La réception de la fairness algorithmique par le droit

Ronan Pons

LA RECEPTION DE LA *FAIRNESS* ALGORITHMIQUE PAR LE DROIT

Ronan PONS[1]

Préambule :

Résumé : Depuis plusieurs années, le mouvement interdisciplinaire visant à défendre la *fairness* algorithmique a gagné en popularité et en adeptes. En participant à cette lutte contre les biais algorithmiques, les juristes ont intégré ce concept au sein du droit, le plus souvent au travers de la notion de non-discrimination algorithmique. Désormais cette préoccupation se retrouve dans la proposition de règlement européen sur les systèmes d'IA. Ses dispositions prometteuses nécessitent encore des éclaircissements mais déjà de nombreux enjeux sont identifiables, notamment sur les modalités de satisfaction de ces obligations. Le texte européen souhaite évaluer les systèmes au travers de métriques statistiques, objets déjà existants dans la littérature scientifique. Cependant leur utilisation en tant que preuve du respect des exigences juridiques n'est pas sans risques. L'article aborde deux difficultés que sont la cohérence inter-métriques et leur adéquation au droit de la non-discrimination. Ces-dernières devront être résolues avant l'exploitation de ces métriques par le droit, sous peine de priver le futur système de certification européen de toute efficacité contre la discrimination algorithmique.

Mots-clés : discrimination algorithmique ; biais ; règlement sur les systèmes d'intelligence artificielle ; droit à la non-discrimination ; preuve juridique ; indicateurs de biais ; droit de l'intelligence artificielle ; interdisciplinarité

Titre anglais : Algorithmic fairness through the law

Abstract : Over the past years, the interdisciplinary movement called « Fair AI » has grown in popularity and members. Participating in this fight against algorithmic bias, legal experts have incorporate the concept into the law, often through the notion of algorithmic non-discrimination. Today this issue can be found in the European proposal for an artificial intelligence act. Its promising provisions still require clarification, but many issues can be identified already, especially on the means to comply with these obligations. The European text wants to evaluate the AI systems through statistical metrics, which exist in the scientific literature. However, their use as proof of compliance with legal requirements is not harmless. The paper addresses two difficulties, the coherence between metrics and their adequacy to non-discrimination law. Those problems must be solved before the use of these metrics by the law, otherwise the future European certification system will be ineffective against algorithmic discrimination.

Keywords : algorithmic discrimination ; bias ; AI Act ; non-discrimination law ; legal evidence ; fairness metrics ; artificial intelligence law ; interdisciplinary

[1] Doctorant en droit de l'Université Toulouse 1 Capitole, coordinateur de la chaire *Law, Accountability and IA* de l'institut ANITI (Toulouse) et membre de la chaire *Accountable AI in a Global Context* (Ottawa). Adresse mail : ronanpons@outlook.fr

1. INTRODUCTION

Depuis plusieurs années maintenant, l'image de neutralité et de perfection des technologies algorithmiques et de leurs décisions s'amenuise aux yeux des citoyens proches du numérique. Des affaires telles que COMPAS[2], Google Photos[3] ou encore Street Bump[4] ont fait voler en éclat le mythe d'objectivité pour le remplacer par la méfiance et l'esprit critique. Cette prise de conscience publique a coïncidé avec une explosion de la popularité du courant de recherche connu sous le nom de *« algorithmic fairness »*. Au départ purement informatique, les enjeux sociaux associés ont rapidement propagé la problématique au-delà des sciences dures et c'est aujourd'hui une recherche pluridisciplinaire qui est menée pour répondre aux questions de *fairness* dans les technologies de prise de décision.

La volonté de concevoir et garantir des systèmes d'IA *fair* est partagée par les institutions publiques au travers de nombreux documents de travail (rapports, résolutions, etc.). En Avril 2021, cette préoccupation a fait son apparition dans un texte règlementaire au niveau européen : la proposition de règlement sur les systèmes d'intelligence artificielle de la Commission européenne[5]. L'apport de cet article se veut alors double. Premièrement il s'agira d'analyser les dispositions pertinentes du texte pour l'efficacité de la lutte contre la discrimination algorithmique. Deuxièmement il s'agira d'aborder les enjeux juridiques dissimulés derrière la stratégie d'encadrement élaborée par la proposition européenne. Le système de certification des systèmes d'IA à haut risque créé un besoin de preuves nécessaires à l'obtention de cette certification. En matière de biais, le choix de ces éléments de certification recèle plusieurs défis juridiques qui peuvent impacter considérablement la lutte contre la discrimination algorithmique et *in fine* la notion même de discrimination en droit.

La présence d'une multitude de domaines de recherches au sein de la communauté *« algorithmic fairness »* ne présente pas que des avantages. Parmi les inconvénients de cette interdisciplinarité se trouve la difficulté de définir de façon univoque la notion de *fairness*. Chaque communauté au sein du mouvement possède sa ou ses propres significations pour ce terme. En droit, et encore plus en droit français, le terme de *fairness* s'avère difficile à manipuler de par son caractère polysémique. Il importe donc d'adresser dans un premier temps les difficultés de recevoir le terme interdisciplinaire de *fairness* en droit et de le traduire en des termes aux significations plus familières.

