Introduction au dossier « Équité en apprentissage automatique »

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

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

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

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

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

1. bilel.benbouzid198@gmail.com

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

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

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

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

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

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

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

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

Références

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

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