REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Vol 10 N°1 2023 *https://doi.org/10.20416/LSRSPS.V10I1.7*

Maël Pégny

État de l'art : l'équité algorithmique

SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES (SPS)
École normale supérieure
45, rue d'Ulm
75005 Paris
www.sps-philoscience.org

REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Maël Pégny

État de l'art : L'équité algorithmique

Sommaire

Cet article présente un état de l'art critique du sous-champ de l'éthique de l'Intelligence Artificielle nommé "équité algorithmique". Il se concentre en particulier sur un objet récent, mais faisant objet d'une activité intense, à savoir les critères statistiques de l'équité algorithmique. Il s'agit de critères permettant de déterminer si un algorithme ou modèle prédictif de ML ne présente pas de biais défavorable à l'égard d'une population donnée en examinant les caractéristiques statistiques de son comportement entrées-sortie. Nous commençons par présenter les diverses définitions au centre des conversations récentes, ainsi qu'un résultat crucial d'impossibilité d'optimisation simultanée de toutes ces métriques, nommé "résultat d'incompatibilité." Nous discutons ensuite les leçons philosophiques à tirer de cette pluralité de définitions et du résultat d'incompatibilité. Nous finissons par une discussion critique portant sur les difficultés provenant de la représentation par des variables mathématiques de réalités sociales complexes comme l'appartenance identitaire à une population donnée, ou la caractérisation d'un discours comme discours de haine.

Mots clés : Intelligence artificielle, équité, statistiques, machine learning, algorithme.
Keywords: Artificial intelligence, equity, statistics, machine learning, algorithm.

1. Introduction

L'association de la définition de la justice et des mathématiques a des racines extrêmement anciennes, puisqu'Aristote déjà associait deux de ses trois notions de justice, la justice distributive et la justice réparatrice, à des notions mathématiques, respectivement l'égalité géométrique (égalité entre proportions ou rationnels) et l'égalité arithmétique (égalité entre nombre entiers[1]). Cette association prend une nouvelle jeunesse avec les tentatives récentes, dans la recherche en éthique de l'IA, de définir des métriques statistiques de l'équité pour les algorithmes. La recherche très dynamique sur ces métriques vise à garantir qu'un algorithme ne produit pas d'effets discriminatoires à l'égard d'une population donnée. Outre l'exemple du risque de récidive que nous discuterons ci-dessous, les algorithmes peuvent être employés dans des décisions aussi variées et sensibles que la diffusion des offres d'emplois, la reconnaissance faciale, les recommandations, la prioritisation des soins médicaux, la prédiction des activités criminelles à des fins de ciblage de l'action policière (*predictive policing*)[2].

Ce travail de revue de la littérature sur les métriques de l'équité en IA vise à mieux comprendre le statut philosophique de ces « définitions mathématiques de l'équité » et la pertinence éthique et politique de cette association entre concepts éthiques et définitions mathématiques[3]. Nous commencerons par une succinte revue[4] des principaux résultats de la

1 *Aristote, Ethique à Nicomaque, livre V.*

2 *Voir respectivement Anja Lambrecht et Catherine Tucker, « Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads », Management science 65, no 7 (2019): 2966-81. Tobias Schnabel et al., « Recommendations as treatments: Debiasing learning and evaluation », in international conference on machine learning (PMLR, 2016), 1670-79. Jacqueline G. Cavazos et al., « Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? », IEEE transactions on biometrics, behavior, and identity science 3, no 1 (2020): 101-11. Heidi Ledford, « Millions of Black People Affected by Racial Bias in Health-Care Algorithms », Nature 574, no 7780 (octobre 2019): 608-9, https://doi.org/10.1038/d41586-019-03228-6. Julia Dressel et Hany Farid, « The Accuracy, Fairness, and Limits of Predicting Recidivism », Science Advances 4, no 1 (17 janvier 2018): eaao5580, https://doi.org/10.1126/sciadv.aao5580. Julia Dressel et Hany Farid, « The Dangers of Risk Prediction in the Criminal Justice System », MIT Case Studies in Social and Ethical Responsibilities of Computing, no Winter 2021 (5 février 2021), https://doi.org/10.21428/2c646de5.f5896f9f.*

3 *Pour une discussion de l'articulation entre les compréhensions mathématiques de l'équité et les traditions conceptuelles venues d'autres disciplines, voir Reuben Binns, « What can political philosophy teach us about algorithmic fairness? », IEEE Security & Privacy 16, no 3 (2018): 73-80. Reuben Binns, « Fairness in Machine Learning: Lessons from Political Philosophy », in Conference on Fairness, Accountability and Transparency, 2018, 149-59, http://proceedings.mlr.press/v81/binns18a/binns18a.pdf. Mateusz Dolata, Stefan Feuerriegel, et Gerhard Schwabe, « A Sociotechnical View of Algorithmic Fairness », Information Systems Journal n/a, no n/a, consulté le 19 mai 2022, https://doi.org/10.1111/isj.12370. Pour un examen des perceptions du public profane sur cette mathématisation de l'équité, voir Reuben Binns et al., « "It's Reducing a Human Being to a Percentage" Perceptions of Justice in Algorithmic Decisions », in Proceedings of the 2018 Chi conference on human factors in computing systems, 2018, 1-14.*

4 *Une revue complète de la littérature technique serait impossible : une simple recherche Google Scholar le 19 Mai 2022 pour les expressions « algorithmic fairness » et « fairness metrics » produit respectivement de 4 210 et 3310 résultats. Nous ne mentionnons la littérature technique que lorsqu'elle a une importance*

littérature, à savoir la définition d'une pluralité de concepts d'équité algorithmique (section 1) et la démonstration d'un résultat dit d'incompatibilité entre certaines de ces métriques (section 2), en nous concentrant sur les métriques de l'équité dite « de groupe » (*group fairness*). Nous nous interrogerons ensuite sur le sens philosophique de cette pluralité de métriques (section 3) et sur les difficultés éthiques posées par l'application de tels critères à des variables mal choisies ou mal conçues (section 4).

2. Une pluralité de concepts

La première distinction fondamentale de cette littérature est la distinction entre équité de groupe et équité individuelle (en anglais, *individual fairness*). Tandis que l'équité individuelle s'interroge sur l'égalité de traitement entre deux individus jugés similaires[5], l'équité de groupe s'interroge sur l'équité de traitement entre différentes populations, notamment entre une population qu'on ne juge pas victime de discriminations, que nous appellerons la population favorisée, et une ou des populations discriminées. Comme la discussion historique sur les discriminations porte avant tout sur les pratiques affectant des populations entières, nous allons nous concentrer, comme le fait la littérature, sur l'équité de groupe. L'un des problèmes décisifs est de déterminer ce que l'on compare entre les différentes populations.

2.1 La parité statistique

Pour définir une métrique de l'équité de groupe, il faut répondre à une question politique décisive : quelle est la quantité statistique que l'on se donne comme cible lorsqu'on cherche à obtenir une égalité de traitement entre populations[6] ? Un critère bien antérieur aux discussions actuelles sur l'apprentissage automatique ou *Machine Learning* (ML), offre un exemple simple de compréhension de cette quantité. Selon la parité statistique (*statistical parity*), aussi nommée « parité démographique » (*demographic parity*), cette quantité est la probabilité d'obtenir une décision donnée, qui doit être similaire[7] à travers les populations. En d'autres termes, si un homme a 30% de chances de se voir attribuer un prêt immobilier, alors une femme doit avoir environ 30% de chances d'avoir elle aussi une réponse favorable. On remarquera que ce critère porte sur la distribution des résultats de la prise de la décision, sans s'interroger sur la nature de la procédure donnant cette distribution en sortie : il s'agit là d'un trait important des métriques d'équité de groupe, qui sont dominées par une réflexion sur l'équité distributive plutôt que procédurale[8].

Cette doctrine de l'équité de groupe a l'avantage d'être simple et intuitive, et elle plonge ses racines dans la tradition juridique américaine de l' « incidence inégale » (*disparate impact*). Cette tradition de réflexion juridique sur l'égalité de traitement est à l'origine d'une règle connue en droit américain sous le nom de « règle des quatre cinquièmes » (*four fifths rule*), d'après laquelle une population discriminée, ou « protégée » dans la terminologie juridique américaine, devrait se voir accorder une proportion d'au moins quatre cinquièmes du taux des embauches du groupe le plus favorisé dans la population considérée[9]. Ce critère suppose donc un

historique, une grande valeur pédagogique ou lorsqu'elle contient des passages conceptuels plus accessibles à un lectorat sans compétences informatiques, en privilégiant toujours les références les plus récentes. Pour la littérature venue de la philosophie et des sciences sociales, notre filet est plus large sans non plus pouvoir prétendre à l'exhaustivité. Nous privilégions les références introductives et de portée large, avec toujours une faveur pour les références récentes, et bien sûr une faveur pour les références illustrant les commentaires critiques particuliers à ce travail.