Une fois les problématiques de traduction abordées, il sera temps de s'intéresser à la proposition de règlement européen sur les systèmes d'IA. L'article se concentrera sur les

---

[2] Jeff Larson Julia Angwin, "Machine Bias", (23 May 2016), online: *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
[3] "Google apologises for Photos app's racist blunder", *BBC News* (1 July 2015), online: <https://www.bbc.com/news/technology-33347866>.
[4] Kate Crawford, "Think Again: Big Data", online: *Foreign Policy* <https://foreignpolicy.com/2013/05/10/think-again-big-data/>.
[5] *Règlement établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes législatifs de l'Union (proposition)*, 2021/0106.

dispositions spécifiques aux biais présents dans les systèmes d'IA. Enfin il sera temps de discuter des enjeux et défis juridiques dissimulés derrière la certification européenne des systèmes d'IA à haut risque proposé par la Commission européenne. Deux enjeux seront présentés : la cohérence des preuves de certification entre elles et leur adéquation avec le droit de la non-discrimination.

2. LA NOTION DE *FAIRNESS* EN DROIT

2.1 Les significations de *fairness* en droit : l'exemple du règlement général à la protection des données (RGPD).

Lors de l'apparition ou l'intégration d'un nouveau concept dans un domaine scientifique, définir ce concept clairement, sans aucune ambigüité avec le lexique déjà existant, est toujours une tâche difficile. La première difficulté rencontrée dans l'utilisation du mot *fairness* en droit provient de l'origine de la notion. Le terme de *fairness*, utilisé dans le contexte du mouvement d'*algorithmic fairness*, n'est pas issu de la communauté juridique. Il possède une sens propre qui diffère de celle accordée par d'autres communautés dans d'autres contextes. En droit, le terme de *fairness* possède déjà des significations spécifiques, ce qui peut amener à des confusions lors de son utilisation. Au sein même du droit du numérique, le terme de *fairness* est déjà vecteur de sens. Pour illustrer la polysémie du mot, rien ne vaut un exemple. Celui-ci est issu du droit du numérique, plus particulièrement du droit des données personnelles[6].

Le règlement général à la protection des données à caractère personnel[7] (RGPD ci-après), ou *General data protection regulation (GDPR)* dans sa version anglaise, est un règlement européen voté le 27 avril 2016 et entré en vigueur le 25 Mai 2018. Parmi ses 173 considérants et ses 99 articles, le terme de *fairness* ou *fair* est utilisé à plusieurs reprises pour renvoyer à des idées différentes. Au travers de cet unique texte, la *fairness* peut être interprétée de différentes manières[8]. Premièrement, la *fairness* peut renvoyer au concept de *fair balancing* qui signifie l'équilibre « entre les intérêts et la nécessité des finalités »[9] du traitement[10]. Deuxièmement, le concept de *fairness* est aussi utilisé au sens de la *procedural fairness*. Cette interprétation désigne les mesures pratiques à prendre pour améliorer l'objectif associé au mot *fair*[11], à savoir dans le RGPD l'objectif de transparence ou de licéité. Enfin la

---

[6] Le concept de *fairness* n'est pas apparu en droit des données personnelles avec le règlement européen à la protection des données. Voir Gianclaudio Malgieri, *The concept of fairness in the GDPR: a linguistic and contextual interpretation* (Barcelona, Spain: Association for Computing Machinery, 2020). P.2.

[7] *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE)*, OJ L 2016.

[8] Damian Clifford & Jef Ausloos, "Data Protection and the Role of Fairness" (2018) 37 Yearbook of European Law 130–187.

[9] Traduit de l'anglais : *« Fair balancing is based on proportionnality between interests and necessity of purposes […] »*. Malgieri, *supra* note 6.

[10] Cette signification de *fairness* se retrouve dans l'article 6.2 du RGPD.

[11] Malgieri, *supra* note 6.

fairness peut être lue comme un synonyme à la non-discrimination. Cette signification est utilisée particulièrement lorsqu'il est question de décisions algorithmiques[12]. La *fairness* est une notion anglophone vague dont il est impossible, et pas nécessairement souhaitable, de retirer une signification unique et transversale. La difficulté liée à ce flou notionnel est exacerbée lorsqu'il faut traduire *« fairness »* pour tenter de l'intégrer dans le droit français.

2.2 La traduction française de *fairness.*

Le terme de *fairness* est souvent traduit en français par « équité » ou « justice sociale »[13]. Ces traductions ont l'avantage de conserver le caractère flou du concept, une caractéristique bien pratique lorsque les discussions sont pluridisciplinaires, comme c'est le cas dans le domaine de l'*algorithmic fairness*. En effet, tout le monde a une idée de ce qui est socialement juste ou ce qui est équitable sans pour autant en avoir une définition précise. Ainsi une communauté pluridisciplinaire peut travailler ensemble à atteindre cet objectif sans être bloquée instantanément dans des débats sémantiques. Dans ce numéro, le choix a été fait de traduire *fairness* en équité, transformant alors l'expression *« algorithmique fairness »* en « équité algorithmique ». Comme dit précédemment, cette traduction est fréquente et ce n'est en aucun cas une erreur. Au contraire, le choix de traduction est en réalité tout à fait adéquat car il s'agit d'un numéro impliquant des contributions multidisciplinaires.

