraints, par les problèmes d'équité qu'ils doivent traiter, de faire converger deux postures souvent distinctes : la recherche exigeante de la fiabilité des procédures de calcul et le souci de transparence du caractère construit et politiquement situé des opérations de quantification[3]. Le FairML est un objet original du point de vue de la sociologie historique de la quantification (Desrosières, 2013 ; Martin, 2020) en ceci qu'il implique, pour les *data scientists*, une posture *de facto* constructiviste : les chercheurs de ce domaine s'intéressent non seulement à ce que la qualité de la prédiction *dit* du rapport de la machine à la réalité, mais aussi, d'un même mouvement, à ce que la machine *fait* au réel en générant des décisions plus ou moins justes. Comment se manifeste concrètement ce constructivisme *de facto* sur la prise en compte de l'équité dans le *machine learning* ? Et quels sont ses effets sur le plan épistémologique ? C'est ce questionnement qui a guidé l'écriture de cet article.

Après avoir localisé socialement la formation du domaine FairML et montré le cadre épistémologique particulier dans lequel il s'inscrit, nous verrons dans un second temps comment, concrètement, les chercheurs parviennent à penser à la fois construction mathématique et construction sociale des approches, à l'issue de controverses sur les métriques d'équité et leur

2. Notons néanmoins, comme nous l'avons souligné dans l'introduction de ce numéro, que le débat n'est pas nouveau. Il est au cœur d'une controverse épistémologique en économie où s'opposent les approches positivistes et normatives. On trouve une analyse fouillée de ce problème épistémologique dans un numéro spécial de la *Revue Philosophique de Louvain* (Larue and Mueller, 2018).
3. Cette distinction est bien analysée par Alain Desrosières, notamment dans Chiapello and Desrosières (2006).

statut dans l'apprentissage machine. Nous montrerons enfin que les approches du FairML tendent vers une forme d'objectivité spécifique, celle du « jugement exercé » (Daston and Galison, 2010), reposant sur une perspective partielle et une justification *raisonnablement* partiale du concepteur de la machine – cette dernière devenant dès lors elle-même située politiquement.

1. Quand l'informatique crée de la philosophie morale : genèse du domaine du FairML

Invitant les philosophes à apprendre de ce que les sciences disent du réel, Bachelard considérait que les *sciences font la philosophie*, et invitait en retour les scientifiques à dialoguer avec les philosophes (Bachelard, 1934). Mais que se passe-t-il si la situation s'inverse, lorsque les objets de recherche relèvent de l'expertise philosophique (comme la morale, l'éthique et le politique dans notre cas) ? Appelle-t-on, de la même manière, les philosophes à rentrer dans un dialogue fécond avec les scientifiques (ici les informaticiens) qui développent des connaissances mathématiques sur des objets proprement philosophiques ? Il est facile d'admettre que dans ce cas de figure l'appel de Bachelard ne va plus de soi – les philosophes ayant plutôt tendance à considérer que personne d'autres qu'eux ne peut traiter de ces objets et, si ce n'est eux, qu'il faut se méfier et alerter des dangers de toute tentative de réductionnisme scientiste. Or, dans le domaine du FairML, la méfiance des philosophes semble avoir des effets vertueux sur les scientifiques qui s'attèlent aux questions d'équité. Elle contribue à façonner une proximité politique plus grande des scientifiques avec leur objet. Nous allons d'abord essayer de comprendre quelles ont été les conditions sociales de cette proximité au politique, puis nous montrerons comment celle-ci implique de penser les « biais » comme des normes de justice sociale.

1.1 Formation du domaine du FairML

Le FairML n'est pas apparu au hasard de l'espace scientifique. Ce sont les chercheuses qui s'intéressent aux problèmes de protection des données (*privacy*) qui vont mettre à l'agenda scientifique les problèmes d'équité. À la croisée des problèmes techniques et politiques, la structuration du domaine de la *fairness* s'apparente à celui de la *privacy* où la philosophie, les sciences sociales et les sciences computationnelles cherchent à s'unifier autour d'un projet commun. Cette structuration correspond à un dialogue constant entre d'une part, les travaux en philosophie morale des techniques (*value sensitive design*) tournés vers la compréhension de la dimension politique des systèmes informatiques (Friedman and Hendry, 2019 ; Nissenbaum, 2001) et d'autre part, la recherche en informatique qui cherche à traduire des concepts abstraits comme la vie privée et l'équité en langage mathématique (Kearns and Roth, 2019).

On comprend mieux cette structuration du domaine en présentant les trois chercheuses souvent présentées comme les pionnières de la recherche sur la *fairness* (et ce n'est sans doute pas un hasard s'il s'agit de trois femmes) : Helen Nissenbaumn, Cynthia Dwork et Latanya Sweeney.

Helen Nissenbaum est philosophe (on lui doit des concepts clefs comme celui de *contextual privacy*), mais elle a aussi contribué à concrétiser la notion de *value in design* en développant des logiciels libres de protection de la vie privée[4]. Elle appelle depuis longtemps à la prise en compte de l'équité dans la conception des outils numériques (Introna and Nissenbaum, 2000). Elle est aussi considérée comme la première chercheuse (avec Batya Friedman, une autre femme, spécialiste notoire de l'intégration de contraintes morales et politiques dans la conception des systèmes informatiques) à avoir posé la question des « biais » en informatique

4. Notamment *TrackMeNot* (pour la protection contre le profilage basé sur la recherche Web) et *AdNauseam* (protection contre le profilage basé sur les clics publicitaires).

comme un problème de « valeur »[5].

Cynthia Dwork est informaticienne (elle est célèbre pour ses contributions mathématiques et statistiques en cryptographie, protection des données et équité), notamment pour la notion de *differential privacy* (Dwork *et al*., 2006), tout en militant également pour une sensibilité politique plus grande dans la conception des techniques (Dwork and Mulligan, 2013). Par sa notoriété en *computer science*, Cynthia Dwork a joué un rôle important dans la constitution du FairML comme spécialité de recherche.

Enfin, Latanya Swenney est une informaticienne qui en plus d'avoir développé d'importants algorithmes d'anonymisation[6], a été une des premières à dénoncer les discriminations dans la publicité en ligne et à alerter sur les liens entre technologie informatique et racisme structurel (Sweeney, 2013).

Ces trois chercheuses montrent bien comment peuvent cohabiter et s'entremêler trois vertus épistémiques différentes : l'ouverture d'esprit vers la philosophie, la rigueur du raisonnement mathématique et l'engagement politique. C'est dans cet esprit que s'est formé le réseau de chercheurs en FairML et sa conférence scientifique annuelle – l'ACM Fairness, Accountability and Transparency (FAccT) Conference[7] – autour de laquelle se constitue un réseau interdisciplinaire. Cette interdisciplinarité n'est pas rhétorique. La philosophie morale et politique n'y apparaît pas seulement comme un cadre à l'intérieur duquel il est possible d'étudier des manières de mesurer la discrimination et de la prévenir. Les chercheurs en informatique s'impliquent aussi dans les débats sur les « métriques », adoptant ainsi une position politique. C'est en quelque sorte un laboratoire du co-constructivisme. Au sein de l'ACM FAccT, tout le monde s'accorde à reconnaître que les machines ne sont pas moralement neutres, qu'il est possible d'identifier en elles des tendances à promouvoir ou à rétrograder des valeurs et des normes morales particulières[8]. Et ceci a un effet direct sur le travail des *computer scientists* qui cherchent à comprendre ce que les algorithmes font à la société, ce que l'équité fait à l'algorithme en retour, et vice-versa. Ce qui semble tout à fait nouveau dans le domaine du FairML, c'est cette cohabitation improbable entre le réalisme métrologique des statistiques et le constructivisme des sciences sociales.