5 Pour une discussion sur les relations entre équité de groupe et équité individuelle, voir Reuben Binns, « On the Apparent Conflict Between Individual and Group Fairness », in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT '20 (Barcelona, Spain: Association for Computing Machinery, 2020), https://doi.org/10.1145/3351095.3372864. Pour un article cherchant à dépasser l'opposition entre équité de groupe et équité individuelle, et examinant les effets de l'appartenance à plusieurs sous-groupes, voir Michael Kearns et al., « Preventing fairness gerrymandering: Auditing and learning for subgroup fairness », in International Conference on Machine Learning (PMLR, 2018), 2564-72.*

6 Il existe de multiples présentations des métriques de l'équité, dont certaines sont bien plus détaillées que ce que nous pouvons proposer ici. Sans prétendre à l'exhaustivité, on peut recommander pour un public plus familier du droit et de l'éthique Doaa Abu-Elyounes, « Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness », University of Illinois Journal of Law, Technology & Policy 2020, no 1 (2020): 1-54. Pour une présentation mathématique plus détaillée, voir Shira Mitchell et al., « Algorithmic fairness: Choices, assumptions, and definitions », Annual Review of Statistics and Its Application 8 (2021): 141-63. Pour une presentation mathématique condensée, voir Sahil Verma et Julia Rubin, « Fairness definitions explained », in 2018 IEEE/ACM International Workshop on Software Fairness (Fairware) (IEEE, 2018), 1-7. Pour une longue présentation mêlant mathématiques, considérations conceptuelles et études de cas concrètes, voir Ninareh Mehrabi et al., « A survey on bias and fairness in machine learning », ACM Computing Surveys (CSUR) 54, no 6 (2021): 1-35. Pour une somme sur le sujet, voir le livre en ligne Solon Barocas, Moritz Hardt, et Arvind Narayanan, Fairness and Machine Learning. Limitations and opportunities. (fairmlbook.org, 2019), https://fairmlbook.org/.

7 En termes plus mathématiques, on dira que la différence entre la probabilité d'avoir un prêt lorsqu'on est un homme et cette même probabilité lorsqu'on est une femme devra être inférieure à une borne ε. Pour contourner la lourdeur de l'expression mathématique, comme pour éviter un effet indu d'intimidation intellectuelle, nous parlerons simplement de „similarité" pour désigner cette notion.

8 Pour un article faisant exception à la règle et s'intéressant à l'équité procédurale du ML, voir Nina Grgić-Hlača et al., « Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning », in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

9 La règle des quatre cinquièmes n'est pas une règle juridique établissant à elle seule l'existence de discriminations. Elle est un indicateur, défini par le Equal

respect approximatif de la parité statistique, ce qui montre que la réflexion sur les métriques statistiques de l'équité a déjà exercé une influence sur le droit positif.

On entend parfois que la parité statistique permet de garantir que la proportion de membres d'une population donnée obtenant un bien social sera similaire à sa proportion dans la population totale. Ainsi, si la parité statistique était appliquée, la moitié des postes de responsabilités serait occupée par des femmes. Mais il s'agit là d'un raccourci intellectuel. La parité statistique ne peut à elle seule rectifier les problèmes de sous-représentation des populations discriminées dans tous les secteurs de la société. Elle est avant tout un critère de décision qui ne peut rectifier les inégalités présentes dans les données de cette décision. S'il y a seulement 3 femmes qui candidatent pour 60 hommes à des postes d'ouvrier du bâtiment, il y aura au final 20 hommes engagés pour une seule femme, même si les deux populations ont une chance similaire d'embauche à 33%. L'entreprise du bâtiment ne peut corriger toutes les inégalités d'accès genrées à certaines professions qui se sont exercées bien en amont de sa décision d'embauche : elle peut seulement faire sa part en n'ajoutant pas d'injustice dans la décision qui lui revient. Pour que l'application de la parité statistique permette d'avoir une proportion de détenteurs d'un bien social similaire aux proportions dans la population globale, il faudra déjà que les proportions dans les données soient similaires à celles de la population globale. Cette vérité fondamentale du *unfairness in, unfairness out* va nous accompagner dans toute notre réflexion sur les métriques de l'équité. Elle montre d'emblée les limites d'une pensée de l'équité qui ne se concentre que sur le moment de la prise de décision, et ignore le contexte en amont.

En outre, confondre la parité statistique comme similarité de la probabilité d'une décision à travers les populations et similarité des proportions de populations en sortie avec les proportions dans la population globale peut avoir des effets pervers. Dans son ouvrage *Automating Inequality*[10], Virginia Eubanks raconte que les programmes d'aide aux chômeurs du *Civilian Conservation Corps* avaient une borne de 10% pour les Noirs au nom de cette interprétation faussée de la parité statistique, alors que ceux-ci pouvaient représenter 80% des chômeurs dans certaines villes du Nord. L'interprétation erronée de la parité statistique peut ainsi mener dans certains contextes à une véritable discrimination à l'égard d'une population donnée, dont les besoins spécifiques sont ignorés par ce critère. Dans d'autres cas de figure, il peut arriver qu'une population victime de discriminations avérées puisse être représentée dans la possession d'un bien social donné à une proportion supérieure à sa part dans la population globale : on dira alors qu'elle est « statistiquement surreprésentée » dans la possession de ce bien, sans qu'une telle terminologie n'implique un quelconque jugement de valeur politique. Cette surreprésentation statistique dans la possession d'un bien social ne signifie en aucun cas l'absence de discrimination : elle peut même constituer un effet secondaire de cette discrimination. Il n'étonnera personne que plus de la moitié des sages-femmes sont des femmes[11]. Les femmes sont donc statistiquement surreprésentées parmi les sages-femmes, ce qui est parfaitement cohérent avec leur enfermement dans la fonction reproductive et maternelle. Il n'en reste pas moins que la profession de sage-femme n'a rien de déshonorant, et que même dans un monde où règnerait l'égalité de genre la plus parfaite, nombre de femmes choisirait de l'exercer. Mais à l'heure actuelle elles peuvent occuper plus de la moitié des postes de sages-femmes sans que cela soit le signe d'une discrimination qu'il faille corriger.

Une telle métrique a aussi été critiquée parce qu'elle ignore la possibilité de traitement différentiel justifié entre populations. Dans le cas célèbre du logiciel de score de risque de récidive COMPAS, sur lequel nous reviendrons plus bas (section 1.2), hommes et femmes ont ainsi été traités séparément. L'argument était que si les femmes ont un plus faible taux de récidive que les hommes, pourquoi le pourcentage de libération conditionnelle ne serait-il pas nettement plus élevé pour elles ? En imposant une similarité de traitement entre une population particulière et le reste de la population, la parité statistique noie les particularités de la population considérée dans la masse, et lui ferme ainsi la porte à un traitement différencié qui pourrait être à la fois justifié et favorable à cette population discriminée. Là encore, de telles critiques s'inscrivent dans une profonde tradition juridique et politique, puisqu'elles invoquent des arguments typiques de la querelle entre partisans de la discrimination positive et partisans de l'égalité de traitement par un voile d'ignorance sur les caractéristiques sensibles des populations, deux positions en incompatibilité philosophique et méthodologique profonde. Il s'agit là aussi d'un débat politique qui accompagne la réflexion sur les métriques de l'équité[12].

Pour répondre à ces critiques, il serait possible d'envisager une forme souple de la parité statistique : au lieu d'imposer une probabilité d'accès à un bien social similaire entre toutes

Employment Opportunity Coordinating Council (EEOC) dans ses Uniform Guidelines on Employee Selection Procedures à destination des agences fédérales américaines, qui peut être employé comme indice (evidence) de la présence d'effets discriminants (disparate impact ou adverse impact). Irwin Greenberg, « An analysis of the EEOCC "Four-Fifths" rule », Management Science 25, no 8 (1979): 762-69.