Toutefois, du point de vue du droit, la notion d'« équité » possède déjà ses propres significations. Ainsi le dictionnaire Cornu donne pas moins de six définitions de l'équité en droit. L'équité renvoie alors autant à la *« justice fondée sur l'égalité*[14] *»* qu'à une *« atténuation, modification apportées au Droit, à la loi, en considération de circonstances particulières*[15] *»*. Il peut s'agir aussi une *« manière de résoudre les litiges en dehors des règles de droit*[16] *»*. Tout comme la *fairness*, l'équité possède de nombreuses définitions. Peu importe le domaine du droit, les auteurs français sont d'accord sur la difficulté de définir cette notion précisément[17]. L'objectif dans cet article n'est pas de lister l'ensemble des significations du terme équitéPar conséquent ici, le choix sera fait d'exploiter le terme de *fairness* au travers d'un concept juridique plus clair et moins équivoque : la non-discrimination. Ainsi la problématique *« algorithmic fairness »* devient celle de la discrimination algorithmique.

---

[12] Le considérant 71 du RGPD explique que l'exigence de *fairness* dans les décisions automatisées nécessite que le responsable de traitement *« prévienne, entre autres, les effets discriminatoires à l'égard des personnes physiques fondées sur la l'origine raciale ou ethnique, les opinions politiques, la religion ou les convictions, l'appartenance syndicale, le statut génétique ou l'état de santé, ou l'orientation sexuelle, ou qui se traduisent par des mesures produisant un tel effet. ».*

[13] Ces termes sont ici exprimés dans leur sens commun.

[14] Gérard Cornu, *Vocabulaire juridique*, 11e édition ed (Paris: Presses Universitaires de France - PUF, 2016).

[15] *Ibid*.

[16] *Ibid*.

[17] Voir notamment Jean-Marc Sorel, "Équité" (2017) Répertoire de droit international 24. ; et Christophe Albiges, "Équité civile" (2017) Répertoire de droit civil Actualisé janvier 2019.

3. LE DROIT A LA NON-DISCRIMINATION ET L'INTELLIGENCE ARTIFICIELLE

3.1 Le régime juridique de la non-discrimination

Contrairement à son sens technique, la discrimination juridique n'est pas un synonyme de « différenciation » mais est un terme empreint d'une connotation négative[18]. Ainsi, la discrimination juridique se caractérise par trois éléments cumulatifs[19] :

- Un traitement moins favorable d'une personne ;
- Fondé sur au moins un critère précisé par la loi[20] (origine, état de santé, orientation sexuelle, identité de genre, etc.) ;
- Et qui correspond à une situation reconnue par la loi (embauche, accès au logement, à l'éducation, etc.).

Il s'agit ici de la conception restrictive de la non-discrimination présente dans les textes français en droit du travail et en droit pénal. Il existe aussi une conception plus extensive, utilisée notamment par le Conseil constitutionnel, le Conseil d'Etat et la Cour européenne des droits de l'Homme, considérant qu'est une discrimination toute atteinte non justifiée au principe d'égalité[21].

La discrimination juridique peut prendre deux formes : la discrimination directe et indirecte. La discrimination directe renvoie à la situation dans laquelle une personne est traitée moins favorablement en raison d'un des critères prohibés par la loi. Par exemple, un recruteur qui refuserait un candidat en raison de son orientation sexuelle, sa couleur de peau ou même sa situation financière commettrait une discrimination directe. Dans le cadre d'une décision automatisée, il s'agit du cas où les concepteurs décident d'insérer un critère discriminant dans les caractéristiques à prendre en compte par l'algorithme.

La discrimination indirecte, quant à elle, est plus subtile et se concentre sur le groupe et non sur l'individu[22]. Il s'agit de "*la situation dans laquelle une disposition, un critère ou une pratique apparemment neutre désavantagerait particulièrement des personnes par rapport à d'autres, pour des motifs prohibés [...]*"[23]. Les critères interdits légalement ne sont pas visibles explicitement dans la décision de l'auteur de la discrimination indirecte mais se retrouvent au travers des personnes défavorisées par la mesure. Cette définition offre au juge la capacité de regarder les conséquences réelles d'un acte sans être restreint par un *« l'apparente légalité du traitement »*[24]. Ces variables responsables d'une discrimination algorithmique sont appelées des *proxies* tandis que les variables interdites par le droit deviennent des *variables latentes*.

---

[18] Danièle Lochak, "La notion de discrimination" (2004) 48:1 Confluences Mediterranee 13–23 Bibliographie_available: 0Cairndomain: www.cairn.infoCite Par_available: 1publisher: L'Harmattan.
[19] https://www.defenseurdesdroits.fr/fr/institution/competences/lutte-contre-discriminations
[20] Au total, on compte vingt-cinq critères de discrimination dans la loi. Certains sont issus de la règlementation européenne ou internationale tandis que certains sont propres à la législation française.
[21] Lochak, *supra* note 18.
[22] Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law" (2018) 55 Common Market Law Review 1143-1186 (2018) (Common Market Law Review) 35.
[23] https://www.defenseurdesdroits.fr/fr/institution/competences/lutte-contre-discriminations
[24] "Lutte contre les discriminations", online: <https://www.senat.fr/rap/r07-252/r07-2523.html>.

De la même manière, la présence d'un biais dans les données d'entrainement pourrait transformer le système d'IA en *« pratique apparemment neutre »* entrainant un désavantage particulier[25]. Bien que la possibilité d'une discrimination algorithmique directe existe, c'est la discrimination algorithmique indirecte qui est la plus fréquente[26], en particulier dans les technologies d'apprentissage automatique.