Mais cette cohabitation n'a rien d'un long fleuve tranquille. Sur le plan épistémologique, elle s'exprime comme une sorte de « rationalisme appliqué » dans le sens de Bachelard (1949), où réalisme et idéalisme, empirisme et conventionnalisme, positivisme et formalisme sont pris dans une tension permanente et fragile. De cette tension, particulièrement observable au sein des conférences ACM FAccT, il résulte une double exigence : le recours à l'argument mathématique et l'axiomatisation d'une part, et, d'autre part, une exigence dite de réflexivité (explicitement revendiquée dans le sens bourdieusien par les informaticiens eux-mêmes ; Laufer *et al*., 2022b) qui conduit les chercheurs à examiner sans cesse les dimensions politiquement et socialement situées des axes de recherche abordés, des manières de formuler les problèmes et des algorithmes eux-mêmes. Il est rare, dans l'histoire des sciences, d'observer en un même lieu, en même temps et par les mêmes protagonistes, la mise en œuvre de cette double exigence qui conduit à la revendication de savoirs situés à la manière de Haraway (1988).

5. Dans leur article « Bias in Computer System » (Friedman and Nissenbaum, 1996), elles décrivent trois types de biais dans les systèmes logiciels : les « préjugés préexistants » qui viennent, de manière implicite ou explicite, des personnes jouant un rôle important dans la conception du système, soit le client, soit le concepteur du système ; les « préjugés techniques » qui proviennent, selon elles, « de la quantification du qualitatif, de la discrétisation du continu et de la formalisation de l'informel », autant de réductions qui biaisent inéluctablement les décisions algorithmiques ; enfin, un troisième préjugé, appelé « préjugé émergent », n'apparaît qu'une fois la conception terminée, lorsque le système interagit avec un monde évolutif, donc susceptible de poser d'autres problèmes de biais indépendants de la conception initiale du système.
6. Cf. sa page de présentation : https://dataprivacylab.org/people/sweeney/
7. Avant de devenir une conférence de l'ACM en 2018, les conférences ACM FAccT s'appelaient depuis 2014 FAT /ML comme *Fairness, Accountability and Transparency in Machine learning* : https://www.fatml.org/
8. Cette idée que la technologie incarne des valeurs est directement inspirée des *Sciences and Technology Studies* (STS), qui étudient le développement de la science et de la technologie et leur interaction avec la société. Dans la littérature sur le FairML, on trouve de nombreuses mentions aux STS, même dans les articles qui s'inscrivent en *computer science*.

1.2 Le biais comme norme de justice sociale

L'un des effets de cette cohabitation est d'avoir fait de la question des « biais » un problème plus politique que méthodologique. Traditionnellement, dans une perspective réaliste, le biais est défini comme une erreur systématique (de mesure, de raisonnement, de procédure ou de jugement) qui produit une déviation par rapport à la « vérité ». En *machine learning*, les chercheurs appellent la réalité qui existe en dehors de leurs modèles la « vérité terrain » ou *ground truth*, et le biais est souvent défini comme un écart par rapport à cette vérité (Jaton, 2021). Ainsi, si la maîtrise des biais est essentiellement méthodologique, elle est une tâche scientifique primordiale en *machine learning* car le contrôle des biais est alors une manière de supprimer toute trace de subjectivité afin d'apporter une objectivité « mécanique » aux énoncés scientifiques (Daston and Galison, 2010).

Mais dès lors que la vérité terrain est naturellement biaisée car la société est structurellement injuste et inégale, que les technologies elles-mêmes sont parties intégrantes des structures sociales inégalitaires (en structurant les données) et que cette vérité peut être perçue de manière multiple selon les objectifs et les intérêts de chacun, la question des biais ne peut plus se poser de manière réaliste. Les biais deviennent des normes sociales qu'il faut s'efforcer de contrôler soit pour changer la société, soit pour choisir de ne rien faire. Les acteurs du FairML ont donné au concept de biais un statut et un sens nouveaux, du moins pour les ingénieurs – les sciences sociales ont de longue date intégré cette analyse des liens étroits entre technique de quantification et construction sociale de la réalité (Desrosières, 2002).

Dans un contexte où pour la plupart des systèmes d'IA connexionnistes la collecte de données ne repose pas sur un protocole spécifique, les *data scientists* endossent plus facilement une posture constructiviste. Par exemple, les développeurs de l'entreprise Predpol spécialisée dans la prédiction du crime pour guider les patrouilles de police, reconnaissent que les enjeux éthiques majeurs de la police prédictive sont de localiser les biais dans les données d'entrainement au niveau des interactions sociales qui produisent le signalement des crimes entre la police, le public et les criminels. Plus encore, ils admettent que ces opérations de codage statistique ont, par le biais des systèmes algorithmiques, des effets en retour sur la « réalité » par des boucles de rétroaction (*feedback loop*) : les résultats des prédictions biaisées alimentent à leur tour les données d'apprentissage qui viennent renforcer et augmenter la distribution inégale des arrestations ou de l'offre de sécurité dans la population (Brantingham, 2017).

On trouve une formulation de cette construction algorithmique de la réalité propre à l'usage du *machine learning* dans Mehrabi *et al*. (2019) qui représente les biais dans un processus cyclique en trois étapes : de la génération de données à l'algorithme (par exemple les bais historiques ou structurels qui renvoient aux rapports de pouvoir asymétriques du monde social), de l'algorithme à l'interaction utilisateur (par exemple l'omission de variables qui tiennent souvent aux préjugés et intérêts des développeurs), puis de l'interaction utilisateur aux données (par exemple les biais comportementaux des usagers « non souhaités » par les concepteurs)[9].

Les nombreuses accusations adressées aux systèmes algorithmiques révèlent une « sociologie des biais » qui montre que les biais ne seront jamais éliminés par une rigueur méthodologique accrue. Car le fond de la critique des décisions algorithmiques n'a rien à voir avec la rigueur avec laquelle les données sont collectées, mais avec le *point de vue* (dans le sens d'une « épistémologie du point de vue » ; Flores Espínola, 2012) adopté par le système sur le monde et ses effets rétroactifs. Il est tout bonnement impossible de produire un système de décision neutre et le biais est un problème de philosophie morale pour les utilisateurs des systèmes vis-à-vis de

9. Cité par Jean-Marie John-Mathews dans sa thèse de doctorat en science de gestion soutenue à l'université Paris-Saclay : « L'Éthique de l'Intelligence Artificielle en Pratique. Enjeux et Limites ».

ceux qui sont calculés (nous y reviendrons plus bas). L'objectif principal est d'éviter de prendre parti inconsciemment, tout en plaçant le problème du biais dans un enjeu de compréhension – en termes de philosophie morale et de sociologie des inégalités – de l'interaction de la machine avec le monde. D'où l'idée, partagée par la plupart des chercheurs du FairML, selon laquelle les algorithmes pourraient créer le potentiel pour de nouvelles formes de transparence et donc des opportunités de détecter les discriminations qui ne sont pas disponibles autrement (Kleinberg, Ludwig *et al*., 2016). En effet, pour prévenir la discrimination, nous devons disposer de moyens de la détecter, ce qui peut s'avérer extrêmement difficile lorsque des êtres humains prennent les décisions. Si les algorithmes peuvent accroître le risque de discrimination, ils ont le potentiel de faciliter la détection – et donc la prévention – de la discrimination. Avec le mouvement FairML, les algorithmes sont devenus des acteurs politiques de premier ordre (Abebe *et al*., 2020).

2. De la mesure des biais à leur interprétation

Avec cette manière de considérer les biais, le point de vue constructiviste se substitue désormais à celui réaliste observé de longue date dans les pratiques de quantification (Desrosières, 2014). Bien que « politique », cette posture n'inhibe pas la recherche statistique sur la mesure de l'équité et l'atténuation des biais dans la production des modèles prédictifs. Comment s'opère concrètement cette mise en politique des algorithmes ? Les recherches sur l'équité dans le *machine learning* sont généralement classées en trois grandes approches : l'équité de groupe, l'équité individuelle et l'équité par la causalité (Castelnovo *et al*., 2022). Si la plupart des analyses systématiques de la littérature les opposent les unes aux autres, il faut plutôt les observer dans une dynamique de controverses (Latour, 2014) où, à chaque approche, c'est une logique plus interprétative qui tente de s'imposer, au prix d'une prise en compte de l'équité de moins en moins automatique. En suivant l'évolution du débat scientifique depuis une dizaine d'années, on peut dessiner une ligne de front de la recherche où l'automatisation de la morale se trouve mise en tension avec un sens toujours plus politique de la quantification.

