Métriques d'équité en Apprentissage Automatique et droit de l'Union Européenne en matière de non-discrimination

M. Legast[*1], Y. Yousefi[*2, 3], L. Koutsoviti Koumeri[*4], A. Legay[1], C. Schommer[3], K. Vanhoof[4]

[1] Université catholique de Louvain, ICTEAM
[2] Università di Bologna, CIRSFID Alma AI
[3] Université du Luxembourg, DP CSCE
[4] Universiteit Hasselt, BINF

Juillet 2023

Résumé

Les modèles d'apprentissage automatique (AA) peuvent présenter des biais discriminatoires envers certains groupes sociaux. Nous étudions à quel point les techniques et définitions d'équité utilisées en AA peuvent garantir le respect du droit de l'UE en matière de non discrimination. À travers des modèles de classification entraînés avec différentes contraintes d'équité, nous évaluons l'efficacité des méthodes de correction de biais et discutons les résultats sous l'angle de l'AA et de l'informatique juridique.

Mots-clés

Apprentissage automatique, Droit de l'Union Européenne, Décision algorithmique, Équité, Non-discrimination

Abstract

Machine Learning (ML) models have been shown to present biases leading to discrimination against certain social groups. Our research studies the extend to which ML techniques and fairness definitions can ensure compliance with the EU non-discrimination legal framework. Using classification models trained with different fairness constraints, we evaluate how effective the bias mitigation process is and discuss the results using both an ML approach and legal informatics methodology.

Keywords

Machine learning, European Union law, Algorithmic decision-making, Fairness, Non-discrimination

[*] All these authors have contributed equally.

1 Introduction

Les décisions algorithmiques prises par des modèles d'Apprentissage Automatique (AA) ont un impact important et croissant sur nos vies. L'implication de l'intelligence artificielle (IA) dans des domaines conséquents amène de nombreuses questions éthiques et légales, telle que la question de l'équité. Il y a notamment un certain nombre d'études qui font état des biais qui peuvent se trouver dans de tels systèmes de décisions et des résultats discriminants qui peuvent en découler [16]. En réaction, un certain nombre de recherches ont été menées dans le domaine de l'équité pour l'IA et l'AA.

Néanmoins, établir une définition de l'équité et garantir des résultats qui ne contiennent pas de biais discriminatoires reste une question ouverte et difficile, que ce soit pour les décisions algorithmiques [20] ou humaines. En informatique, de nombreuses définitions d'équité et de discrimination existent, faisant appel à des notions mathématiques et des concepts éthiques différents. Les différentes métriques d'équité qui en découlent sont utilisées pour donner une évaluation chiffrée du niveau d'équité ou de discrimination d'une BDD ou des prédictions d'un modèle. Plusieurs méthodes pour éviter et réduire les biais ont également été proposées [16]. En droit, la notion d'équité est généralement liée à celles d'égalité et de (non-)discrimination. C'est notamment le cas au niveau de l'Union Européenne (UE) dont le cadre légal est celui qui est pris en considération dans cet article.

Plusieurs analyses de la littérature existante, à la fois au niveau juridique [12] et technique [7], font état d'un écart entre le droit en matière de non-discrimination et la manière dont la recherche en informatique répond à cette question. Ces articles encouragent donc des recherches interdisciplinaires, notamment sur la compatibilité entre les définitions mathématiques et juridiques de l'équité.

Notre recherche répond à ce constat en analysant la concordance entre les méthodes et définitions liées à l'équité dans le domaine de l'AA d'une part et le cadre réglementaire de l'UE en matière de non-discrimination d'autre part. Nous évaluons à quel point les définitions mathématiques d'équité et les métriques et méthodes de correction de biais associées peuvent permettre le respect de ce cadre réglementaire. Nous étudions aussi leur pertinence pour détecter et prouver une discrimination.