3.2 La prise en compte du risque de discrimination par les décisions automatisées

Depuis 2016[27], les textes traitant de l'IA et soulevant le problème de la discrimination se multiplient. Ces initiatives sont élaborées par des acteurs très divers, allant des institutions publiques aux groupements privés en passant mêmes par des organisations internationales[28]. Pour ne citer que les textes les plus connus, on retrouve à l'international la déclaration de Montréal pour un développement responsable de l'intelligence artificielle[29] ainsi que celle de Toronto, plus spécifique au risque de discrimination algorithmique[30]. Au niveau de l'union européenne, le parlement européen a voté une résolution pour une IA éthique[31] tandis que la commission européenne a commandé plusieurs rapports sur la question de l'IA[32], aboutissant à un livre blanc[33]. Chacun de ces textes aborde à sa manière la question de la discrimination algorithmique. Certains traitent la question de façon centrale tandis que d'autres ne l'énumèrent que comme un des nombreux enjeux de l'intelligence artificielle. De même, la question de l'encadrement de l'intelligence artificielle est abordée sous différentes approches. Si le livre blanc ou la proposition de règlement européen aborde la question sous le prisme de la certification technique d'un produit à risque, d'autres préfèrent traiter le cas de l'intelligence artificielle au travers d'une approche par les droits fondamentaux. La France n'est pas non plus en reste sur ce sujet. Déjà en 2017 la Commission nationale de l'informatique et des libertés (CNIL ci-après) publiait un rapport sur les enjeux de l'intelligence artificielle dont une partie entière était dédié aux risques de biais et discriminations[34]. Pour lutter contre ce risque, le rapport Villani proposer la création d'

---

[25] Hacker, "Teaching Fairness to Artificial Intelligence", *supra* note 22.
[26] A titre d'exemple, voir "Pays-Bas. Scandale des allocations familiales : un avertissement qui montre l'urgence d'interdire les algorithmes racistes", (25 October 2021), online: *Amnesty International* <https://www.amnesty.org/fr/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/>.
[27] Jessica Fjeld et al, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI" (2020), online: <https://papers.ssrn.com/abstract=3518482>.
[28] *Projet de recommandation sur l'éthique de l'intelligence artificielle*, RAPPORT DE LA COMMISSION SCIENCES SOCIALES ET HUMAINES (SHS), by Unesco, 41 C/73 (2021).
[29] *Déclaration de Montréal pour un développement responsable de l'IA*.
[30] *The Toronto Declaration : Protecting the right to equality and non-discrimination in machine learning systems* (2018).
[31] *Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes*, 2020.
[32] *Discrimination, Artificial Intelligence and Algorithmic Decision-Making*, by Frederik Zuiderveen Borgesius, www.coe.int (Strasbourg: Conseil de l'Europe, 2018).
[33] *Livre blanc sur l'intelligence artificielle : Une approche européenne axée sur l'excellence et la confiance*, COM(2020) 65 final (Bruxelles: Commission européenne, 2020).
[34] *Comment permettre à l'Homme de garder la main ? Les enjeux éthiques de l'intelligence artificielle.*, Synthèse du débat public animé par la CNIL, Synthèse du débat public animé par la CNIL (CNIL, 2017). P. 31-34.

« études d'impact sur les risques discriminatoires »[35], reprenant ainsi la logique des études d'impact sur la vie privée inscrites dans le RGPD. Enfin l'on peut citer, sans pour autant être exhaustif, les rapports du Défenseur des droits : *« Algorithmes : prévenir l'automatisation des discriminations*[36] *»* et *« Technologies biométriques : l'impératif respect des droits fondamentaux*[37] *»*. C'est dans ce contexte de production massive de documents sur l'encadrement de l'IA qu'a été dévoilé par la Commission européenne la proposition de règlement européen sur les systèmes d'intelligence artificielle en Avril 2021[38].

4. L'ENCADREMENT DES BIAIS ALGORITHMIQUES PAR LA PROPOSITION EUROPENNE DE REGLEMENT SUR LES SYSTEMES D'IA.

4.1 Contexte et objectifs de la proposition européenne.

La proposition de règlement européen sur les systèmes d'IA s'intègre dans les orientations politiques pour 2019-2024 de la Commission européenne d'Ursula Von der Leyen[39]. Dans le préambule, la Commission européenne nous indique les objectifs poursuivis par son texte :

- *« Veiller à ce que les systèmes d'IA mis sur le marché de l'Union et utilisés soient sûrs et respectent la législation en vigueur en matière de droits fondamentaux et les valeurs de l'Union ;*
- *Garantir la sécurité juridique pour faciliter les investissements et l'innovation dans le domaine de l'IA ;*
- *Renforcer la gouvernance et l'application effective de la législation existante en matière de droits fondamentaux et des exigences de sécurité applicables aux systèmes d'IA;*
- *Faciliter le développement d'un marché unique pour des applications d'IA légales, sûres et dignes de confiance, et empêcher la fragmentation du marché. »*[40]

Ces quatre objectifs peuvent être regroupés en deux catégories : la volonté de développer un marché unique européen stable sur les systèmes d'intelligence artificielle et la protection des droits fondamentaux des européens. Ces deux axes se retrouvent plus explicitement dans la présentation des bases juridiques qui sous-tendent ce nouveau texte mais il se dégage alors une priorisation claire entre ces deux objectifs. En effet, la proposition européenne se fonde à la fois sur l'article 114 du Traité de fonctionnement de l'Union européenne (TFUE ci-après) et sur l'article 16 du TFUE. Il apparait alors clairement que l'article 114 relatif à l'établissement et au fonctionnement du marché