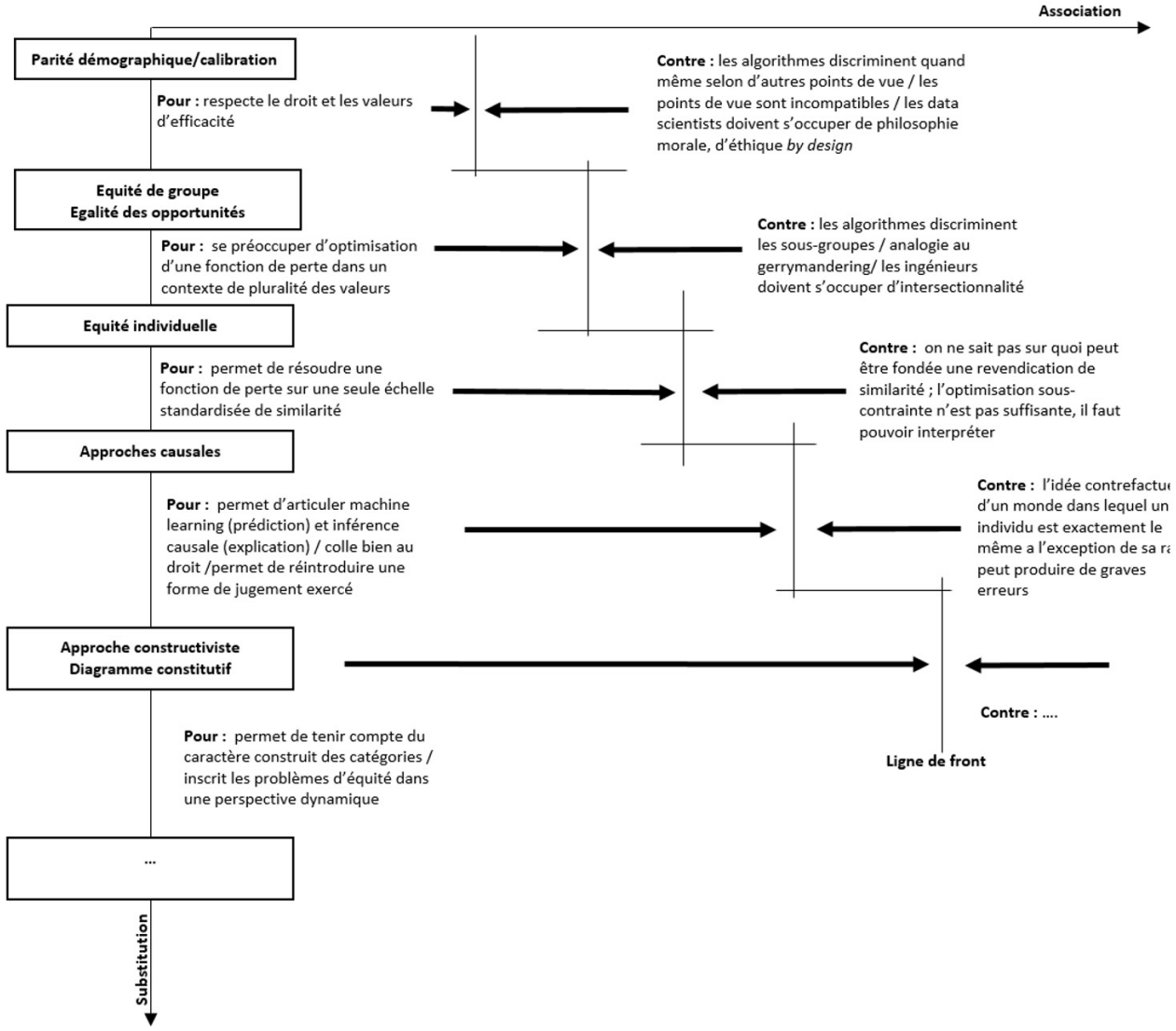


Figure 1 – *Diagramme dogmatique (non fidèle à la temporalité d'apparition des métriques) de la dynamique des controverses, inspiré de ceux de B. Latour (2014). Ce diagramme illustre les approches successives qui tendent à tourner le dos au réalisme métrologique.*

2.1 L'économie morale des métriques d'équité de groupe

C'est la controverse autour du logiciel COMPAS, utilisé par certains tribunaux aux États-Unis pour la prédiction de la récidive, qui a placé l'équité de groupe au premier rang des objets de recherche traités dans le FairML. Les désaccords qui ont opposé les data journalistes de ProPublica aux ingénieurs de Northpointe ont permis d'observer des manières différentes d'évaluer l'équité de l'algorithme : ProPublica a découvert que l'outil d'évaluation des risques COMPAS était biaisé à l'encontre des minorités ethniques en présentant des taux de faux positifs et de faux négatifs déséquilibrés. Si la situation est intuitivement injuste, le débat qui a suivi s'est principalement concentré sur le contraste entre la mesure utilisée par ProPublica et celle que lui a opposée Northpointe qui proposait plutôt d'égaliser la *précision* du modèle entre les groupes (approche appelée calibration). Comment trancher entre une métrique qui renvoie à une valeur d'égalité des chances et une autre à des valeurs (d'ingénieur) d'efficacité du modèle ? Il est impossible pour un modèle de satisfaire les deux métriques en même temps, et les chercheurs ont même formalisé cette impossibilité sous la forme d'un théorème (Kleinberg, Mullainathan *et al.*, 2016).

Cette difficile compatibilité des métriques a été conceptualisée à partir de trois grandes familles d'équité de groupe : la séparation (égalisation du « rappel »), la suffisance (égalisation de la « précision ») et l'indépendance (appelée souvent « parité démographique »). Alors que les deux premières renvoient à des manières différentes d'égaliser les types d'erreurs entre les groupes, la troisième impose une égalité des résultats de la classification algorithmique. L'*égalité des opportunités* (de la famille dite séparation) a été considérée, parmi les métriques de groupe, comme le meilleur moyen de produire de l'équité par Maurice Hardt, l'un des chercheurs les plus influents du domaine (Hardt *et al*., 2016). Hardt avance des arguments mathématiques en faveur de l'égalité des opportunités, en montrant notamment que la métrique d'égalité des opportunités est intéressante car elle maximise mieux l'utilité du modèle que la *parité statistique*. Mais on peut aussi comprendre sa position pour au moins trois raisons politiques.

La première raison est que la calibration ne peut pas être considérée comme une intervention de justice sociale en tant que telle, mais la résultante « normale » du travail bien fait du *data scientist* qui s'assure de la précision du modèle, quels que soient les groupes. En calibrant seulement le modèle, on ne relève pas le défi d'introduire de la morale dans la machine (bien que la précision relève d'une économie morale particulière comme le montre Loraine Daston ; Daston, 1995). La deuxième raison est que l'égalité des opportunités est *task specific*, impliquant ainsi le modélisateur et l'utilisateur dans la compréhension des modèles et leur raffinement. Enfin, la troisième raison est que la parité statistique (donc l'indépendance) peut être considérée comme le contraire même de l'équité. En compensant l'effet (indésirable) de la dépendance des classifications à la variable sensible, elle impose de traiter différents groupes de manière différente. Elle implique une forme de discrimination positive qui ne repose sur aucun principe méritocratique.

En proposant la métrique d'égalité des opportunités, Hardt n'avance pas seulement un argument statistique. Sa posture est aussi politique : il s'agit de prendre position dans l'opposition classique en philosophie morale entre l'égalité des résultats (la parité statistique) et l'égalité des chances (*equality of opportunity*). Notons en passant que la famille des métriques de la parité statistique reste néanmoins considérée par les juristes comme la seule métrique cohérente avec le droit antidiscriminatoire (voir l'encadré plus bas « Les conventions imposées par le droit »).