10 *Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (Saint Martin's Press, 2018).*

11 *« Nursing and Midwifery », consulté le 15 mai 2022, https://www.who.int/news-room/fact-sheets/detail/nursing-and-midwifery.*

12 *Pour une introduction à ce débat, voir la section II de l'opus cité d'Abu Elyounes, qui présente la distinction entre l'approche selon laquelle l'équité doit être aveugle (the unaware approach) contre celle défendant que l'équité doit être fondée sur une prise en compte explicite des appartenances de groupe des individus (fairness through awareness).*

les populations, on pourrait envisager que certaines populations discriminées aient droit *a minima* à une probabilité similaire, et éventuellement à une probabilité supérieure dans certaines circonstances. On trouve là aussi une idée essentielle du débat sur les métriques de l'équité, à savoir qu'une conception donnée d'une telle métrique ne doit pas être conçue comme une définition mathématique de l'équité, mais comme un instrument statistique dont la pertinence devrait être évaluée en fonction du contexte d'usage. Dans le cas des critiques de la parité statistique au sens strict que nous venons d'examiner, on a vu qu'il fallait prendre en compte des informations telles que la surreprésentation statistique dans la possession d'un bien social donné –comme le cas des sages-femmes- la sous-représentation statistique dans la possession d'un bien social donné –comme dans le cas des Afro-Américains plus fréquemment sans emploi- et enfin, et surtout, si la population considérée doit être considérée victime de discriminations systématiques ou non. Elle dépend aussi d'autres éléments spécifiques de contexte propres aux populations considérées et à l'usage de l'algorithme examiné : si l'algorithme montre une publicité pour le Viagra à 90% des hommes visitant un site Web, l'équité n'impose pas forcément qu'il la montre à la même proportion de femmes. L'équité n'impose pas plus que l'on montre autant les publicités pour les matchs de football aux femmes qu'aux hommes : elle impose plutôt qu'on les montre autant aux femmes intéressées par le football qu'aux hommes intéressés par le football, deux populations bien plus difficiles à isoler[13]. L'équité ne peut donc être atteinte en appliquant de manière rigide un critère numérique à travers les populations délimitées par diverses informations sensibles : elle nécessite une prise en compte du contexte d'usage et une réflexion sur la population ciblée. Une telle conception contextuelle de l'emploi d'une métrique peut être défendue au nom d'un certain pluralisme. On qualifiera ici de « pluralisme » la conception défendant qu'il n'y a pas à choisir la bonne métrique ou la bonne conjonction de métriques parmi les différentes définitions proposées dans la littérature, mais qu'il faut plutôt comprendre dans quel contexte une métrique, ou une conjonction de métriques, peut être appliquée de manière pertinente pour promouvoir une décision équitable : l'objet du débat n'est pas de trouver la définition juste de la métrique de l'équité, mais de concevoir une série d'outils statistiques permettant d'évaluer les effets d'une décision algorithmique en fonction du contexte. Nous allons voir que cette question du pluralisme des métriques est essentielle aux débats en cours.

2.2 L'égalité de l'exactitude

Nous n'allons pas nous attarder plus avant sur les débats entourant la parité statistique, dans la mesure où celle-ci constitue moins le centre du débat actuel que d'autres métriques plus conçues plus spécifiquement pour les modèles du ML. Une des caractéristiques les plus fondamentales du débat sur les métriques d'équité en ML est qu'elles doivent porter sur des modèles statistiques prédictifs. Il ne s'agit plus seulement d'évaluer les effets d'une procédure ou d'une politique publique par des indicateurs statistiques *a posteriori,* mais de comparer la prédiction faite *a priori* avec les résultats *a posteriori*. Un tel tournant conceptuel permet de prendre en compte les probabilités d'erreur du modèle lui-même. La question des relations entre performances prédictives et équité mène à s'intéresser par conséquent à des quantités différentes de la parité statistique.

L'une des premières métriques d'intérêt est l'exactitude (*accuracy*[14]) de l'algorithme, soit le taux de prédictions couronnées de succès données par cet algorithme. Par exemple, un modèle prédisant le défaut de paiement aura une exactitude de 80% si 80% des résultats prédisent correctement soit des futurs défauts, soit des remboursements bien effectués. Pour réfléchir sur l'équité d'un modèle prédictif, on s'interroge d'abord sur ses chances de succès, et l'on examine ensuite comment ces chances de succès, si elles varient en fonction de l'appartenance à une population, ont des chances d'affecter l'accès au bien social. L'équité de l'exactitude (*accuracy equity*) affirme donc que l'exactitude du modèle devrait être la même pour toutes les populations : si un modèle a une exactitude de 75% pour une population donnée, il doit avoir la même exactitude pour toutes les autres populations. Heidari et al 2019[15] soulignent avec justesse que la discussion de la répartition des erreurs implique la prise en compte des limitations épistémiques du modèle. En d'autres termes, plutôt que de raisonner sur un modèle parfait, qui n'existe guère dans la pratique, mieux vaut inclure la présence des erreurs dans la réflexion sur l'équité.

13 *Nous reprenons cet exemple à Jon Kleinberg, Sendhil Mullainathan, et Manish Raghavan, « Inherent Trade-Offs in the Fair Determination of Risk Scores », in 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), éd. par Christos H. Papadimitriou, vol. 67, Leibniz International Proceedings in Informatics (LIPIcs) (Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017), 43:1-43:23, https://doi.org/10.4230/LIPIcs.ITCS.2017.43.*

14 *La traduction française de l'anglais accuracy dans un contexte statistique est une tâche redoutable. Dans le contexte de la théorie de la mesure, l'accuracy est traduit par "exactitude" (distance d'une valeur de mesure à la valeur réelle) et opposée à la "précision" (precision, dispersion des valeurs d'une même mesure répétée). Dans le contexte des modèles prédictifs binaires, l'accuracy peut être traduit par "précision" et opposée au "rappel" (recall mais aussi precision…), soit le taux de prédictions réussies parmi le taux de prédictions positives. Dans tous les cas, la confusion menace.*

15 *Hoda Heidari et al., « A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity », in Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19 (New York, NY, USA: ACM, 2019), 181-90, https://doi.org/10.1145/3287560.3287584.many existing definition of algorithmic fairness, such as predictive value parity and equality of odds, can be interpreted as special cases of EOP. In this respect, our work serves as a unifying moral framework for understanding existing notions of algorithmic fairness. Most importantly, this framework allows us to explicitly spell out the moral assumptions underlying each notion of fairness, and interpret recent fairness impossibility results in a new light. Last but not least and inspired by luck egalitarian models of EOP, we propose a new family of measures for algorithmic fairness. We illustrate our proposal empirically and show that employing a measure of algorithmic (un*

La controverse autour du logiciel COMPAS[16] a eu le mérite de montrer que l'équité de l'exactitude est insuffisante. Cette controverse a eu une immense importance pour la littérature sur les métriques de l'équité, et elle mérite donc quelques mots de présentation. COMPAS est un logiciel propriétaire d'IA utilisé dans le système judiciaire américain pour calculer un score de risque de récidive d'un détenu, qui est pris en compte par les juges dans des décisions comme le maintien en détention préventive ou la remise de peine[17]. Ce logiciel a été accusé par l'association ProPublica d'être injuste envers les détenus Afro-Américains, en commettant plus d'erreurs défavorables à leur encontre qu'envers les détenus blancs[18]. L'entreprise a répliqué que l'exactitude de l'algorithme était identique pour toutes les populations, et que les scores attribués par son logiciel correspondaient à des statistiques similaires de récidive pour toutes les populations[19]. Pour comprendre le fond de cette controverse, il faut saisir les différences entre ces métriques de l'équité algorithmique.

L'exactitude de l'algorithme était bien similaire entre les Afro-Américains et les Blancs, mais cette performance globale cachait une disparité éthiquement pertinente dans les taux d'erreurs. Les Afro-Américains subissaient beaucoup plus d'erreurs défavorables les classant comme individus à haut risque. La notion de décision favorable n'est pas elle-même une notion mathématique : elle dépend de la relation de désir que les individus ont envers la décision prise. Mais ceci entraîne qu'on ne peut mettre toutes les erreurs de l'algorithme sur le même plan. L'équité de l'exactitude est donc une métrique insuffisante au sens où elle ne permet pas de rendre compte de cette asymétrie morale entre les erreurs.

La prise en compte simultanée de l'inéluctabilité des erreurs, ainsi que de l'asymétrie morale entre erreurs favorables et erreurs défavorables, invite à proposer de nouvelles métriques. Deux autres métriques ont été conçues pour rendre compte de cette asymétrie entre erreur favorable et erreur défavorable : l'égalité des opportunités et l'égalité des chances. Pour simplifier la présentation, nous la limiterons au cas des décisions binaires, soit les décisions se bornant à déterminer possession d'une propriété par l'individu concerné, comme « présenter un haut risque de récidive » ou « être un client solvable » : les résultats mathématiques peuvent être généralisés à plus de deux classes. Avec cette simplification, l'égalité des opportunités (en anglais *equality of opportunity* ou *equalized opportunity*) affirme que les taux de faux négatifs devraient être identiques à travers toutes les populations considérées. Si cette métrique est satisfaite, tous les individus méritant un résultat devraient avoir une chance égale de se voir attribuer ce résultat, puisque le taux de refus injustifiés (faux négatifs) sont similaires à travers toutes les populations. L'égalité des chances (*equalized odds*) généralise cette approche aux personnes recevant un résultat négatif : les taux de faux négatifs et de faux positifs sont identiques à travers les populations. C'est la raison pour laquelle on peut considérer que la métrique d'égalité des chances est en réalité la conjonction de deux métriques. Comme le remarquent bien Heidari et al. 2019, cette métrique part du constat qu'un individu peut recevoir tout aussi bien un avantage qu'un désavantage par erreur, et que ces avantages et désavantages erronés doivent aussi être répartis équitablement. L'équité ne signifie donc pas seulement que chaque personne a une chance égale de recevoir ce qu'elle mérite, mais aussi qu'elle a une chance égale de recevoir ce qu'elle ne mérite pas.