---

[35] *Donner un sens à l'intelligence artificielle : Pour une stratégie nationale et européenne*, by Cédric Villani (2018). P. 147.
[36] *Algorithmes : prévenir l'automatisation des discriminations* (Défenseur des Droits, 2020).
[37] *Technologies biométriques : l'impératif respect des droits fondamentaux* (Défenseur des droits, 2021).
[38] *Législation sur l'intelligence artificielle*, *supra* note 5.
[39] https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf
[40] *Législation sur l'intelligence artificielle*, *supra* note 5. P. 3

intérieur est le but premier du texte. En réglementant au niveau européen, la Commission espère éviter *« une fragmentation du marché intérieur »* et *« une diminution substantielle de la sécurité juridique tant pour les fournisseurs que pour les utilisateurs des systèmes d'IA*[41] *»*. L'article 16 relatif à la protection des données à caractère personnel, présenté brièvement en cinq petites lignes, semble apparaitre comme un petit bonus à la sécurité juridique du marché européen.

4.2 Le choix d'encadrement des biais algorithmiques dans les systèmes d'IA à haut risque par la certification.

Le règlement proposé par la Commission européenne aborde l'encadrement de l'intelligence artificielle par une approche technique de certification. Expliqué simplement, les fournisseurs systèmes d'IA dits à haut risque devront satisfaire aux exigences énoncées par le texte afin d'obtenir la certification obligatoire pour "accéder" au marché européen[42]. La proposition se concentre donc quasi essentiellement sur des exigences *ex ante* de conception, privilégiant ainsi le fournisseur et les utilisateurs des systèmes aux personnes affectées par les décisions.

Malgré l'apparente subsidiarité de l'objectif de protection des droits fondamentaux dans la proposition, certaines dispositions se révèlent pertinentes pour la lutte contre la discrimination algorithmique. Pour la première fois en Europe, le droit décide d'aborder explicitement la question des biais dans les outils d'intelligence artificielle. De façon regrettable, la proposition de règlement n'a pas recours au terme « discrimination » dans ses articles, ce-dernier étant limité aux considérants. Toutefois, biais et discrimination sont deux concepts liés sans pour autant être identiques. La présence de biais dans la conception du système d'IA peut entrainer des décisions discriminantes pour les individus. Par conséquent, l'encadrement des biais a nécessairement des effets sur la discrimination algorithmique.

4.3 Les dispositions relatives aux biais algorithmiques dans les données des systèmes d'IA

La proposition européenne de législation sur l'IA énonce donc des obligations relatives aux biais, en particulier ceux présents dans les données utilisées par les fournisseurs de ces systèmes d'IA à haut risque. L'article 10, intitulé *« Données et gouvernance des données »*, porte sur les jeux de données utilisés par les technologies d'apprentissage automatique[43]. Les exigences concernent les jeux de données d'entrainement, de validation et de test impliquées dans la conception des systèmes d'IA à haut risque. La Commission européenne impose plusieurs pratiques dont *« un examen permettant de repérer d'éventuel biais*[44] *»*. La

---

[41] *Ibid*. P. 7.

[42] *« Le présent règlement s'applique : a) aux fournisseurs, établis dans l'Union ou dans un pays tiers, qui mettent sur le marché ou mettent en service des systèmes d'IA dans l'Union; (b) aux utilisateurs de systèmes d'IA situés dans l'Union; (c) aux fournisseurs et aux utilisateurs de systèmes d'IA situés dans un pays tiers, lorsque les résultats générés par le système sont utilisés dans l'Union. »* - Article 2.1

[43] Pour reprendre les termes exacts du texte : *« les techniques qui impliquent l'entrainement de modèles aux moyens de données […]»*.

[44] Article 10.2 (f)

problématique des biais est ici explicitement inscrite dans la proposition. Cependant, ce n'est pas la seule mesure pertinente pour lutter contre les biais algorithmiques et, *in fine*, la production de décisions discriminantes. Le texte impose la mise en œuvre de *« pratiques appropriées »* concernant la collecte de données, les opérations de pré-traitement des données, l'adéquation de ces données à l'objectif poursuivi, les hypothèses pertinentes formulées par les concepteurs sur les données[45]. L'absence de précision ou d'exemples sur lesdites *pratiques appropriées* est regrettable. Toutefois, le texte met en lumière non seulement la présence de biais dans ces jeux de données mais aussi l'importance des opérations réalisées par l'équipe de conception sur ces données. La prise en compte de ces étapes est bienvenue car sont des vecteurs importants de création et reproduction des biais dans les données utilisées[46]. Toujours sur les trois types de jeux de données énoncés, l'article 10 exige qu'ils soient *« pertinents, représentatifs, exempts d'erreurs et complets »*. Pour satisfaire ces objectifs, la proposition dispose que :

> *« Ils possèdent les propriétés statistiques appropriées, y compris, le cas le cas échéant, en ce qui concerne les personnes ou groupes de personnes à l'égard desquels le système d'IA à haut risque est destiné à être utilisé. »*.