Notre méthode de recherche combine des expérimentations avec des modèles de prédiction en classification et l'usage de méthodologie d'informatique juridique, qui interprète et adapte le concept légal d'équité aux paradigmes des nouvelles technologies et inversement [18]. Pour comparer plusieurs scénarios, nous entraînons des modèles avec diffé-

rentes contraintes d'équité. Nous analysons ensuite, à l'aide de plusieurs métriques, le niveau de biais dans les bases de données (BDD) d'entraînement et dans les prédictions émises par ces différents modèles et leurs équivalents non contraints. Nous discutons les résultats obtenus d'un point de vue technique et au regard du cadre réglementaire de l'UE sur la non-discrimination.

2 L'équité, ses approches et enjeux

L'équité est un concept pour lequel il est difficile de donner une définition unique, claire et sans ambiguïté. Cette notion peut être interprétée différemment suivant notamment les contextes, les cultures et les individus [11].

Plusieurs approches co-existent, tant au niveau éthique ou légal qu'en AA. Certaines définitions se complètent et peuvent être appliquées simultanément tandis que d'autres sont incompatibles entre elles [3], [15]. Le choix de la définition d'équité considérée ainsi que la manière de l'implémenter est donc particulièrement important, à la fois dans le développement d'IA et au niveau du cadre réglementaire.

2.1 L'équité dans le droit de l'UE

L'équité dans le droit de l'UE est généralement envisagée via un cadre réglementaire de non-discrimination qui promeut l'égalité. La notion juridique d'équité dans l'UE découle de l'article 21 de la Charte des droits fondamentaux de l'UE et de l'article 14 de la Convention européenne des droits de l'homme. Ces deux textes interdisent les discriminations fondées sur certains attributs sensibles comme l'origine sociale, la religion ou le sexe.

Sous ce cadre réglementaire, les **discriminations directes** et **indirectes** fondées sur des critères protégés sont illégales [1], à moins qu'un objectif légitime, jugé approprié et nécessaire, puisse objectivement et raisonnablement le justifier [17], [22]. Nous utiliserons les termes ***discrimination explicable*** et *attribut explicatif* pour parler d'une différence de traitement avec une telle justification et de la caractéristique qui la justifie, selon la terminologie de [14].

Si les discriminations directes et indirectes sont explicitement abordées, les discriminations fondées sur d'autres critères ou sur la combinaison de plusieurs de ces critères (discriminations intersectionelles [5]) sont moins protégées et sont peu présentes dans la jurisprudence. Ces différentes notions sont autant d'aspects qui doivent être pris en compte pour la réalisation d'IA équitables.

À un autre niveau, on retrouve le principe d'**égalité formelle** qui considère qu'il faut traiter tous les individus de la même manière pour éviter les discriminations. Cette notion est à distinguer de l'**égalité matérielle** (*substantive equality*) qui implique de tenir compte du contexte social et des inégalités historiques pré-existantes pour les corriger et atteindre une égalité effective. D'après la jurisprudence de la Cour de justice de l'Union européenne (CJUE), l'objectif du droit anti-discrimination n'est pas seulement de garantir une égalité formelle, mais aussi d'atteindre une égalité matérielle, ce qui nécessite de tenir compte des différences entre groupes de population [6].

Pour développer des IA qui participent à la réduction des inégalités sociales et soient en cohérence avec le droit de l'UE, ces notions doivent être prises en compte dans l'AA, en particulier dans le choix des métriques d'équité. Dans cette optique, Wachter et al. [21] catégorise ces métriques en deux groupes. Les **métriques conservatrices de biais** (*bias preserving*) reproduisent les performances historiques avec un taux d'erreur par chaque groupe semblable à celui des données d'entraînement (qui contiennent généralement des biais [16]). Les **métriques transformatrices de biais** (*bias transforming*) comparent les taux de résultats favorables entre les différents groupes et prennent en compte les biais sociaux en nécessitant une décision explicite quand aux biais qui devraient être présents dans le système. Ce deuxième type de métrique est d'avantage en adéquation avec le principe d'égalité substantive et donc le droit anti-discrimination de l'UE.

Enfin, nous abordons également le caractère **contextuel** de l'équité dans les décisions juridiques. Cet aspect contextuel implique que chaque situation peut être traitée différemment dans différents contextes et en considérant des éléments différents. Ainsi, la manière dont les tribunaux considèrent l'égalité et l'équité peut varier d'un cas à l'autre. En effet, prendre une décision équitable demande généralement d'effectuer un jugement sur base de multiples facteurs spécifiques à une situation précise [11]. Cette approche contextuelle se retrouve dans la jurisprudence de l'UE [22].