En tout état de cause, l'application de l'égalité des chances entraîne la similarité des taux d'erreurs favorables et défavorables à travers les populations, et apporterait donc une solution au problème vu dans le cas COMPAS. En revanche, elle ne résout pas la question décisive du coût des erreurs, qui n'est pas forcément symétrique entre le favorable et le défavorable, et peut dépendre du contexte. Comme il est bien connu dans la réflexion sur les tests médicaux[20], les tests pour des maladies graves nécessitant une détection et

16 *Julia Angwin, Surya Mattu, Lauren Kirchner, Jeff Larson. There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, May 23, 2016.*

17 *Pour une description sociologique des pratiques des juges employant ces logiciels, voir Angèle Christin, Alex Rosenblat, et Danah Boyd, « Courts and predictive algorithms », Data & CivilRight, 2015. Angèle Christin, « Algorithms in Practice: Comparing Web Journalism and Criminal Justice », Big Data & Society 4, no 2 (décembre 2017): 1-14, https://doi.org/10.1177/2053951717718855.*

18 *Julia Angwin et al., « Machine Bias », ProPublica, 23 mai 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=siiaBuUx_5-LH2f_432kxejIHJI-dlxM.*

19 *La réponse formelle de la compagnie propriétaire du logiciel, Northpointe (renommée depuis lors Equivant) peut être téléchargée à la page suivante equivant, « Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity », equivant, 1 décembre 2018, https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/. Des auteurs venus du système judiciaire américain ont aussi pu prendre la défense de COMPAS et des autres outils actuariels d'évaluation du risque employé par ce même système : Alejandro Flores, Kristin Bechtel, et Christopher T. Lowenkamp, « False Positives, False Negatives, and False Analyses: a Rejoinder to " Machine Bias: There's Software Used across the Country to Predict Future Criminals. and It's Biased against Blacks. " », Federal Probation Journal 80, no 2 (2016): 38-46. Cara Thomson, « Myths and Facts. Using Risk and Need Assessments to Enhance Outcomes and Reduce Disparities in the Criminal Justice System » (National Institute of Corrections & Community Corrections Collaborative Network, mars 2017).*

20 *Il ne s'agit là que d'un élément de discussion dans un débat bien plus complexe. Dans une discussion des tests pour les maladies génétiques chez les nouveaux nés, R. Rodney Howell rappelle que si les faux négatifs sont « absolument inacceptables », les faux positifs ont aussi un coût, comme la saturation du système administratif et médical s'ils sont en très grand nombre, ou la baisse du seuil des vigilances des médecins, un risque grave quand un délai de traitement peut être fatal R. Rodney Howell, « The High Price of False Positives », Molecular Genetics and Metabolism, Special Issue Articles Authored by SIMD Presidents and their Colleagues, 87, no 3 (1 mars 2006): 180-83, https://doi.org/10.1016/j.ymgme.2005.10.004.*

un traitement rapide sont conçus avant tout pour éviter les faux négatifs. L'asymétrie entre le coût des erreurs est la raison de ce choix : il est dommageable de faire croire à tort à une personne qu'elle est atteinte d'une maladie grave, mais il est encore plus dommageable de ne pas détecter une maladie nécessitant une intervention rapide. La réflexion initiée par le cas COMPAS a donc mené à une prise en compte de l'asymétrie entre erreurs favorables et erreurs défavorables, mais la prise en compte des asymétries de coût entre les erreurs par une métrique d'équité serait bien plus complexe.

2.3 La calibration

La dernière des métriques parmi les plus discutées en IA est la calibration. La calibration d'un modèle générant une probabilité garantit que cette probabilité se traduise de la même manière en termes de statistiques réelles à travers les populations. En termes simples, si un modèle prédit qu'une propriété est présente dans 30% d'une population, il se peut que dans les statistiques réelles elle ne soit présente que dans 27% des cas, mais il faut que cet écart entre prédiction probabiliste et réalité statistique soit le même pour toutes les populations. La situation à éviter serait qu'une probabilité de 30% se traduise en 40% pour les statistiques d'une population et en 20% pour une autre, car de tels écarts fluctuants entre prédiction et réalité selon la population pourrait être une source d'erreurs dommageables[21]. Pour reprendre l'exemple des algorithmes calculant un risque de récidive, si les détenus blancs se voyant attribué un risque de 30% récidivent dans les faits dans 40% des cas, et que les Afro-Américains se voyant attribués la même probabilité ne récidivent que dans 20% des cas, il est évident que le modèle sous-estime le risque de récidive des Blancs et surestiment celui des Afro-Américains, ce qui mènera à un traitement défavorable de ces derniers. La calibration est particulièrement importante pour les nombreux modèles de ML attribuant une note de risque (*risk score*) qui traduit une certaine fourchette de probabilité : elle garantit alors que la note signifie la même chose à travers toutes les populations[22]. La calibration était la métrique satisfaite par le logiciel COMPAS. Elle était donc à la base de la défense de l'entreprise propriétaire Equivant contre les critiques de l'association ProPublica : cette défense affirmait qu'un score de risque bien calibré ne pouvait être accusé de biais raciste à l'égard des Afro-Américains. Comme on vient de le voir, cette réponse constitue un changement de sujet : la métrique employée par la défense d'Equivant diffère de celle employée implicitement par la critique de ProPublica, à savoir l'égalité des chances. Mais comme nous allons le voir à présent en présentant le théorème d'incompatibilité des métriques, ce changement est d'autant plus problématique que ces métriques non seulement ont une signification différente, mais ne peuvent être optimisées en même temps.

3. Les résultats d'incompatibilité : l'impossible équité ?

3.1 Une première présentation intuitive du résultat

L'un des résultats fondateurs de la littérature en métriques de l'équité est la démonstration qu'il est en règle générale impossible d'optimiser en même temps calibration et égalité des chances. Pour énoncer le résultat avec plus de rigueur, il faut introduire le concept de « taux de base » (*base rate*) pour une population : celui-ci désigne le taux réel dans une population de la variable prédite par le modèle. Si le modèle tâche de prédire le taux de contamination dans une population, le taux de base d'une population sera le taux de personnes contaminées dans cette population. Le théorème d'incompatibilité des métriques affirme que si le modèle prédictif est imparfait, et les taux de base diffèrent entre deux populations, alors il est impossible d'optimiser simultanément égalité des chances et calibration pour ces deux populations[23]. Il faut souligner que le résultat est valable pour tout modèle statistique prédictif, et n'est donc en aucun restreint à des techniques propres à l'IA contemporaine.

Les seules exceptions à ce résultat sont donc la possession d'un modèle parfait, soit un modèle ne produisant aucune prédiction erronée, ou l'égalité entre les taux de base des deux populations considérées. Dans le cas d'un modèle parfait, obtenir des taux d'erreurs similaires entre populations ne pose pas problème, puisqu'il n'y a pas d'erreur : malheureusement, les modèles parfaits n'existent guère dans la vie réelle. L'égalité des taux de base entre deux populations signifie qu'une prédiction donnée du modèle aurait nécessairement le même écart aux taux de base des deux populations. Si un tel cas de figure est bien entendu possible d'un point de vue théorique, il ne saurait être la règle. Lorsqu'on s'intéresse par exemple à la comparaison entre une population majoritaire et une minorité discriminée, le problème fondamental

21 *La calibration est aussi nommée "calibration par groupe'* (calibration per group) *pour indiquer qu'elle doit être respectée pour chaque groupe séparément.*

22 *D'autres auteurs comme Hedden Brian Hedden, « On statistical criteria of algorithmic fairness », Philosophy and Public Affairs 49, no 2 (2021). préfèrent abandonner la référence à la sémantique et parler de „valeur probante"* (evidential import) *de la note, mais ce type de nuances ne nous concernent pas au niveau intuitif et introductif de cette présentation.*

23 *Kleinberg, Mullainathan, et Raghavan, « Inherent Trade-Offs in the Fair Determination of Risk Scores ». Le résultat avait déjà été prouvé dans le cas binaire par Alexandra Chouldechova, « Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments », Big Data 5, no 2 (juin 2017): 153-63, https://doi.org/10.1089/big.2016.0047.*

est précisément que les taux de base des variables d'intérêt sont souvent différents puisque les populations sont sociologiquement différentes, et qu'en outre ces différences peuvent faire partie de l'héritage historique de la discrimination. Dans le cas des Afro-Américains, leur plus fort taux de récidive est ainsi le fruit de leur profil socio-économique et de l'attention disproportionnée que leur portent les forces de l'ordre, qui augmente considérablement leurs chances d'arrestation[24]. Et plus la discrimination est forte et systématique, plus elle affecte de variables et plus elle les affecte fortement. Un individu membre d'une population discriminée peut voir affecter ses chances d'être arrêté ou de finir en prison, d'avoir un logement ou un emploi, de faire des études, de se marier en dehors de sa communauté d'origine, d'atteindre une certaine espérance de vie, etc. La discrimination rend donc plus difficile l'application des métriques statistiques d'équité que l'on souhaiterait pourtant employer pour lutter contre elle. Dans l'immense majorité des cas pratiques, on se retrouvera donc face à la situation vue dans le cas COMPAS, où un arbitrage est nécessaire entre les plus importantes métriques de l'équité.