L'indication des *« propriétés statistiques »* renvoie à l'idée que des outils statistiques seront utilisés pour apporter les éléments de preuve du respect de ces exigences. En utilisant l'expression *« les personnes ou groupes de personnes »* affectées par le système d'IA, le texte fait écho avec les individus ou catégories d'individus protégées par le droit à la non-discrimination. En une phrase, le texte met en lien l'enjeu de la discrimination algorithmique et les éléments souhaités protéger les individus. Aujourd'hui les statistiques sont déjà utilisées dans la preuve d'une discrimination indirecte[47]. Cette idée est renforcée par la proposition européenne qui rappelle que la protection des libertés fondamentales, si elle n'est pas oubliée totalement, passera par des spécifications techniques des systèmes d'IA à haut risque. Ces exigences sur les jeux de données varieront en fonction de la destination du système d'IA à haut risque mais également en fonction de la zone géographique affectée par les résultats du système[48]. De nombreuses précisions et/ou modifications sont à attendre avant le vote définitif du texte mais il est rassurant de voir que la Commission européenne a bien pris en considération la place centrale des données dans la lutte contre la discrimination algorithmique.

4.4 L'autorisation de traitement des données sensibles pour lutter contre les biais

L'une des difficultés principales dans la lutte contre les discriminations et biais, qu'ils soient "traditionnels" ou algorithmiques, réside dans l'accès aux données de discriminations.

---

[45] Article 10.2 de la proposition de règlement

[46] Solon Barocas & Andrew D Selbst, "Big Data's Disparate Impact" (2016), online: <https://papers.ssrn.com/abstract=2477899> et Hacker, "Teaching Fairness to Artificial Intelligence", *supra* note 22.

[47] Hacker, "Teaching Fairness to Artificial Intelligence", *supra* note 22.

[48] *« Les jeux de données d'entraînement, de validation et de test tiennent compte, dans la mesure requise par la destination, des caractéristiques ou éléments propres au contexte géographique, comportemental ou fonctionnel spécifique dans lequel le système d'IA à haut risque est destiné à être utilisé »* - Article 10 alinéa 4.

Comment détecter une discrimination envers les femmes s'il est impossible d'obtenir les informations sur le sexe des individus sujets à la décision ? Au sein de la communauté *Fair AI*, des outils ont été développés afin d'identifier les biais dans la conception des systèmes d'IA sans pour autant avoir accès à la variable dite "sensible". Malgré les efforts des chercheurs, ces outils ne sont pas aussi performants dans leur mission que ceux ayant accès à ces variables interdites[49]. L'article 10.5 du règlement vient offrir une piste de solution à cet épineux problème de détection de la discrimination :

> *« Dans la mesure où cela est strictement nécessaire aux fins de la surveillance, de la détection et de la correction des biais en ce qui concerne les systèmes d'IA à haut risque, les fournisseurs de ces systèmes peuvent traiter des catégories particulières de données à caractère personnel visées à l'article 9, paragraphe 1, du règlement (UE) 2016/679, à l'article 10 de la directive (UE) 2016/680 et à l'article 10, paragraphe 1, du règlement (UE) 2018/1725, sous réserve de garanties appropriées pour les droits et libertés fondamentaux des personnes physiques, y compris des limitations techniques relatives à la réutilisation ainsi que l'utilisation des mesures les plus avancées en matière de sécurité et de protection de la vie privée, telles que la pseudonymisation, ou le cryptage lorsque l'anonymisation peut avoir une incidence significative sur l'objectif poursuivi. »*

Cette disposition offre aux fournisseurs de systèmes d'IA une exception aux règlementations européennes existantes sur les données qui interdisent le traitement de catégories particulières de données. Ces données, aussi appelées « données sensibles », sont en principe interdites à traiter notamment en raison du fort risque de discrimination associé. Le RGPD offre une liste des données sensibles à son article 9, il s'agit entre autres de *« l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques […] »*.

Cette disposition est bienvenue car elle vient répondre aux besoins des acteurs de lutte contre les biais et discriminations algorithmiques. Toutefois cette possibilité de traiter ces données crée un vrai risque sur les droits et libertés des personnes concernées. Ainsi il ne s'agit pas d'une solution universelle utilisable à volonté.

Premièrement cet article 10.5 nécessite une condition pour être exploité : *Dans la mesure où cela est strictement nécessaire aux fins de la surveillance, de la détection et de la correction des biais en ce qui concerne les systèmes d'IA à haut risque […] »*[50]. En droit des données personnelles, le concept de nécessité est représenté par le principe de minimisation des données. Cela signifie qu'est considéré comme nécessaire le traitement de données si et seulement s'il n'existe pas d'alternative aussi efficace dans

[49] *Reconciling Legal and Technical Approaches to Algorithmic Bias*, SSRN Scholarly Paper, by Alice Xiang, papers.ssrn.com, SSRN Scholarly Paper ID 3650635 (Rochester, NY: Social Science Research Network, 2021).
[50] Emphase ajoutée par l'auteur.

l'atteinte de la finalité mais qui nécessiterait moins de données personnelles[51]. Ainsi il incombera aux fournisseurs de systèmes d'IA d'attester de l'absence d'alternatives moins intrusives (pour les droits et libertés des personnes concernées) et aussi performantes dans la détection de biais algorithmiques. L'efficacité des techniques de débiaisage ne nécessitant pas l'accès aux données sensibles déterminera la possibilité de récolter lesdites données sensibles. Également, le critère de nécessité porte sur le recours aux données sensibles mais aussi à la quantité de données sensibles nécessaires au débiaisage des données d'entrainement. Ce n'est pas parce que la situation nécessite le recours à cet article 10.5 que le fournisseur de système d'IA aura le droit de traiter librement des données sensibles, chacune de ces données récoltées devra être justifié au regard de la finalité de détection et correction des biais.