2.2 Des machines « intersectionnelles » : des sous-groupes à l'individu

Mais quelles que soient les métriques utilisées, l'équité de groupe est formulée à un niveau agrégé, et il a été montré que cette agrégation peut produire une forme de « *fairness gerrymandering* » (Kearns and Roth, 2018) – la notion de *gerrymandering* faisant référence au redécoupage électoral de certaines circonscriptions aux États-Unis en faveur de certains partis politiques. La métaphore du *gerrymandering* est une manière de signifier comment l'optimisation des prédicteurs, pour produire l'équité entre les groupes (par exemple, en imposant l'indépendance des résultats en fonction des variables de genre), se fait au prix de la discrimination de sous-groupes (par exemple les femmes d'une certaine classe d'âge). Autrement dit, à l'intersection de groupes protégés qui se chevauchent, des sous-groupes peuvent être discriminés. Sur le plan algorithmique, c'est une sorte de régression infinie, dans laquelle, comme le disent simplement Kearns et Roth, « *despite avoiding discrimination by race, gender, age, income, disability, and sexual orientation in isolation, we find ourselves with a model that, for example, unfairly treats disabled gay Hispanic women over age fifty-five making less than $50,000 annually* ».

En suivant cette logique intersectionnelle toujours plus granulaire dans la constitution des groupes, on arrive au niveau de l'individu. Dans cette optique, Dwork propose pour la première fois en 2012 une métrique d'équité individuelle pour prévenir les phénomènes de *gerrymandering* (Dwork *et al*., 2012). L'équité individuelle ne se calcule plus en fonction de l'appartenance à une catégorie, mais en mesurant une distance interindividuelle. Cette métrique intéresse les

chercheurs car elle est agnostique quant au sens qu'elle donne à la similarité, ce qui lui permet d'être spécifique à chaque tâche. Comme le soulignent Dwork et ses collaborateurs dans leur article séminal, les métriques d'équité individuelle reposent toujours sur un choix politique et contextuel :

> « *Our approach is centered around the notion of a task-specific similarity metric describing the extent to which pairs of individuals should be regarded as similar for the classification task at hand.1 The similarity metric expresses ground truth. When ground truth is unavailable, the metric may reflect the "best" available approximation as agreed upon by society. Following established tradition [Raw01], the metric is assumed to be public and open to discussion and continual refinement. Indeed, we envision that, typically, the distance metric would be externally imposed, for example, by a regulatory body, or externally proposed, by a civil rights organization.* » (Dwork, Ibid.)

Mais si les chercheurs justifient cette approche par son caractère procédural, il reste difficile de définir les critères de similarité entre les individus. L'équité individuelle nécessite des hypothèses fortes sur les relations entre le choix des *features* et les classes à prédire, ce qui n'est pas une tâche triviale pour le modélisateur. Les chercheurs se demandent encore actuellement si les notions individuelles d'équité peuvent être rendues pratiques. Plus encore, comme le souligne Jean-Marie John Mathews dans sa thèse, l'équité individuelle s'inscrit dans une politique qui est propre au *machine learning* : « on passe d'une classification nominale selon l'appartenance à une catégorie, à une classification ordinale interindividuelle [...] On laisse à l'algorithme le soin de fabriquer cette distance interindividuelle dans les espaces latents des réseaux de neurones. L'équité semble donc bien pouvoir être atteinte, mais comment la vérifier lorsqu'il n'est pas possible d'interpréter nominalement l'espace dans laquelle elle est vérifiée ? Quelles sont ces nouvelles catégories vis-à-vis desquelles l'algorithme nous garantit d'être agnostique ? Pour des raisons d'intelligibilité, l'*ordinalité* pure semble avoir des limites et il faudrait revenir à une forme de raisonnement nominal (Fourcade, 2016) » (Mathews, 2022).

Les conventions imposées par le droit

Les juristes voient le problème de l'équité algorithmique d'un point de vue différent. Selon eux, l'encadrement juridique des algorithmes de classification vise la recherche d'une pure égalité arithmétique des résultats de prédiction entre les individus, ou du moins à permettre la dénonciation du caractère plus ou moins excessif des inégalités de situation dans les prédictions afin de limiter de trop grandes disparités entre les groupes. Cet objectif trouve une traduction juridique aux États-Unis et en Europe, respectivement par les notions de « *disparate impact* » et « discriminations indirectes » qui toutes deux impliquent un usage de la statistique comme instrument de preuve.

Une comparaison des différences entre ces deux notions et leur implication dans la manière d'envisager la *fairness* dans le *machine learning* dépassent le cadre de cet article (pour une analyse approfondie, consulter Kirat *et al*., 2022). Prenons seulement pour exemple le cas européen à partir du débat suscité par les travaux de l'*Oxford Internet Institute* (Wachter *et al*., 2021). Ces analyses juridiques montrent que les métriques de parité statistique (critiquées comme nous l'avons vu plus haut par les chercheurs en *machine learning*) correspondent davantage à la conception juridique de l'équité car elles façonnent des décisions algorithmiques en faveur de mesure de compensation et de redressement. Elles sont porteuses d'un potentiel important de révision critique des conduites et des conventions sociales qui se logent dans les données d'apprentissage. Autrement dit, il s'agit de contraindre les algorithmes

afin qu'ils rapprochent les groupes vulnérables des groupes privilégiés. Cette famille de métrique est envisageable dans le droit car il s'agit d'intégrer un système algorithmique d'aide à la décision dans un projet de changement de l'état du monde. Du point de vue du droit, selon Wachter et ses collaborateurs, seules les métriques relevant d'une conception correctrice de la justice sociale sont mobilisables. Plus précisément, à partir d'une analyse jurisprudentielle de la CJUE, ils montrent que la législation européenne adopte une « égalité contextuelle ». Dans ce contexte, la mesure d'équité technique qui représente la traduction juridique la plus proche du « *gold standard* » de la Cour de justice européenne pour évaluer la discrimination est la « disparité démographique conditionnelle » (CDD). Cette métrique correspond à la métrique parité statistique, mais elle ajoute une contrainte conditionnelle qui s'exprime par une ou plusieurs variables. C'est le caractère conditionnel de cette métrique qui la rend plus adaptable à l'« égalité contextuelle » qui s'exprime dans la jurisprudence.

Dans le débat sur la *fairness* dans le *machine learning*, les juristes de l'*Oxford Internet Institute* ne proposent pas seulement de privilégier une métrique par rapport à une autre. C'est la manière de poser le problème de la *fairness* qui est modifiée. Les systèmes, selon eux, ne peuvent pas et ne doivent pas être conçus pour détecter, évaluer et corriger automatiquement les décisions discriminatoires, indépendamment des orientations et de l'interprétation locales du pouvoir judiciaire. Ce qu'il faut, c'est plutôt un « système d'alerte précoce » pour la discrimination automatique. Pour ce faire, il faut concevoir des systèmes capables de produire automatiquement ou systématiquement les types de preuves statistiques nécessaires pour que le pouvoir judiciaire puisse prendre des décisions normatives en toute connaissance de cause, et pour que les contrôleurs du système détectent systématiquement les discriminations potentielles avant qu'elles ne se produisent. En d'autres termes, ce qu'il faut, ce sont des normes techniques cohérentes qui s'alignent sur les procédures de référence du pouvoir judiciaire pour évaluer les discriminations algorithmiques. La confrontation entre les métriques d'égalité d'opportunité et celle de parité statistique renvoie respectivement à deux démarches différentes, respectivement ce qui relève de la lutte algorithmique pour la justice sociale, toujours discutable, et la détermination algorithmique de ce qui est juste selon le droit jurisprudentiel, en évolution permanente. La situation est d'autant plus trouble que les systèmes d'IA doivent réaliser en même temps ces deux projets contradictoires.