2.2 Enjeux particuliers liés à l'AA

Plusieurs publications, notamment dans le domaine juridique, pointent de nouvelles difficultés pour détecter et comprendre les discriminations algorithmiques par rapport aux cas de discriminations humaines plus évidents [22]. Les discriminations algorithmiques ont souvent un caractère opaque ou intangible [19] et il peut être plus difficile pour les victimes d'avoir un élément de comparaison qui permet de repérer une discrimination [22]. À ceci s'ajoute souvent un manque d'accès aux données et algorithmes utilisés. Détecter ces discriminations et prouver leur occurrence en justice est donc une tâche ardue [10], [23].

En plus de cela, les décisions algorithmiques peuvent être fondées sur des nouveaux attributs ou des catégorisations moins évidentes. Ceci peut mener à des discriminations reposant sur de nouveaux éléments et pas seulement sur les critères légalement protégés. Les IA peuvent donc avoir des comportements discriminatoires injustes, mais pas illégaux, ce qui constitue un frein à l'accomplissement de l'objectif juridique d'égalité contre lequel les législations actuelles ne sont pas nécessairement suffisantes [23].

Enfin, notons que la caractéristique contextuelle de l'équité ajoute une difficulté supplémentaire au développement d'IA équitables et compatibles avec le cadre réglementaire de l'UE. D'après [22], la prise de décisions algorithmiques au niveau légal et éthique doit être conditionnée au fait que les systèmes soient capables de reproduire l'approche juridique d'égalité contextuelle.

2.3 Traitement de l'équité en classification

Dans le domaine de l'AA, de nombreux travaux ont proposé des manières de détecter, mesurer et corriger les discriminations [16]. Dans notre recherche, nous considérons en particulier le problème canonique de la classification. Ce problème consiste en la prédiction de la classe d'une nouvelle observation en utilisant les connaissances apprises à partir d'autres observations pour lesquelles la classe était connue. Le problème de la classification équitable considère un ou plusieurs attribut(s) protégé(s) duquel (ou desquels) les prédictions ne peuvent pas dépendre.

Pour ce problème uniquement, plus de 90 définitions d'équité et méthodes de correction des biais ont été recensées [13]. Il n'existe cependant pas de méthode ni de définition ou métrique qui réponde à tous les contextes et toutes les contraintes d'équité à prendre en considération. Certaines méthodes et métriques peuvent être combinées pour améliorer les résultats, mais d'autres sont incompatibles entre elles. Il est donc important de considérer les spécificités des définitions et des méthodes correspondantes afin de faire le choix le plus approprié selon le contexte.

À ce stade de notre recherche, nous analysons deux définitions d'équité en particulier, qui sont les deux suivantes :

Demographic Parity (DP) correspond à une équivalence entre la probabilité d'obtenir une prédiction $\hat{y}$ favorable pour une personne appartenant au groupe privilégié ($G = priv$) ou au groupe défavorisé ($G = def$) [9].

$$P(\hat{y} = +|G = def) = P(\hat{y} = +|G = priv)$$

Cette définition d'équité est la plus couramment utilisée dans la recherche pour l'équité en classification [13]. Elle est fondée sur le principe d'équité de groupe qui requiert un traitement égal à l'échelle des différents groupes [1] [9].

La différence ou le ratio entre les deux probabilités constituent des métriques d'équité associées à cette définition. Dans le cas de DP, ces métriques permettent de détecter les discriminations directes et indirectes. Il s'agit également de métriques transformatrices de biais [21] qui peuvent donc être utilisées pour mesurer le niveau d'égalité matérielle.

Cette approche qui évalue le traitement différencié global a néanmoins été critiquée, notamment parce qu'elle ne permet pas de différencier entre discrimination illégale et explicable et que son utilisation peut également mener à des discriminations inverses [14] [22].