Conscient des risques pour les personnes concernées, le texte européen impose des *« garanties appropriées pour les droits et libertés fondamentaux des personnes physiques […]. »*. Ces garanties appropriées sont déjà exigées dans le cadre juridique européen sur les données[52] et se pose alors la question de l'articulation entre ces textes et le règlement sur l'IA. Est-ce que le niveau de protection exigé entre ces différents textes est équivalent ? Ou est-ce que les mesures exigées dans la proposition européenne seront plus élevées que celles demandées par les autres textes ? Dans un avis joint du comité européen de protection des données et du contrôleur européen à la protection des données[53], ces-derniers soulèvent le besoin de clarté dans l'articulation des textes, notamment sur les mesures de protection à mettre en place autour de ce traitement de données sensibles.

Enfin cette disposition souffre également d'une limite. Les données sensibles utilisables au travers de cet article sont au nombre de neuf. Si un fournisseur de système d'IA a recours à cette disposition, il peut en théorie détecter des biais sur neuf types de discriminations possibles. Or le droit européen liste quatorze critères discriminants tandis que le droit français monte au-delà de vingt-cinq critères[54]. L'article 10.5 n'est donc qu'une avancée partielle dans la lutte contre la discrimination algorithmique. Il est essentiel que les législateurs, juges et fournisseurs d'accès aux systèmes d'IA soient conscient de cet aspect lacunaire de la disposition. Sinon l'entrée en vigueur du texte entrainerait la relégation au second plan, voir l'oubli total, des discriminations algorithmiques qui ne seraient pas liées à une donnée sensible.

4.5 Des obligations juridiques satisfaites par des tests techniques.

[51] Pour illustrer le principe de minimisation du RGPD et l'idée de « nécessité », voir l'avis de la CNIL sur l'utilisation de la reconnaissance faciale dans deux lycées français : https://www.cnil.fr/fr/experimentation-de-la-reconnaissance-faciale-dans-deux-lycees-la-cnil-precise-sa-position

[52] Les textes en question sont directement référencés dans l'article 10.5 de la proposition de règlement.

[53] *Avis conjoint 05/2021 de l'EDPB et du CEPD sur la proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle)* (2021). P. 23.

[54] note 34 ou https://www.defenseurdesdroits.fr/fr/institution/competences/lutte-contre-discriminations

Au-delà du contenu des obligations, la technicité de l'encadrement des systèmes d'IA à haut risque passe aussi par les modalités de satisfaction à ces exigences. Ces systèmes d'IA devront être testés afin de garantir que *« les systèmes d'IA à haut risque de façon cohérente à leur destination et qu'ils sont conformes aux exigences […] »*. La réussite à ces tests détermine leur conformité aux obligations juridiques du règlement. A ce propos, l'article 9.7 de la proposition européenne indique que *« les tests seront effectués sur la base de métriques et de seuils probabilistes préalablement définis, qui sont adaptés à la destination du système d'IA à haut risque »*. La procédure d'évaluation pour la certification, qu'elle soit effectuée en interne par ou par un organisme externe, portera donc aussi bien sur le respect des exigences organisationnelles[55] que la réussite du système aux tests techniques.

Ces outils techniques de test ouvrent une question attendue par de nombreux concepteurs : quelles sont les objectifs à atteindre pour obtenir la certification nécessaire à l'utilisation du système d'IA dans le marché européen ? La Commission européenne crée ainsi un besoin, celui de nouvelles preuves pour la certification des systèmes d'IA.

5 LES MODES DE PREUVES D'ABSENCE DE BIAIS DANS LE FUTUR DROIT EUROPEEN DE L'IA.

5.1 Les indicateurs de *fairness* existants dans la littérature scientifique

La communauté scientifique n'a pas attendu le texte de la Commission européenne pour élaborer des indicateurs de biais algorithmiques. Aujourd'hui la littérature scientifique a établi une multitude d'indicateurs pour détecter la présence de biais dans un système de prise de décision automatisée[56]. Ces métriques poursuivent toutes le même objectif : celui de détecter des biais. Cependant chacune de ces méthodes le fait de d'une manière qui lui est propre. Les différences peuvent se trouver dans les moyens de contrôler et/ou dans l'objet à contrôler, à savoir la définition du concept de biais. Parmi ces indicateurs, deux sont particulièrement populaires : la parité statistique ou *disparate impact* et l'égalité du taux d'erreur ou *equality of odds*.

La parité statistique consiste à comparer la probabilité d'acceptation d'un sous-groupe de population par rapport à un autre. En d'autres termes, il s'agit d'observer si une catégorie de population (définie selon certains critères) obtient une proportion de réponses positives (ou négatives) égales à celles d'une autre catégorie de population. Pour illustrer cela, prenons un exemple simplifié de l'utilisation de cet indicateur pour contrôler les biais entre hommes et femmes d'un algorithme servant à accorder des prêts bancaires. Sur les femmes qui ont fait une demande de prêt, combien se sont vues obtenir ledit prêt ? Ensuite la même question est posée pour l'autre catégorie de population, les hommes. Enfin on compare les proportions d'accord de prêt dans les deux catégories. Si les femmes n'obtiennent pas le même taux de

[55] A titre d'exemples : l'obligation de mettre en place un système de gestion de la qualité (article 17), d'établir une documentation technique (article 18).