2.3 Des diagrammes causaux aux diagrammes constitutifs : vers un mode opératoire conventionnaliste ?

Pour pallier les problèmes que présentent les métriques d'équité individuelle, une équipe de jeunes chercheurs de l'*Alan Turing Institute* (Kusner *et al*., 2018) propose alors une nouvelle métrique dite d'équité contrefactuelle – une approche causale qui, selon les auteurs, offrent « *a natural way to define a similar individual* ». Cette métrique repose, selon leur définition, sur l'intuition qu'une décision est juste envers un individu si elle est la même dans le monde réel et un monde contrefactuel où l'individu appartient à un groupe démographique différent. S'inspirant de la théorie causale de Judea Pearl et ses collègues[10], l'article de Kusner et ses collaborateurs marque l'entrée d'un nouveau domaine d'études où s'est développée une multitude de notions de l'équité reposant sur les causes – l'équité contrefactuelle n'étant qu'une notion parmi

10. L'entrée du FairML dans les approches causales renvoie à un mouvement plus large en *machine learning* qui cherche à prendre en compte les aspects causaux dans les modèles pour construire des algorithmes interprétables, et le *causal fairness* est une classe de problèmes parmi d'autres. Elle est présentée comme un nouveau paradigme se substituant aux approches observationnelles. On passe d'une recherche sur les bons critères d'équité à celle sur le bon processus causal de génération de données du modèle prédictif.

d'autres. En effet, pour chacune des notions débattues au sein des approches contrefactuelles (*individual equalized contrefactual odds, path specific causal fairness, equal effort fairness*, etc.), qui sont autant d'alternatives aux métriques vues plus haut, on retrouve les différents systèmes de valeurs qui entourent les débats sur l'équité. Mais l'approche causale est aussi une valeur en soi : elle est motivée par l'idée que les questions relatives à la justice et à la discrimination sont de nature causale, et qu'il faut s'intéresser aux raisons causales des modèles d'injustice pour élaborer des algorithmes qui les corrigent . Sa politique est de s'opposer au positivisme du *machine learning* qui ne se soucie pas d'explication causale.

De manière générale, dans cette approche, il s'agit d'utiliser la technique des *graphes orientés acycliques* afin de montrer comment des comparaisons contrefactuelles de traitements des personnes appartenant à des groupes sensibles peuvent être intégrées comme des contraintes dans l'apprentissage. C'est au *data scientist* de définir les variables pertinentes en fixant une théorie de la discrimination. Cette théorie lui permet alors de tracer des flèches pour représenter les relations causales entre chaque variable qui forment des chemins vers la variable à prédire (Tremblay, 2022). Il existe plusieurs manières d'aborder les problèmes de causalité et des choix méthodologiques variés sont accessibles sur étagère (interventionniste ou contrefactuelle, par exemple). L'un des principaux problèmes de l'analyste est celui de savoir si la causalité peut être mesurée de manière unique à partir des données d'observation. Or, différents types d'effets causaux sont envisageables : l'effet total sur les interventions, les effets spécifiques au chemin qui permettent de rendre compte des discriminations directes ou indirectes, les effets contrefactuels, etc. Leur identifiabilité est, pour le dire simplement, dépendante des données d'observation accessibles et du niveau de connaissance des mécanismes de la discrimination (cf. Makhlouf *et al*., 2021, sur les critères d'identifiabilité pour décider des mesures d'équité basées sur la causalité).

Cette approche causale est radicalement opposée à celle du *machine learning*. Les variables explicatives de la discrimination et les liens entre ces variables ne sont pas donnés – ils se construisent avec les connaissances spécifiques à chaque problème. C'est le retour de la posture surplombante des experts sur le monde, à partir de leur propre modèle du monde, celui qu'ils jugent eux-mêmes pertinent, ce fameux monde que le *machine learning* prétendait faire émerger du monde lui-même par une quantité de données toujours plus massives. D'une manière plus forte que dans les approches précédentes, l'approche causale rend le développeur de l'algorithme encore plus dépendant des spécialistes d'autres disciplines telles que le droit, l'économie et les sciences humaines et sociales. Elle impose d'intégrer le contexte des données utilisées pour former les algorithmes. Elle donne ainsi une plus grande autonomie aux utilisateurs, évaluateurs et sujets de la décision vis-à-vis de l'algorithme ; c'est du moins en ces termes qu'elle est justifiée par ses promoteurs.

Malgré cet effort supplémentaire pour donner une place toujours plus grande à l'interprétation, la dynamique de la controverse ne s'arrête pas là. L'utilisation de l'inférence causale fait l'objet d'une critique qui exige de s'interroger sur le sens des catégories sensibles utilisées. Dans les analyses contrefactuelles, des catégories sociales sont manipulées comme on manipule un « traitement » dans un modèle causal, par exemple remplacer une molécule par un placébo pour évaluer l'efficacité d'un médicament. Manipuler des catégories sociales comme le sexe et la race dans une analyse contrefactuelle revient à les considérer comme des « choses en soi », donc à envisager les catégories sociales dans une perspective réaliste (Tiercelin, 2011). L'idée d'un monde dans lequel un individu est exactement le même à l'exception de sa race est-elle plausible ? Peut-on dissocier la race d'un individu des autres variables sociales qui le constituent ? Dans le monde social, soutiennent certaines analyses critiques de l'analyse contrefactuelle, les

11. Cette forme de correction des injustices intéresse particulièrement les juristes qui considèrent l'analyse causale comme l'instrument idéal pour la mise en œuvre du principe de « droit à l'explication » formulé dans le RGPD.

choses ont une signification causale « non pas en raison de ce qu'elles sont en elles-mêmes, mais en raison de leur relation avec d'autres choses ». Certains chercheurs appellent à mettre en place une vision plus constructiviste des catégories sociales dans les modèles d'analyse causale afin de tenir compte de leur « consistance » sociale, comme le proposent Lily Hu et Issa Kohler-Hausmann :

> « *The question then becomes: Given how a category is constituted, what algorithmic procedures do we consider fair? Constitutive diagrams of categories like sex and race would proffer explanations of how the meanings of those categories emerge from their constitutive structure; in other words, how the arrangement of complex social relations constitute a given group as what-it-is. Whereas causal diagrams facilitate inquiry into modular counterfactuals and ask how causal effects can be decomposed along different pathways, constitutive diagrams would highlight another counterfactual question: How might the social meaning of a group change if its constitutive elements are altered? That is, after all, the very promise of the antidiscrimination project: "[T]o transform the social meaning of social categories that have—for so long, in so many domains—been infused with disfavor and disadvantage.* » (Hu and Kohler-Hausmann, 2020)

Les « diagrammes constitutifs » sont encore trop jeunes pour qu'on puisse en montrer la portée dans cet article. Mais notons que c'est la solution trouvée pour faire tenir ensemble le caractère conventionnel de la base analytique des mécanismes discriminatoires avec des exigences d'optimisation propre à la pratique de l'apprentissage statistique. Il s'agit d'un pas supplémentaire vers des approches toujours plus interprétatives, qui placent les discussions explicites sur les conventions au premier plan de l'analyse quantitative. La valeur qui sous-tend cette approche est qu'un algorithme est juste, dans un contexte social donné, s'il tient compte de la manière dont les catégories elles-mêmes opèrent dans le monde. Mais si le FairML est en passe d'introduire dans la pratique de quantification un mode opératoire « conventionnaliste » (Boltanski and Thévenot, 1991), du chemin reste encore à parcourir. L'approche de Hu et Kohler-Hausmann s'inscrit néanmoins dans une ontologie réaliste du social, une ontologie des relations (Nef and Berlioz, 2021). Les catégories sont conventionnelles, mais les relations qu'elles tissent préexistent au diagramme constitutif et ont une existence bien réelle. D'où une posture de recherche sur la *fairness* plus proche de l'économie du bien être que de l'économie des conventions. Des propositions plus conventionnalistes existent comme celle de John-Mathews *et al*. (2022), mais elles sont encore peu visibles. Elles montrent cependant que la controverse peut encore évoluer pour faire exister le social dans les systèmes d'IA sous une forme toujours plus plurielle qui implique d'assumer qu'aucune statistique ne peut jamais s'extraire d'un point de vue *biaisé* sur le monde.