3.2 La preuve du théorème

Nous allons présenter la preuve du théorème d'incompatibilité de manière semi-formelle. Ce passage reprend avec quelques réarrangements modestes la présentation de Kleinberg *et al.*

Afin d'énoncer correctement le résultat et sa preuve, il nous faut au préalable fixer la terminologie et la notation. On considère un problème quelconque de classification appliqué à une population. On cherche donc à savoir, à l'aide d'un vecteur de données σ, si un individu satisfait la propriété *P* d'appartenir à une classe donnée, par exemple « les individus présentant un fort risque de défaut de paiement. » La classe des personnes satisfaisant la propriété est nommée par Kleinberg *et al.* la classe positive, et la classe complémentaire la classe négative. Ces appellations n'ont rien à voir avec le caractère désirable ou non de la propriété : par exemple, on peut appartenir à la classe positive en étant classé comme un individu à fort risque de récidive violente. Dans la pratique, on ne connait que rarement la propriété *P* avec certitude : on formule donc un modèle statistique qui attribue à chaque individu, sur la base des données σ, une certaine probabilité p_σ d'appartenir à la classe positive. La fraction des individus possédant effectivement la propriété *P* est nommée « taux de base » (*base rate*).

On fait l'hypothèse que p_σ est connue pour tout σ. Ceci n'est bien sûr pas toujours garantie dans la pratique, mais cela permet d'énoncer un argument *a fortiori* pour Kleinberg *et al.* : même dans le cas favorable où p_σ est parfaitement connue, on ne peut satisfaire les critères d'équité simultanément. Enfin, les individus peuvent appartenir à des groupes notés par un entier *t*. Pour simplifier l'exposition, Kleinberg et al. présentent la preuve dans le cas simple de deux groupes, mais la preuve peut être généralisée à un nombre arbitraire *n* de groupes.

Le calcul d'un score de risque correspond à la répartition des individus dans un intervalle (noté *b* pour l'anglais *bin*) en fonction de σ, et au calcul d'un score v_b pour tous les individus dans cet intervalle. L'appartenance à un groupe donné *t* n'influençant pas la probabilité p_σ, les critères d'équité se reformulent comme des conditions sur p_σ par rapport à l'appartenance à un groupe. La calibration signifie que pour tout groupe *t* et tout intervalle *b* avec son score associé, $v_{b,t,}$ la fraction des personnes du groupe *t* appartenant à la classe positive est égale à v_b. L'égalité des taux d'erreur favorable et défavorable, nommée équilibres pour la classe positive et la classe négative par Kleinberg *et al*, signifie que le score moyen des personnes d'un groupe *t* appartenant à la classe positive (resp. négative) doit être égale au score moyen des personnes de tout groupe *t'* appartenant à la classe positive (resp. négative).

La terminologie et la notation étant fixées, nous pouvons énoncer la preuve du théorème. Soient *x* le score moyen donné à un membre de la classe négative, et *y* le score moyen attribué à un membre de la classe positive. Appliquons à présent nos trois critères d'équité. Les conditions d'équilibre imposent que pour tout groupe *t*, $x = y$. La condition de calibration impose que le score $v_{b,t}$ des personnes du groupe *t* dans l'intervalle *b* soit égal au nombre des personnes dans *b* appartenant à la classe positive. Si on effectue une sommation sur tous les *b*, la somme des scores reçus par les personnes du groupe *t* est notée μ_t.

L'application des trois critères d'équité impose donc que le score total pour le groupe *t* comprenant N_t membres soit décrit par $(N_t - \mu_t)x + \mu_t.y = \mu_t$. Il s'agit là d'une équation linéaire : chaque paire de groupes doit donc satisfaire un système de deux équations linéaires en *x, y*. La satisfaction de nos critères d'équité pour deux groupes correspond donc aux solutions de ce système linéaire.

Il n'existe que deux solutions à ce système d'équations. Dans un cas, les deux lignes sont identiques : tous les couples (*x, y*) satisfont l'équation. Si on cherche à interpréter ce résultat mathématique en termes de conditions pesant sur les groupes, cela correspond à l'égalité des taux de base entre groupes, $\mu_1 = \mu_2$. Si les taux de base sont différents, les deux lignes sont différentes et se croisent au point (0, 1). Cette deuxième solution correspond au cas où le score moyen pour la classe négative est 0 et le score moyen pour la classe positive

24 *Ojmarrh Mitchell et Michael S. Caudy, « Examining racial disparities in drug arrests », Justice Quarterly 32, no 2 (2015): 288-313.*

est 1, soit une prédiction parfaite. Pour résumer, les trois critères d'équité sont simultanément satisfaits si et seulement si les taux de base entre groupes sont égaux ou si la prédiction est parfaite.

Il importe de souligner qu'il existe également une version approchée du théorème. Pour tout $\varepsilon > 0$, on peut formuler une version approchée des critères d'équité où l'égalité désirée entre groupes n'est satisfaite qu'à une erreur ε près. Pour tout $\delta > 0$, on peut formuler une version approchée de l'égalité des taux de base et de la prédiction parfaite : l'égalité des taux de base est satisfaite à une erreur δ près, et $y = 1 - \delta$. La version approchée du théorème énonce que la satisfaction approchée des trois critères d'équité n'est possible que par une version approchée de l'égalité des taux de base ou de la prédiction parfaite.

4. Le débat philosophique sur le sens des métriques et de leur incompatibilité

4.1 De l'incompatibilité au pluralisme ?

Le résultat d'incompatibilité est d'autant plus significatif qu'il porte sur deux métriques de l'équité simples et naturelles. Il semble naturel de désirer à la fois que les notes aient la même signification pour tout le monde et que les erreurs favorables et défavorables soient réparties de manière équitable. Néanmoins, dans les cas pratiques normaux où on a affaire à des populations ayant des propriétés statistiques différentes et des modèles imparfaits, cet objectif d'optimisation simultanée est impossible à atteindre[25].

Ces métriques de l'équité sont donc bien divergentes dans un sens profond : il est nécessaire de choisir entre elles. L'analyse statistique montre ici une immédiate pertinence politique, dans la mesure où elle structure la conception des normes : sans elle, il est fort possible que certains législateurs auraient pu essayer d'appliquer deux critères incompatibles à la fois, pour ne découvrir l'impossibilité d'une telle entreprise qu'*a posteriori*, une fois la loi appliquée à de vastes populations avec les difficultés qu'on imagine.

Non seulement la communauté travaillant sur l'équité algorithmique ne prétend donc pas donner la définition mathématique de l'équité, mais elle souligne au contraire l'incompatibilité de différentes conceptions naturelles de l'équité. Un problème philosophique majeur consiste alors à savoir quel sens exact attribuer à cette incompatibilité.

Le résultat d'incompatibilité a pu être utilisé par les partisans d'une interprétation pluraliste des métriques de l'équité. Avant d'expliquer cet usage philosophique du résultat mathématique, il nous faut préciser que ce pluralisme est devenu une position défendue ou à tout le moins discutée par plusieurs auteurs[26], mais ne fait cependant pas l'unanimité. Dans un article récent[27], Hedden a défendu la thèse que seule la calibration doit compter comme une condition statistique nécessaire de l'équité. Son argumentation est basée sur la construction d'un exemple sophistiqué d'un modèle menant manifestement à un résultat équitable, mais ne satisfaisant que le critère de calibration. Mieux encore, l'algorithme violerait les autres critères même dans le cas idéalisé où les taux de base des populations étaient identiques, ce qui montrerait la vanité de ces critères même dans les cas marginaux où ils sont censés être compatibles avec la calibration. Toujours selon cet argument, l'impossibilité de satisfaire simultanément les métriques de calibration et d'égalité des chances ne constitue donc pas un véritable problème d'équité, puisque cette dernière ne constitue pas une véritable métrique de l'équité. En outre, puisque l'algorithme est optimal, il n'est pas même possible de dire que le viol d'un des critères contribuerait dans une certaine mesure à rendre l'algorithme moins équitable (voir Hedden, note 36). Cet argument doit cependant affronter deux problèmes. Le premier est la pertinence de son contre-exemple, qui est si fortement idéalisé que sa reconnaissance comme exemple pose question. Le deuxième est l'inférence faite par l'auteur de l'existence d'un contre-exemple à toutes les métriques sauf la calibration à la conclusion que la calibration est « la bonne métrique[28]. » Nous allons voir tout de suite qu'un autre point de vue philosophique parvient à une toute autre conclusion.