Conditional Demographic Disparity (CDD) considère que l'équité est atteinte lorsque la proportion de personnes défavorisées ($G = def$) parmi celles qui obtiennent une prédiction $\hat{y}$ favorable équivaut à leur proportion parmi celles qui en obtiennent une défavorable, en considérant un attribut explicatif R [22].

$$P(G = def|\hat{y} = +, R = r) = P(G = def|\hat{y} = -, R = r)$$

La prise en compte d'un tel attribut a pour but de corriger les limitations de DP précitées [14] et de se rapprocher de l'approche contextuelle du droit [22]. Il s'agit également d'une définition fondée sur l'équité de groupe, qui permet de mesurer des discriminations directes et indirectes et dont les métriques associées sont considérées comme transformatrices de biais [21].

1. Ceci le distingue de l'équité individuelle qui requiert un traitement similaire entre individus semblables. [16]

3 Méthodologie

Nous considérons le problème de la classification binaire avec un unique attribut protégé binaire. Notre objectif est d'analyser l'impact de la définitions d'équité considérée sur les résultats, à la fois au niveau de la mesure des biais et de leur correction.

Pour ce faire, nous entraînons et comparons plusieurs modèles de classification entraînés avec une correction de biais imposée via une contrainte sur l'équité. Cette contrainte prend la forme d'une fonction qui mesure le niveau d'équité du modèle et dont la valeur ne peut pas descendre sous un certain seuil. Nous utilisons pour cela le méta-algorithme présenté dans [4] qui permet d'imposer une contrainte appartenant à un large choix de métriques d'équité. Cela nous permet de comparer des modèles de prédiction entraînés avec des contraintes correspondant à différentes définitions, tout en gardant le reste de l'algorithme inchangé. Nous employons l'implémentation disponible via AIF360 [2] qui utilise l'algorithme du gradient (*gradient descent*) pour l'apprentissage.

Nous considérons différents scénarios qui correspondent à la combinaison d'une contrainte d'équité, d'une BDD d'apprentissage et du choix d'un attribut protégé. Cela nous donne par exemple le scénario de la prédiction du risque de récidivisme pour des justiciables (BDD COMPAS [8]) avec une correction du biais racial selon la définition d'équité DP. Pour chacun de ces scénarios, nous créons différents modèles en faisant varier la force de la contrainte (c'est à dire le seuil minimal d'équité imposé) de 0 (pas de correction des biais) à 1 (contrainte correspondant à une équité "parfaite").

Nous mesurons ensuite le niveau de biais avec plusieurs métriques, notamment DP et CDD, pour chacun des modèles ainsi créés et pour les données considérées comme la vérité terrain (*ground truth*). A ce stade, nous avons utilisé DP comme contrainte d'équité et allons inclure des contraintes basées sur d'autres définitions dans le futur.

4 Résultats et contribution attendues

Cette recherche apporte plusieurs contributions. Tout d'abord, nous comparons l'impact de différentes contraintes d'équité et niveau de contraintes sur les performances des modèles, et ce, en considérant plusieurs approches de l'équité.

Nous apportons également une analyse juridique aux résultats d'expérimentations sur la mesure et réduction des biais, qui sont généralement abordés principalement d'un point de vue technique. Nous considérons notamment le choix du niveau de contrainte pour obtenir un bon compromis entre la précision et l'équité du modèle. Nous explorons une approche juridique pour le choix de valeurs seuils

impliquant de tels compromis. Nous tenons compte pour cela de l'approche contextuelle du droit et de l'objectif du cadre réglementaire anti-discrimination de l'UE d'atteindre une égalité matérielle. Cette approche nécessite de ne pas se limiter à l'égalité formelle et à la correction technique des biais, mais également de reconnaître et lutter contre les différences sociales historiques et actuelles entre différents groupes, comme le souligne la jurisprudence de la CJUE.
Enfin, nous suggérons l'introduction de marges concernant les valeurs d'équité et de précision minimales autorisées dans la législation à venir sur l'intelligence artificielle (*Artificial Intelligence Act*). Celles-ci permettraient d'avoir des lignes directrices claires pour le développement des IA. Cela participerait également à la protection des personnes impactées par les discriminations algorithmiques, à la fois en prévenant l'occurrence de telles discriminations et en facilitant le recours à la justice lorsqu'elles se produisent néanmoins.