3. En guise de conclusion

Puisque la controverse n'est pas close, nous ne sommes pas en mesure de conclure cet article. Notons néanmoins que l'on voit poindre une forme d'objectivité spécifique dans le domaine du FairML. La controverse décrite ci-dessus montre que l'enjeu est de construire des approches ni sur l'héritage « réaliste » que l'on reproche souvent au *machine learning*, ni sur la recherche d'un équilibre optimal entre des conceptions de l'équité totalement antagonistes, mais bien sur la base d'une perspective partielle et d'une justification *partiale* des machines prédictives. Il y a deux visées éthiques centrales dans le FairML : d'une part, contre une conception univoque de la performance, un pluralisme axiologique et, d'autre part, contre l'objectivité mécanique, la primauté au jugement exercé par l'expert. Alors que depuis le XIX^{e} siècle, la « valeur de neutralité », bien plus que celle de « vérité » comme l'a montré Ted Porter, est un puissant moteur du développement des pratiques statistiques, la valeur de « partialité raisonnable » sous-tend, en quelque sorte, le développement de l'apprentissage automatique au XXI^{e} siècle. Pour être éthiquement acceptable, un algorithme ne peut être que raisonnablement partial. Et

pour devenir « scientifique », paradoxalement, le FairML tourne progressivement le dos à toute conception « réaliste » du machine learning, c'est-à-dire à toute prétention d'optimisation des enjeux d'équité en dehors d'un savoir situé (Haraway, 1988).

Références

Abebe R., Barocas S., Kleinberg J. *et al*. (2020), « Roles for computing in social change », *in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 27 January 2020), FAT* '20, Association for Computing Machinery, pp. 252-260.

Bachelard G. (1934), *Le Nouvel Esprit Scientifique*, Paris, Presses Universitaires de France.

Bachelard G. (1949), *Le rationalisme appliqué*, Paris, Presses Universitaires de France.

Boltanski L. et Thévenot L. (1991), *De la justification : les économies de la grandeur*, Paris, Gallimard.

Brantingham P. J. (2017), « The Logic of Data Bias and its Impact on Place-Based Predictive Policing », *Ohio State Journal of Criminal Law*, 15, p. 473.

Castelnovo A., Crupi R., Greco G., Regoli D., Penco I., and Cosentini A. (2022), « A clarification of the nuances in the fairness metrics landscape », *Scientific Reports*, 12.

Chiapello È. et Desrosières A. (2006), « La quantification de l'économie et la recherche en sciences sociales : paradoxes, contradictions et omissions. Le cas exemplaire de la "Positive accounting theory" », *in* Eymard-Duvernay F. (éd.), *L'économie des conventions, méthodes et résultats. Tome 1. Débats*, Paris, La Découverte, coll. « Recherches », pp. 297-310.

Crawford K. (2021), *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale University Press.

Daston L. (1995), « The Moral Economy of Science », *Osiris*, 10, 2nd Series, pp. 2-24.

Daston L. and Galison P. (2010), *Objectivity*, Zone Books.

Desrosières A. (2002), *The Politics of Large Numbers: A History of Statistical Reasoning*, Harvard University Press.

Desrosières A. (2013), *Pour une sociologie historique de la quantification : L'Argument statistique I*, Presses des Mines via OpenEdition.

Desrosières A. (2014), *Prouver et gouverner : Une analyse politique des statistiques publiques*, Paris, La Découverte.

Dwork C. and Mulligan D. K. (2013), « It's Not Privacy, and It's Not Fair », *Stanford Law Review Online*, 66, p. 35.

Dwork C., McSherry F., Nissim K. *et al*. (2006), « Calibrating Noise to Sensitivity in Private Data Analysis », in Halevi S. and Rabin T. (eds.), *Theory of Cryptography*, Berlin, Heidelberg, Springer, Lecture Notes in Computer Science, pp. 265-284.

Dwork C., Hardt M., Pitassi T. *et al*. (2012), « Fairness through awareness », *in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (New York, NY, USA, 8 January 2012), ITCS '12, Association for Computing Machinery, pp. 214-226.

Eubanks V. (2018), *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Publishing Group.

Flores Espínola A. (2012), « Subjectivité et connaissance : réflexions sur les épistémologies du "point de vue », *Cahiers du Genre*, 53(2), pp. 99-120.

Fourcade M. (2016), « Ordinalization: Lewis A. Coser memorial award for theoretical agenda setting 2014 », *Sociological Theory*, 34(3), pp. 175–195.

Friedman B. and Hendry D. G. (2019), *Value Sensitive Design: Shaping Technology with Moral Imagination*, MIT Press.

Friedman B. and Nissenbaum H. (1996), « Bias in computer systems », *ACM Transactions on Information Systems*, 14(3), pp. 330-347.

Haraway D. (1988), « Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective », *Feminist Studies*, 14(3), pp. 575-599.

Hardt M., Price E., and Srebo N. (2016), « Equality of Opportunity in Supervised Learning », in Lee D., Sugiyama M., Luxburg U., Guyon I., and Garnett R. (eds.), *Advances in Neural Information Processing Systems*,
https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Hu L. and Kohler-Hausmann I. (2020), « What's Sex Got To Do With Fair Machine Learning? », *in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 27 January 2020), FAT* '20, Association for Computing Machinery, pp. 513.

Introna L. and Nissenbaum H. (2000), « Defining the Web: the politics of search engines », *Computer*, 33(1), pp. 54-62.

Jaton F. (2021), « Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application », *Big Data & Society*, 8(1),
https://doi.org/10.1177/20539517211013569.

John-Mathews J.-M., Cardon D., and Balagué C. (2022), « From Reality to World. A Critical Perspective on AI Fairness », *Journal of Business Ethics*, 178(4), pp. 945-959.

Kearns M. and Roth A. (2019), *The Ethical Algorithm: The Science of Socially Aware Algorithm Desig*n, Oxford University Press.

Kearns M., Neel S., Roth A., and Wu, Z. S. (2018), « Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness », *arXiv*, https://arxiv.org/abs/1711.05144.

Kirat T., Tambou O., Do V., and Tsoukiàs A. (2022), « Fairness and Explainability in Automatic Decision-Making Systems. A challenge for computer science and law », *arXiv*,
https://arxiv.org/abs/2206.03226.

Kleinberg J., Ludwig J., and Mullainathan S. (2016), « A Guide to Solving Social Problems with Machine Learning »,
https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning.

Kleinberg J., Mullainathan S., and Raghavan M. (2016), « Inherent Trade-Offs in the Fair Determination of Risk Scores », *arXiv*, http://arxiv.org/abs/1609.05807.

Kleinberg J., Ludwig J., Mullainathan S, and Sunstein C. R. (2018), « Discrimination in the Age of Algorithms », *Journal of Legal Analysis*, 10, pp. 113-174, https://doi.org/10.1093/jla/laz001.

Kusner M. J., Russell C., Loftus J. R., and Silva R. (2018), « Causal Interventions for Fairness », *arXiv*, http://arxiv.org/abs/1806.02380.

Larue L. et Mueller T. M. (2018), « La Normativité en Science Economique. Une perspective pratique, historique et philosophique », *Revue Philosophique de Louvain*, 116, p. 147.

Latour B. (2014), *Cogitamus : six lettres sur les humanités scientifiques*, Paris, La Découverte.

Laufer B., Jain S., Cooper A. F., Kleinberg J., and Heidari H. (2022), « Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings, and Future Prospects », *in 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 20 June 2022), FAccT '22, Association for Computing Machinery, pp. 401-426.

Makhlouf K., Zhioua S., and Palamidessi C. (2021), « Survey on Causal-based Machine Learning Fairness Notions », *arXiv*, http://arxiv.org/abs/2010.09553.