4.2 Impossible équité ou contextualisme ?

Pour le comprendre, il faut discuter plus avant le sens du pluralisme. Même si l'on admet le pluralisme des métriques de l'équité, son sens exact dépend du statut épistémique attribué à ces métriques. L'optimisation d'une métrique ou d'une

25 *Kleinberg, Mullainathan, und Raghavan, „Inherent Trade-Offs in the Fair Determination of Risk Scores".*

26 *Outre l'article de Heidari et al. 2019 déjà mentionné, on pourra trouver une discussion de l'interprétation pluraliste dans Will Fleisher, « Evidence of Fairness: On the Uses and Limitations of Statistical Fairness Criteria », SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 30 novembre 2021), https://doi.org/10.2139/ssrn.3974963. Binns, « What can political philosophy teach us about algorithmic fairness? » Derek Leben, « Normative principles for evaluating fairness in machine learning », in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, 86-92. La position pluraliste peut aussi être mentionnée dans des articles techniques comme A. Feder Cooper et Ellen Abrams, « Emergent Unfairness in Algorithmic Fairness-Accuracy Tradeoff Research », in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21) (Virtual Event, 2021).*

27 *Hedden, « On statistical criteria of algorithmic fairness ».*

28 *Je tiens à remercier Michele Loi de m'avoir fait réalisé ce point crucial.*

conjonction de métriques doit-elle être conçue comme une condition nécessaire, une condition suffisante, ou une condition nécessaire et suffisante de l'équité d'un algorithme ? Ou doit-elle être conçue simplement comme un indice (*evidence*) plaidant en faveur de l'équité de cet algorithme[29] ? Si on suppose que l'optimisation conjointe des deux métriques de calibration et d'égalité des chances constitue une condition nécessaire de l'équité, alors le théorème d'incompatibilité des métriques montre l'impossibilité pratique d'une procédure prédictive équitable : on peut alors parler d' « impossibilité de l'équité[30]. » Il s'agit là d'une interprétation d'une force terrible, dans la mesure où elle impose des arbitrages lourds à toute action basée sur un modèle statistique prédictif et visant à une distribution équitable de ses résultats. Dans une telle perspective, l'emploi contextuel d'une métrique plutôt que d'une autre ne constituerait donc qu'un pis-aller politique face à l'impossibilité d'obtenir l'équité désirée, même si ce pis-aller pourrait être guidé par des considérations de principe.

Cette interprétation pessimiste d'impossibilité de l'équité repose cependant toute entière sur l'hypothèse que seule une conjonction des métriques considérées par le résultat d'incompatibilité pourrait constituer une condition nécessaire de l'équité. Une version particulière du pluralisme, qu'on peut qualifier de « contextualisme », considérerait au contraire que la pertinence d'une métrique dépend elle-même du contexte. Comme on l'a vu dans notre examen de la parité statistique stricte, une métrique pertinente dans certains contextes pourrait devenir contre-productive dans d'autres : on ne saurait donc considérer que son optimisation est une condition nécessaire de l'équité qui devrait s'appliquer dans tous les cas de figure. De telles réflexions ne portent pas que sur la parité statistique, dont on a vu qu'elle n'était pas au centre des débats actuels. Un défenseur de la discrimination positive pourrait ainsi défendre dans le cas du logiciel COMPAS qu'il est équitable de violer le critère de calibration pour garantir l'égalité des chances entre Afro-Américains et Blancs[31]. Sans forcément nier toute pertinence au critère de calibration, qui pourrait être accepté dans d'autres cas de figure, il doit donc nier *contra* Hedden que la calibration soit une condition nécessaire de l'équité dans un cas de figure où l'égalité des chances est posée comme métrique cardinale. Pour un tel contextualisme fort, les résultats d'incompatibilité ne montrent pas que l'équité est impossible, puisqu'il existe des cas où l'on peut abandonner une des métriques concernées sans renoncer à l'équité et sombrer dans un pis-aller politique. Un tel contextualisme peut parfaitement accepter la validité de l'exemple sophistiqué de Hedden, mais pour en tirer une toute autre conclusion. Cet exemple montrera juste un contexte où la métrique de calibration est valide, et où les autres métriques peuvent être négligées. Dans d'autres contextes, ce sera au tour de la calibration d'être prise en défaut par un contexte, et l'exemple de Hedden ne montre donc nullement que la calibration soit la « bonne » métrique de l'équité. Pour récapituler, il ne s'agit pas pour le contextualiste de nier qu'il puisse y avoir des cas de figure où l'on pourrait souhaiter l'application simultanée des trois métriques, et où le résultat d'incompatibilité nous mettra bien face à un dilemme. Il s'agit juste de nier que ce soit le cas général, et d'affirmer qu'il existe des contextes où il est éthiquement légitime de privilégier certaines métriques et de négliger d'autres, et où leur incompatibilité n'impliquera donc nullement une quelconque impossibilité de l'équité.

Dans cette perspective, la tâche du contextualiste est d'articuler les hypothèses permettant de déterminer la métrique pertinente selon le contexte. Le travail d'Heidari et al. va ainsi jusqu'à attribuer un sens moral à l'incompatibilité des conceptions de l'équité de groupe, en voyant cette incompatibilité comme la conséquence d'hypothèses morales bien distinctes, et qui n'ont aucune raison d'être toujours satisfaites simultanément. Nous n'avons guère l'espace nécessaire pour présenter les finesses techniques et philosophiques de cet article, qui tâche de concevoir un cadre unifié de conceptualisation des métriques fondé sur la distinction entre les variables dont les individus peuvent être tenus responsables et celles qui ne sont que le fruit des circonstances, distinction sur laquelle nous allons revenir dans la section suivante. La caractérisation de leur position comme « contextualiste » est de mon fait, mais a été acceptée par l'un des auteurs de l'article, Michele Loi, qui a eu la gentillesse de relire une présentation de ce travail. L'interprétation des résultats d'incompatibilité n'est donc pas purement une affaire de mathématiques et exige le travail délicat de ramener à la surface un ensemble d'hypothèses philosophiques implicites[32].

29 *Cette question du statut épistémique des métriques est placée au cœur de l'opus cité de Fleisher.*

30 *Sorelle A. Friedler, Carlos Scheidegger, et Suresh Venkatasubramanian, « The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making », Communications of the ACM 64, no 4 (2021): 136-43. Pour une autre discussion critique de cette idée fondée sur un idéal d'égalité substantielle, voir Ben Green, « Impossibility of What? Formal and Substantive Equality in Algorithmic Fairness », Formal and Substantive Equality in Algorithmic Fairness (July 9, 2021), 2021.*

31 *Pour une illustration pédagogique de cette décision voir Karen Hao et Jonathan Stray, « Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game », MIT Technology Review, 17 octobre 2019, https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.*

32 *Cependant, comme l'a fait remarquer un relecteur anonyme, les chercheurs en IA pourraient s'interroger sur la possibilité d'automatiser la reconnaissance du*

5. Le difficile problème de la définition des variables

Le problème de la délimitation des variables dont un individu peut être tenu pour responsable, et la pertinence d'employer celles qui ne le sont pas dans une décision sensible, sont deux problèmes centraux de l'algorithmisation de la décision. Si l'on souhaite attribuer un bien social sur une base méritocratique, il faut en effet réussir à sélectionner des variables dont les individus peuvent être tenus pour responsables. Par exemple, une personne ne doit pas se voir refuser un emploi parce qu'elle est née dans un quartier mal famé ou parce que son père a séjourné en prison. Ceci reste vrai même si un modèle statistique nous affirme que ces valeurs de variables sont corrélées à la probabilité d'être un mauvais employé, parce qu'on ne peut prétendre embaucher une personne sur la base de ses mérites et la rejeter ensuite pour des choses qui ne sont pas de son fait. Tout échec dans la délimitation des variables dont on peut être tenu pour responsable condamnera la procédure à être inéquitable, même si elle satisfaisait par exemple une métrique comme l'égalité des chances, car la procédure de décision ne fait tout simplement pas ce qu'elle est censée faire. La littérature en éthique algorithmique le reconnait pleinement, puisque cette distinction est reconnue par Heidari et al.

Ce qui est vrai pour les variables dont nous sommes moralement responsables est encore plus vrai pour le choix des populations discriminées ou des attributs sensibles. Comme le remarque fort justement Zliobaite 2017[33], le rôle de l'analyse des données n'est pas de dire quelle forme de distinctions entre populations est juste ou injuste. Elle doit simplement prendre comme entrées les caractéristiques considérées comme sensibles par le droit, et aussi le cas échéant par la philosophie politique, les sciences sociales, les décideurs politiques et l'ensemble du débat politique agitant la société dans son ensemble. L'équité algorithmique ne peut donc que prendre en entrée un ensemble de positions sur les problèmes les plus centraux de la discrimination, comme la délimitation des populations discriminées, la nature des dimensions de la discrimination et le type de mesure à adopter dans les politiques publiques.

