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

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

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

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

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

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

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

1. bilel.benbouzid198@gmail.com

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

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

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

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

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

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

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

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

1.1 Formation du domaine du FairML

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

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

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

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

comme un problème de « valeur »[5].

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

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

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

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

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

1.2 Le biais comme norme de justice sociale

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

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

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

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

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

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

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

2. De la mesure des biais à leur interprétation

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

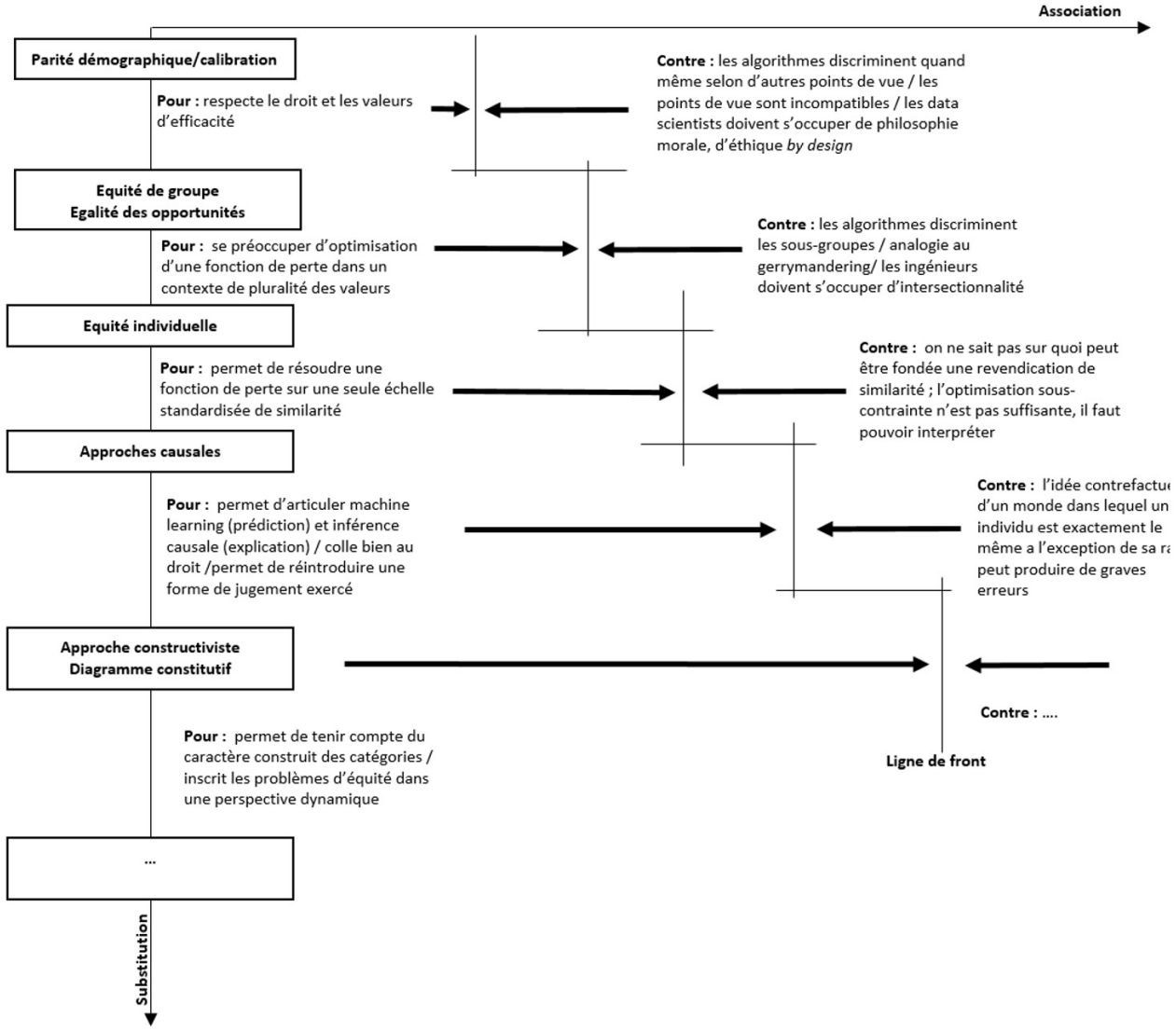

Figure 1 – *Diagramme dogmatique (non fidèle à la temporalité d'apparition des métriques) de la dynamique des controverses, inspiré de ceux de B. Latour (2014). Ce diagramme illustre les approches successives qui tendent à tourner le dos au réalisme métrologique.*

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

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

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

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

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

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

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

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

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

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

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

Les conventions imposées par le droit

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3. En guise de conclusion

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

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

Références

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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