Martin O. (2020), *L'empire Des Chiffres: Sociologie de La Quantification*, Malakoff.

Mehrabi N., Morstatter F., Saxena N., Lerman K., and Galstyan A. (2019), « A Survey on Bias and Fairness in Machine Learning », *arXiv*, http://arxiv.org/abs/1908.09635.

Nef F. et Berlioz S. (2021), *La nature du social : de quoi le social est-il fait ?*, Le Bord de l'eau.

Nissenbaum H. (2001), « How computer systems embody values », *Computer*, 34(3), pp. 120–119.

Noble S. U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press.

O'Neil C. (2016), *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, Crown.

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

Porter T. M. (1996), *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press.

Sweeney L. (2013), « Discrimination in Online Ad Delivery », *arXiv*, https://arxiv.org/abs/1301.6822.

Tiercelin C. (2011), *Le ciment des choses* (1re édition), Paris, Editions Ithaque.

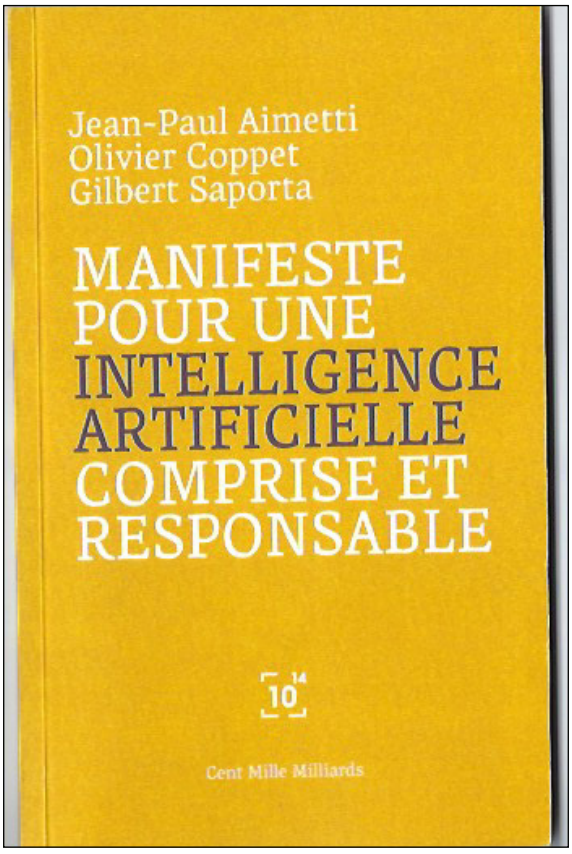
Tremblay V. (2022), « Équité algorithmique : perspective interdisciplinaire et recommandations pour statisticiens et autres scientifiques de données », https://hal.science/hal-03663226.

Wachter S., Mittelstadt B., and Russell C. (2021), « Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI », *Computer Law & Security Review*, 41, p. 105567, https://doi.org/10.1016/j.clsr.2021.105567.

Manifeste pour une intelligence artificielle comprise et responsable

de Jean-Paul AIMETTI, Olivier COPPET et Gilbert SAPORTA(2022)

Jean-Jacques DROESBEKE[1]
Université libre de Bruxelles



Livre (90 pages)
Édition : Cent Mille Milliards – 2022
ISBN : 978-2-85071-207-4

1. Jean-Jacques.Droesbeke@ulb.be

Voilà un beau titre pour ce petit essai de 90 pages ! À l'heure où l'intelligence artificielle s'invite quotidiennement dans nos conversations, nos médias et parfois nos travaux, le souci des auteurs de cet opuscule ne peut que nous interpeller.

La première moitié de l'ouvrage propose un certain nombre d'aphorismes. On y trouve quatre parties bien distinctes. La première s'interroge sur la définition de l'Intelligence artificielle (IA). Les auteurs en ont choisi une sans surprise : « Algorithmes et systèmes conduisant à des raisonnements et des actions automatisés, visant à remplacer et améliorer des activités généralement attribuées à l'intelligence et au comportement humains » (p. 16). Si vous ne connaissez pas la signification du mot « algorithme », une petite étoile vous renvoie à quelques mots d'explication dans un chapitre réservé à ce type d'éclaircissement. La voie choisie par les auteurs a fait ses preuves : raconter brièvement l'origine de l'automatisation et suivre quelques étapes importantes du développement de cette IA. Nous avons apprécié cette envie de ne pas recourir sans arrêt à des termes anglais quand des expressions adéquates existent en français ! Par ailleurs, le biomimétisme a ses limites et les auteurs ne se privent pas de nous le rappeler.

La deuxième partie est intitulée « Petit guide pratique pour ne pas dire n'importe quoi… ». Il concerne d'abord le concept de « données», matière première dont se nourrit l'IA. Les algorithmes sont ensuite sujets à des réflexions diverses pour souligner leur fragilité et la nécessité de ne pas regarder les résultats de leur application comme une vérité absolue. L'usage de l'IA en entreprise retient ensuite l'attention des auteurs qui soulignent ses limites. Si aujourd'hui, c'est encore « la faute à l'informatique» quand quelque chose ne tourne pas rond, demain, ce sera « la faute à l'IA ». L'humain est de plus en plus indispensable et doit le rester.

La troisième partie nous rappelle que « l'IA est la meilleure et la pire des choses ». Tout le monde sait que « Data science sans conscience n'est que ruine de l'humanité », mais il est bon de le rappeler fréquemment, ce que font ici les auteurs.

Enfin, se pose la question de l'avenir de l'IA. L'imagination permet des scénarios de toute espèce dont les conséquences ne sont pas nécessairement profitables au développement harmonieux de nos sociétés. Les auteurs ne s'aventurent pas à proposer des solutions mais soulignent tout l'enjeu de conserver la maîtrise la plus grande possible du recours à une robotisation dont on mesure mal les effets. Les questions qu'ils posent en soulignent toute l'importance.

Le reste de ce petit ouvrage s'adresse « aux plus curieux » en tentant de définir des termes courants comme « algorithme » — nous en avons déjà parlé — et l'IA elle-même, mais aussi des expressions courantes comme « data science », « boite noire », « cookies » et « GAFAM ou BATX ». Deux exemples simples illustrent le propos, l'un dans le domaine médical, l'autre dans l'apprentissage d'un agent conversationnel, encore appelé « chatbot » ou « dialogueur » auquel nous avons toutes et tous été confronté(e)s dans nos recherches de contacts téléphoniques. Quelques pages revenant sur l'histoire de l'IA, sur les biais possibles de l'IA et les précautions dont il faut s'entourer quand on y recourt achèvent ce petit opuscule, sans oublier quelques références bibliographiques et la présentation des auteurs.

Ces derniers sont trois mathématiciens qui savent de quoi ils parlent. Le premier, Jean-Paul Aimetti, est Professeur émérite du Conservatoire national des arts et métiers, président de l'Académie des sciences commerciales et de l'ISC, grande école de commerce. Son profil est relativement proche de celui de Gilbert Saporta, Professeur émérite du même établissement et Président d'honneur de la Société française de Statistique. Le troisième auteur, Olivier Coppet, est aussi statisticien. Il s'est forgé une expérience importante dans des projets d'enquêtes, de panels et de bases de données, complétée ensuite par son intérêt pour la qualité des données et l'éthique de leurs usages.

Ce petit ouvrage se lit facilement et avec plaisir. Il nous rappelle avec insistance l'obligation pour chacune et chacun d'entre nous de mettre l'humain au centre des débats et des actions de notre époque. Le choix de la citation de Piaget qui ouvre l'ouvrage n'est pas anodin : « L'intelligence, ce n'est pas ce que l'on sait, mais ce que l'on fait quand on ne sait pas ».