Le problème de la délimitation des variables prises en entrées, et du sens qu'on leur confère, est cependant décisif pour comprendre les effets politiques de l'emploi d'un algorithme. Si bien intentionnée et humble que soit l'attitude défendue par Zliobaite, elle doit en effet affronter un problème politique majeur, à savoir que l'équité algorithmique risque d'être appliquée à des populations dont la délimitation a été forcée par les institutions, et qui sont donc elles-mêmes problématiques. L'un des exemples les plus célèbres et les plus tragiques d'une telle catégorisation forcée est la division de la population rwandaise en deux « ethnies » Hutu et Tutsi. Du point de vue des historiens contemporains[34], une telle ethnisation de cette catégorie sociale complexe est aussi ridicule que l'ancienne théorie dite de la « double nationalité » voulant que le petit peuple français descendait des Gaulois tandis que les aristocrates français descendait des envahisseurs germaniques, théorie a qui d'ailleurs inspirée directement les colonisateurs belges de la région. Malgré sa complète absence de fondement historique, les autorités coloniales, puis le Rwanda indépendant ont continué à appliquer cette division de la population, engendrant une division politique et des rancunes tout à fait réelles, avec la fin tragique que l'on sait. Dans un tel contexte, appliquer une métrique de l'équité comme par exemple l'égalité des chances aux populations hutu et tutsi n'aurait guère contribué à faire régner l'équité : elle aurait conféré une légitimité scientifique à un non-sens complet, et contribuer ainsi à monter les populations les unes contre les autres. L'acceptation des variables par les concepteurs d'un modèle, et en particulier des variables décrivant des populations, est en elle-même un geste pourvu d'une portée éthique et politique. Nul bien ne peut provenir de l'application de critères statistiques à des divisions dépourvues de sens, et le scientifique ne peut espérer participer à un œuvre de justice en acceptant aveuglément les catégorisations de populations en entrées, car l'application des métriques d'équité n'a de sens éthique que si elle prend pour objet des catégories douées de sens.

Un autre exemple du même problème peut être trouvé dans une enquête sur les règles de modération du discours de haine (*hate speech*) par Facebook (désormais Méta) courant 2017[35]. Facebook définissait alors le discours de haine comme un discours dirigé contre une population délimitée par une variable protégée ou une conjonction de variables protégées (*protected variables*), l'équivalent américain des informations sensibles du droit européen. Une telle définition bureaucratique avait de nombreuses conséquences discutables. Un discours dirigé contre les « hommes blancs » était ainsi qualifié de discours haineux, parce que la population visée est délimitée par deux variables protégées, la

contexte dans lequel s'applique le choix de métrique. Il s'agit là d'une question difficile, qui à notre connaissance n'est pas explorée dans la littérature.

33 Indrė Žliobaitė, « Measuring discrimination in algorithmic decision making », Data Mining and Knowledge Discovery 31, no 4 (2017): 1060-89.

34 Dominique Franche. Rwanda. Généalogie d'un génocide. Editions Mille et une nuits, 1997.
Pour une présentation du sens social véritable des catégories « hutu » et « tutsi », et de leur instrumentalisation et déformation par les autorités coloniales et par les régimes post-indépendance au Burundi et au Rwanda, voir Stephen B. Isabyrie et Kooros M. Mahmoudi, « Rwanda, Burundi and their tribal wars », Social Change 31, no 4 (décembre 2001): 49-69.

35 Ariana Tobin Julia Angwin, « Facebook (Still) Letting Housing Advertisers Exclude… », ProPublica, novembre 2017, https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

race et le genre. En revanche, un discours dirigé contre les enfants noirs n'était pas considéré comme discours de haine, parce qu'il ne s'agit que d'un sous-groupe d'une population désignée par une variable protégée. Une telle classification, outre qu'elle ne tient pas compte de la position des populations dans les rapports de domination établis, fait beau jeu du fait que si l'on déverse sa haine en ligne contre des enfants noirs, c'est probablement parce qu'ils sont noirs. Dans un autre exemple, un appel à tuer tous les islamistes suite à un attentat terroriste n'est pas considéré comme un discours de haine, parce qu'il ne désigne qu'un sous-groupe d'une population désignée par une variable protégée de confession, et non la population entière. Une telle décision est problématique à un double titre. D'abord, parce que la confusion entre musulmans, musulmans pratiquants, islamistes et islamistes djihadistes est au cœur des discours de haine à l'encontre des musulmans, et qu'il est donc légitime pour la population musulmane entière de se sentir inquiéter par un tel discours. Ensuite, parce qu'elle ne tient aucun compte du contenu du discours et de ses effets psychologiques : il est difficile de supposer qu'un appel au meurtre invite à faire des distinctions politiques fines, plutôt qu'à se déchaîner sur le premier musulman venu pour passer ses nerfs. Indépendamment des positions des uns et des autres sur ce qui doit constituer un discours de haine, on voit à nouveau que l'application d'une métrique de l'équité comme par exemple l'égalité des chances à l'algorithme de modération des discours de haine n'est en aucun cas une véritable garantie de traitement équitable, si la classification des données linguistiques comme « discours de haine » est mal conçue. Là encore, *garbage in, garbage out :* l'application d'une métrique de l'équité à des classes mal choisies vient légitimer un non-sens et donner l'illusion d'un travail. Il ne s'agit pas là des problèmes classiques dans la littérature du ML sur la qualité des données ou de leur collecte, mais d'une interrogation, plus familière des sciences sociales, sur le choix et le sens même des catégories employées pour classer des populations.

6. Conclusion

La réflexion en cours sur les métriques statistiques permettant de déterminer si un algorithme est équitable est à bien des égards originale et digne d'intérêt philosophique. Ces métriques ne sont pas des indicateurs statistiques grossiers et réducteurs d'une réalité sociale complexe, dont l'histoire des statistiques montre malheureusement bien des exemples[36]. Elles ne sont bien sûr pas non plus, et ne prétendent pas non plus être, une forme de « définition mathématique de l'équité. » Elles formalisent diverses notions intuitives et naturelles de l'équité, et des résultats comme le théorème d'incompatibilité contribuent à montrer la complexité des relations entre ces intuitions. Elles ont ainsi une véritable portée conceptuelle en même temps que pratique. Il reste cependant beaucoup de travail à faire pour préciser les hypothèses morales qui sous-tendent chacune de ces métriques, et pour examiner leur pertinence respective selon le contexte d'usage. Une fertilisation croisée de l'éthique de l'IA et des traditions critiques portant sur les catégories statistiques[37], les catégorisations des appartenances identitaires ou les concepts juridiques[38] est encore nécessaire, pour éviter que ces réflexions ne se perdent dans la pratique par application à des variables dépourvues de pertinence.

RÉFÉRENCES

Abu-Elyounes, Doaa. « Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness ». *University of Illinois Journal of Law, Technology & Policy* 2020, nº 1 (2020): 1-54.

Angwin, Julia, Jeff Larson, Surya Mattu, et Lauren Kirchner. « Machine Bias ». ProPublica, 23 mai 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=siiaBuUx_5-LH2f_432kxejIHJI-dlxM.

Barocas, Solon, Moritz Hardt, et Arvind Narayanan. *Fairness and Machine Learning. Limitations and opportunities*. fairmlbook.org, 2019. https://fairmlbook.org/.

Binns, Reuben. « Fairness in Machine Learning: Lessons from Political Philosophy ». In *Conference on Fairness, Accountability and Transparency*, 149-59, 2018. http://proceedings.mlr.press/v81/binns18a/binns18a.pdf.

Binns, Reuben. « On the Apparent Conflict Between Individual and Group Fairness ». In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372864.

Binns, Reuben. « What can political philosophy teach us about algorithmic fairness? » *IEEE Security & Privacy* 16, nº 3 (2018): 73-80.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, et Nigel Shadbolt. « "It's Reducing a Human Being to a Percentage" Perceptions of Justice in Algorithmic Decisions ». In *Proceedings of the 2018 Chi conference on human factors in computing systems*, 1-14, 2018.

36 Pour une critique sociologique de l'abus des métriques, voir par exemple Jerry Z. Muller, The Tyranny of Metrics (Princeton University Press, 2019).

37 Pour ne prendre que deux exemples connus, qu'on songe à la sociologie historique de la quantification Alain Desrosières, Pour une sociologie historique de la quantification: L'Argument statistique I (Presses des MINES, 2008); Alain Desrosières, Gouverner par les nombres: L'argument statistique II (Presses des Mines via OpenEdition, 2013). Pierre Bourdieu et al., Sur l'état. Cours au Collège de France (1989-1992). (Seuil, 2012). (En particulier la première leçon et la leçon du 7 Février 1991). Pour une discussion du statut délicat des statistiques officielles pour un problème sociologique comme la « déviance », voir John I. Kitsuse et Aaron V. Circourel, « A note on the uses of official statistics », Social Problems 11 (1963): 131-39.