Références

[1] M. BELL, "The right to equality and non-discrimination," *Economic and Social Rights under the EU Chatter of Fundamental Rights : A Legal Perspective*, p. 91-110, 2003.

[2] R. K. E. BELLAMY, K. DEY, M. HIND et al., *AI Fairness 360 : An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*, 2018.

[3] R. BERK, H. HEIDARI, S. JABBARI, M. KEARNS et A. ROTH, "Fairness in criminal justice risk assessments : The state of the art," *Sociological Methods & Research*, t. 50, n° 1, p. 3-44, 2021.

[4] L. E. CELIS, L. HUANG, V. KESWANI et N. K. VISHNOI, "Classification with fairness constraints : A meta-algorithm with provable guarantees," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, p. 319-328.

[5] K. W. CRENSHAW, "Mapping the margins : Intersectionality, identity politics, and violence against women of color," in *The public nature of private violence*, Routledge, 2013, p. 93-118.

[6] M. DE VOS, "The European Court of Justice and the march towards substantive equality in European Union anti-discrimination law," *International Journal of Discrimination and the Law*, t. 20, n° 1, p. 62-87, 2020.

[7] M. DOLATA, S. FEUERRIEGEL et G. SCHWABE, "A sociotechnical view of algorithmic fairness," *Information Systems Journal*, t. 32, p. 754-818, 2021.

[8] J. DRESSEL et H. FARID, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, t. 4, 2018.

[9] C. DWORK, M. HARDT, T. PITASSI, O. REINGOLD et R. ZEMEL, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, sér. ITCS '12, New York, NY, USA : ACM, 2012, p. 214-226.

[10] P. HACKER, "Teaching fairness to artificial intelligence : existing and novel strategies against algorithmic discrimination under EU law," *Common Market Law Review*, t. 55, n° 4, 2018.

[11] N. HELBERGER, T. ARAUJO et C. H. de VREESE, "Who is the fairest of them all ? Public attitudes and expectations regarding automated decision-making," *Computer Law & Security Review*, t. 39, 2020.

[12] D. HELLMAN, "MEASURING ALGORITHMIC FAIRNESS," *Virginia Law Review*, t. 106, n° 4, p. 811-866, 2020.

[13] M. HORT, Z. CHEN, J. ZHANG, F. SARRO et M. HARMAN, "Bias Mitigation for Machine Learning Classifiers : A Comprehensive Survey," *ArXiv*, 2022.

[14] F. KAMIRAN, I. ŽLIOBAITĖ et T. CALDERS, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making," *Knowledge and information systems*, t. 35, n° 3, p. 613-644, 2013.

[15] J. KLEINBERG, S. MULLAINATHAN et M. RAGHAVAN, "Inherent trade-offs in the fair determination of risk scores," *arXiv*, 2016.

[16] N. MEHRABI, F. MORSTATTER, N. SAXENA, K. LERMAN et A. GALSTYAN, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, t. 54, n° 6, 115 :1-115 :35, 2021.

[17] D. MOECKLI et al., "Equality and non-discrimination," *International human rights law*, p. 189-208, 2010.

[18] G. SARTOR, "Informatica giuridica," *Il diritto nella società dell'informazione*, 2006.

[19] A. SIAPKA, "The Ethical and Legal Challenges of Artificial Intelligence : The EU response to biased and discriminatory AI," *SSRN*, 2018.

[20] S. VERMA et J. RUBIN, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness*, IEEE, ACM, 2018, p. 1-7.

[21] S. WACHTER, B. MITTELSTADT et C. RUSSELL, "Bias preservation in machine learning : the legality of fairness metrics under EU non-discrimination law," *W. Va. L. Rev.*, t. 123, p. 735, 2020.

[22] S. WACHTER, B. MITTELSTADT et C. RUSSELL, "Why fairness cannot be automated : Bridging the gap between EU non-discrimination law and AI," *Computer Law & Security Review*, t. 41, 2021.

[23] Y. YOUSEFI, "Notions of Fairness in Automated Decision Making : An Interdisciplinary Approach to Open Issues," in *EGOVIS 2022, Proceedings*, Vienna, Austria : Springer, 2022, p. 3-17.