Hommage à André Vanoli : praticien, réformateur et historien de la comptabilité nationale

Quentin DUFOUR[1]
CMH, ENS-EHESS, UMR CNRS 8097

André Vanoli est né le 22 octobre 1930 à Portel (Pas-de-Calais), et mort le 20 février 2022 à Paris. Comptable national dès la fin des années 50, historien de sa propre discipline, il a été un acteur de premier plan dans le développement de la comptabilité nationale tant en France qu'au niveau des institutions internationales. Il a également produit des travaux théoriques et historiques dont l'influence s'étend au-delà des praticiens de la comptabilité nationale, notamment en économie, en histoire et en sociologie.

Issu d'une famille modeste, il rentre à l'Institut d'Études Politiques de Paris en 1948 dans le but d'intégrer l'École Nationale d'Administration (ENA). C'est durant cette période qu'il s'engage au sein du Parti Communiste Français (PCF), en apportant notamment une expertise sur les sujets économiques. L'année 1956 constitue un tournant dans sa trajectoire : de ses propres mots, il est à la fois écarté du PCF du fait d'un désaccord avec la ligne officielle du parti, et empêché pour des raisons politiques de présenter l'ENA (Vanoli, 2002).

Après sa rencontre avec Claude Gruson, André Vanoli intègre en 1957 le Service des Études Économiques et Financières (SEEF), entité du ministère des Finances alors chargée de développer une comptabilité nationale et de fournir des prévisions économiques dans le cadre du Plan. À l'époque, le SEEF est un organe relativement nouveau et en pleine ébullition (Terray, 2003). D'une part, il participe à la construction d'un système d'information économique dont les concepts ne sont pas encore stabilisés au niveau international. Il produit un système original pensé en lien avec la comptabilité d'entreprise (Touchelay, 2016). Dans ce cadre, Vanoli est affecté aux premiers comptes des biens et services au titre de l'agriculture et de la chimie. D'autre part, la comptabilité nationale est étroitement connectée à un projet politique : sous l'influence de la pensée keynésienne, marquée par les années de guerre, l'équipe dirigée par Claude Gruson entend « programmer l'espérance » de la France et fonder sa puissance grâce à la comptabilité nationale (Fourquet, 1980). En 1962, l'organisation du SEEF est bouleversée. Une partie, qui comprend Vanoli, intègre l'Insee pour se focaliser sur la confection des comptes passés. Elle rejoint ainsi Claude Gruson devenu directeur général de l'institut un an plus tôt. Le reste des équipes conserve son affectation au ministère des Finances et devient la Direction de

1. quentin.dufour@ens.psl.eu

la Prévision.

De 1962 jusqu'à sa retraite en 1995, André Vanoli reste à l'Insee, et continue à travailler sur la comptabilité nationale. Il fait partie des rares personnes à avoir constitué une expérience sur le long terme des problématiques et des enjeux de la discipline. Il est possible de souligner trois éléments marquants de sa carrière professionnelle. Premièrement, il a participé à la coordination statistique au niveau national. Depuis les années 60 et jusqu'à la fin des années 80, la comptabilité nationale joue un rôle clé dans l'organisation de l'Insee : elle est présente dans les différentes divisions de production et dispose d'une division de synthèse. Conjuguées à la centralité de la comptabilité nationale à l'Insee, les prises de responsabilité de Vanoli l'amènent ainsi à occuper des postes décisifs de l'organisation en étroite relation avec le directeur général. Il prend notamment la tête de la direction de la coordination statistique et comptable. Il est par ailleurs un des fondateurs du Cnis en 1984, dont il sera secrétaire général. Deuxièmement, il a été une des chevilles ouvrières du renouvellement des concepts de la comptabilité nationale au niveau mondial. D'abord dans les années 60, au sein d'un groupe de travail européen réuni à Bruxelles. Il y remet un rapport en 1964 sur le fonctionnement des comptes nationaux à l'Insee qui inspirera pour partie le Système Européen des Comptes de 1970. Ensuite, dans les années 80, décennie durant laquelle il a été l'un des six experts internationaux à piloter la réflexion sur le Système de Comptabilité Nationale de 1993, jalon historique de la standardisation du cadre comptable. Sa position centrale à l'Insee durant les années 80 lui a permis de mobiliser l'institution dans son ensemble afin d'alimenter des travaux collectifs sur le cadre comptable. Troisièmement, il s'est largement investi dans l'animation et la diffusion de la discipline, que ce soit au sein d'associations – il a présidé l'Association de Comptabilité Nationale (ACN), ainsi que l'International Association for Research on Income and Wealth – ou comme expert pour accompagner le développement de systèmes de comptabilité nationale dans différents pays (Colombie, Brésil, Tunisie, Grèce…).

Tout au long de sa carrière, son activité de praticien a été systématiquement accompagnée et alimentée par des réflexions théoriques. Sur plusieurs décennies, il a consulté une quantité impressionnante de travaux académiques et experts sur le sujet (notamment en histoire et en économie), mais il a également initié et participé à des discussions collectives, lors du renouvellement du cadre comptable ou à l'occasion de colloques associatifs. La pièce maitresse de son travail théorique est certainement son ouvrage de 2002, *Une histoire de la comptabilité nationale* (Vanoli, 2002), dont la diffusion s'est étendue avec sa traduction en 2005. Pensé à l'origine comme un texte court, il s'agit finalement d'une somme de plus de 600 pages sur l'histoire de la discipline en France et à l'international, mais également d'une réflexion sur les concepts comptables et les problèmes qu'ils drainent (production, valeur, volume, etc.), dont les questions autour de la controverse sur la croissance et la notion de bien-être. Bien que versant dans une histoire internaliste, ce texte constitue un ouvrage de référence pour les historiens et sociologues de l'économie. Le travail théorique d'André Vanoli ne se limite évidemment pas à son livre de 2002. Ses réflexions ont fait l'objet de publications et d'actes de colloques antérieurs et postérieurs, sur différents sujets. Citons à titre d'exemple la notion de production (Vanoli, 1983), la relation entre comptabilité nationale et comptabilité d'entreprise (Vanoli, 2010), ou encore le rapport aux nouveaux indicateurs de richesse (Vanoli, 2013). Dès les années 90, il a commencé à s'intéresser aux problèmes environnementaux et à leur articulation aux comptes nationaux (Vanoli, 1995). Toujours au fait des évènements qui traversent le monde économique et social, c'est cette thématique qu'il a portée jusqu'à récemment. En 2019, lors du colloque de l'ACN dont il était président d'honneur, il proposait un état des lieux des réflexions sur le sujet, et mettait en avant sa proposition originale autour de la notion de « coûts écologiques non payés » sur laquelle il travaillait depuis plusieurs années.

Références

Fourquet F. (1980), *Les comptes de la puissance. Histoire de la comptabilité nationale et du Plan*, Paris, Éditions Recherches.

Terray A. (2003), *Des francs-tireurs aux experts : l'organisation de la prévision économique au ministère des Finances. 1948-1968*, Paris, Comité pour l'Histoire Économique et Financière.

Touchelay B. (2016), « Private accounting, statistics and national accounting in France: a unique relationship (1920-1960s) », *in* Isabelle B., Jany-Catrice F., and Touchelay B. (eds.), *The social sciences of quantification. From politics of large numbers to target-driven policies*, Springer, pp. 141-148.

Vanoli A. (1983), « Les tracés divers de la notion de production », *Economie et statistique*, 158, pp. 61-73.

Vanoli A. (1995), « Reflection on Environmental Accounting Issues », *Review of Income and wealth*, 41(2), pp. 113-137.

Vanoli A. (2002), *Une histoire de la comptabilité nationale*, Paris, La Découverte.

Vanoli A. (2010), « Is National Accounting Accounting? National Accounting between Accounting, Statistics and Economics », *Comptabilité(s)* [online], 1, http://journals.openedition.org/comptabilites/226.

Vanoli A. (2013), « Chapitre 12. Comptabilité nationale, statistiques et indicateurs de développement durable : état de l'art et des réflexion », *in* Vivien F.-D. (éd.), *L'évaluation de la durabilité*, Versailles, Editions Quæ, pp. 239-265.