38 Pour un article appelant à l'intégration des travaux techniques en éthique de l'IA avec les perspectives constructivistes sur les appartenances identitaires, notamment de race, voir Alex Hanna et al., « Towards a critical race methodology in algorithmic fairness », in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, 501-12.

REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Bourdieu, Pierre, Patrick Champagne, Rémi Lenoir, Franck Poupeau, et Marie-Christine Rivière. *Sur l'état. Cours au Collège de France (1989-1992)*. Seuil, 2012.

Cavazos, Jacqueline G., P. Jonathon Phillips, Carlos D. Castillo, et Alice J. O'Toole. « Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? » *IEEE transactions on biometrics, behavior, and identity science* 3, nº 1 (2020): 101-11.

Chouldechova, Alexandra. « Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments ». *Big Data* 5, nº 2 (juin 2017): 153-63. https://doi.org/10.1089/big.2016.0047.

Christin, Angèle. « Algorithms in Practice: Comparing Web Journalism and Criminal Justice ». *Big Data & Society* 4, nº 2 (décembre 2017): 1-14. https://doi.org/10.1177/2053951717718855.

Christin, Angèle, Alex Rosenblat, et Danah Boyd. « Courts and predictive algorithms ». *Data & CivilRight*, 2015.

Cooper, A. Feder, et Ellen Abrams. « Emergent Unfairness in Algorithmic Fairness-Accuracy Tradeoff Research ». In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Virtual Event, 2021.

Desrosières, Alain. *Gouverner par les nombres: L'argument statistique II*. Presses des Mines via OpenEdition, 2013.

Desrosières, Alain. *Pour une sociologie historique de la quantification: L'Argument statistique I*. Presses des MINES, 2008.

Dolata, Mateusz, Stefan Feuerriegel, et Gerhard Schwabe. « A Sociotechnical View of Algorithmic Fairness ». *Information Systems Journal* n/a, nº n/a. Consulté le 19 mai 2022. https://doi.org/10.1111/isj.12370.

Dressel, Julia, et Hany Farid. « The Accuracy, Fairness, and Limits of Predicting Recidivism ». *Science Advances* 4, nº 1 (17 janvier 2018): eaao5580. https://doi.org/10.1126/sciadv.aao5580.

Dressel, Julia, et Hany Farid. « The Dangers of Risk Prediction in the Criminal Justice System ». *MIT Case Studies in Social and Ethical Responsibilities of Computing*, nº Winter 2021 (5 février 2021). https://doi.org/10.21428/2c646de5.f5896f9f.

equivant. « Response to ProPublica: Demonstrating Accuracy Equity and Predictive Parity ». equivant, 1 décembre 2018. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Saint Martin's Press, 2018.

Fleisher, Will. « Evidence of Fairness: On the Uses and Limitations of Statistical Fairness Criteria ». SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 30 novembre 2021. https://doi.org/10.2139/ssrn.3974963.

Flores, Alejandro, Kristin Bechtel, et Christopher T. Lowenkamp. « False Positives, False Negatives, and False Analyses: a Rejoinder to « Machine Bias: There's Software Used across the Country to Predict Future Criminals. and It's Biased against Blacks. " ». *Federal Probation Journal* 80, nº 2 (2016): 38-46.

Friedler, Sorelle A., Carlos Scheidegger, et Suresh Venkatasubramanian. « The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making ». *Communications of the ACM* 64, nº 4 (2021): 136-43.

Green, Ben. « Impossibility of What? Formal and Substantive Equality in Algorithmic Fairness ». *Formal and Substantive Equality in Algorithmic Fairness (July 9, 2021)*, 2021.

Greenberg, Irwin. « An analysis of the EEOCC "Four-Fifths" rule ». *Management Science* 25, nº 8 (1979): 762-69.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, et Adrian Weller. « Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning ». In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018.

Hanna, Alex, Emily Denton, Andrew Smart, et Jamila Smith-Loud. « Towards a critical race methodology in algorithmic fairness ». In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501-12, 2020.

Hao, Karen, et Jonathan Stray. « Can You Make AI Fairer than a Judge? Play Our Courtroom Algorithm Game ». *MIT Technology Review*, 17 octobre 2019. https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

Hedden, Brian. « On statistical criteria of algorithmic fairness ». *Philosophy and Public Affairs* 49, nº 2 (2021).

Heidari, Hoda, Michele Loi, Krishna P. Gummadi, et Andreas Krause. « A Moral Framework for Understanding Fair ML Through Economic Models of Equality of Opportunity ». In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 181-90. FAT* '19. New York, NY, USA: ACM, 2019. https://doi.org/10.1145/3287560.3287584.

Howell, R. Rodney. « The High Price of False Positives ». *Molecular Genetics and Metabolism*, Special Issue Articles Authored by SIMD Presidents and their Colleagues, 87, nº 3 (1 mars 2006): 180-83. https://doi.org/10.1016/j.ymgme.2005.10.004.

Isabyrie, Stephen B., et Kooros M. Mahmoudi. « Rwanda, Burundi and their tribal wars ». *Social Change* 31, nº 4 (décembre 2001): 49-69.

Julia Angwin, Ariana Tobin. « Facebook (Still) Letting Housing Advertisers Exclude… ». *ProPublica*, novembre 2017. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

Kearns, Michael, Seth Neel, Aaron Roth, et Zhiwei Steven Wu. « Preventing fairness gerrymandering: Auditing and learning for subgroup fairness ». In *International Conference on Machine Learning*, 2564-72. PMLR, 2018.

Kitsuse, John I., et Aaron V. Circourel. « A note on the uses of official statistics ». *Social Problems* 11 (1963): 131-39.

Kleinberg, Jon, Sendhil Mullainathan, et Manish Raghavan. « Inherent Trade-Offs in the Fair Determination of Risk Scores ». In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, édité par Christos H. Papadimitriou, 67:43:1-43:23. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Lambrecht, Anja, et Catherine Tucker. « Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads ». *Management science* 65, nº 7 (2019): 2966-81.

Leben, Derek. « Normative principles for evaluating fairness in machine learning ». In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 86-92, 2020.

Ledford, Heidi. « Millions of Black People Affected by Racial Bias in Health-Care Algorithms ». *Nature* 574, nº 7780 (octobre 2019): 608-9. https://doi.org/10.1038/d41586-019-03228-6.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, et Aram Galstyan. « A survey on bias and fairness in machine learning ». *ACM Computing Surveys (CSUR)* 54, nº 6 (2021): 1-35.

Mitchell, Ojmarrh, et Michael S. Caudy. « Examining racial disparities in drug arrests ». *Justice Quarterly* 32, nº 2 (2015): 288-313.

LATOSENSU
REVUE
DE LA SOCIÉTÉ
DE PHILOSOPHIE
DES SCIENCES

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, et Kristian Lum. « Algorithmic fairness: Choices, assumptions, and definitions ». *Annual Review of Statistics and Its Application* 8 (2021): 141-63.

Muller, Jerry Z. *The Tyranny of Metrics*. Princeton University Press, 2019.

« Nursing and Midwifery ». Consulté le 15 mai 2022. https://www.who.int/news-room/fact-sheets/detail/nursing-and-midwifery.

Schnabel, Tobias, Adith Swaminathan, Ashudeep Singh, Navin Chandak, et Thorsten Joachims. « Recommendations as treatments: Debiasing learning and evaluation ». In *international conference on machine learning*, 1670-79. PMLR, 2016.

Thomson, Cara. « Myths and Facts. Using Risk and Need Assessments to Enhance Outcomes and Reduce Disparities in the Criminal Justice System ». National Institute of Corrections & Community Corrections Collaborative Network, mars 2017.

Verma, Sahil, et Julia Rubin. « Fairness definitions explained ». In *2018 ieee/acm international workshop on software fairness (fairware)*, 1-7. IEEE, 2018.

Žliobaitė, Indrė. « Measuring discrimination in algorithmic decision making ». *Data Mining and Knowledge Discovery* 31, nº 4 (2017): 1060-89.

HISTORIQUE

État de l'art soumis le 28 janvier 2022.
État de l'art accepté le 19 décembre 2022.

SITE WEB DE LA REVUE

https://ojs.uclouvain.be/index.php/latosensu

DOI

https://doi.org/10.20416/LSRSPS.V10I1.7

CONTACT ET COORDONÉES

Maël Pégny
Université de Lorraine
maelpegny@gmail.com

SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES (SPS)

École normale supérieure
45, rue d'Ulm
75005 Paris

SOCIÉTÉ DE PHILOSOPHIE DES SCIENCES (SPS)

École normale supérieure
45, rue d'Ulm
75005 Paris
www.sps-philoscience.org