Équité, explicabilité, paradoxes et biais

Gilbert Saporta

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.