[56] Parmi d'autres : Alexandra Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments" (2016) arXiv:161007524 [cs, stat], online: <http://arxiv.org/abs/1610.07524> arXiv: 1610.07524.

réponse positives que les hommes, il y a un biais dans le système de décision qui aboutit à une situation préjudiciable. Seuls les résultats sont pris en compte. Toutefois, il n'est pas nécessaire d'obtenir une égalité parfaite entre les deux groupes pour rester dans le cadre de la loi. Ainsi aux Etats-Unis, la Commission pour l'égalité des chances en matière d'emploi[57] propose depuis longtemps la règle des 4/5 comme une présomption de *disparate impact*. Cette règle énonce que le groupe de population le plus défavorisé dans le processus de décision doit au moins avoir un taux d'acceptation égal à 80% du taux d'acceptation du groupe le plus favorisé[58]. Le taux à partir duquel une situation d'illégalité peut être caractérisé ou présumée peut être défini différemment selon le choix politique du pays en question.

L'égalité du taux d'erreur correspond à comparer non pas le taux de décision positives mais le taux d'erreur entre deux catégories de population. Ainsi il est question d'observer si le système d'IA se trompe plus souvent pour les femmes que pour les hommes (ou inversement) pour reprendre notre exemple. La question posée par cet indicateur est la suivante : Parmi toutes les femmes qui méritaient un prêt, combien l'ont obtenu ? C'est cet indicateur qui a été utilisé par l'équipe de ProPublica pour identifier le biais raciste du logiciel étatsunien COMPAS[59]. Ils ont ainsi prouvé que si commettait bien plus d'erreurs défavorables[60] pour les afro-américains que pour les blancs. Et inversement, la population blanche faisait l'objet de plus d'erreur favorables que la population afro-américaine[61]. Cet indicateur est plus difficile à obtenir que le *disparate impact* car il impose de posséder une réalité de terrain afin de savoir si les individus méritaient ou non la décision rendue par le système d'IA. Une inégalité du taux d'erreur apparait notamment quand une catégorie de population est sous-représentée dans les données d'entrainement, ce qui cause une plus faible fiabilité des résultats produits à l'encontre des individus constituant cette catégorie[62].

5.2 La cohérence des indicateurs de biais dans la certification des systèmes d'IA.

La parité statistique et l'égalité du taux d'erreur font partie de deux grandes familles d'indicateurs élaborées par Wachter et al. [63] en 2021. La parité statistique appartient à la famille des indicateurs modifiant les biais. Autrement dit, les indicateurs comme celui-ci vont imposer des biais aux systèmes d'IA au détriment des biais existants au départ. Ainsi en utilisant le *disparate impact*, on exige du système d'IA qu'il produise une proportion équivalente de décisions positives (ou négatives) entre les différentes catégories de personnes.

---

[57] *US Equal Employment Opportunity Commission* : https://www.eeoc.gov/
[58] Barocas & Selbst, *supra* note 46.
[59] Jeff Larson Mattu Julia Angwin,Lauren Kirchner,Surya, "How We Analyzed the COMPAS Recidivism Algorithm", online: *ProPublica* <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=dvecePGfOFm-2Axng71w4PbbV6wY27tC>.
[60] Le taux d'erreur défavorable correspond au taux de faux positifs, à savoir ceux obtenant un score de récidive élevé alors qu'ils « méritaient » un score faible
[61] Le taux d'erreur favorable correspond au taux de faux négatifs, à savoir ceux obtenant un score de récidive alors qu'ils « méritaient » un score élevé
[62] Pour un exemple sur la reconnaissance faciale, voir : Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (PMLR, 2018).
[63] *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law*, SSRN Scholarly Paper, by Sandra Wachter, Brent Mittelstadt & Chris Russell, papers.ssrn.com, SSRN Scholarly Paper ID 3792772 (Rochester, NY: Social Science Research Network, 2021).

Ces indicateurs ne se soucient pas de la qualité des prédictions, seuls les résultats sont pris en compte. Pour satisfaire cette exigence, le système d'IA peut par exemple accepter des personnes, appartenant à un groupe social défavorisé, qui ne "méritaient" pas un prêt afin de rééquilibrer les proportions d'acceptation avec la catégorie de personne servant de comparaison. L'objectif est ici de produire une égalité ou une quasi-égalité de résultats en "transformant les biais sociaux" en biais dans le taux d'erreur[64]. L'égalité du taux d'erreur, quant à lui, appartient à la famille des indicateurs qui conservent les biais[65], à savoir les biais présents dans la société. En se concentrant sur la précision du système, les inégalités sociales déjà existantes seront reproduites par le système d'IA. Il n'y a pas d'approche fondamentalement meilleure que l'autre, elles représentent toutes les deux une conception de la *fairness* différente.

Cette opposition apparente la parité statistique et l'égalité du taux d'erreur est un phénomène appelé le « dilemme des métriques » qui se vérifie techniquement[66]. En effet, il n'est pas possible pour un système d'IA de réussir un test fondé sur la parité statistique et un test fondé sur l'égalité du taux d'erreur. Réussir l'un signifie échouer l'autre dès lors qu'il existe des inégalités entre les groupes de populations[67]. Par conséquent ces deux indicateurs ne doivent pas être utilisés simultanément pour certifier un système comme non biaisé au sens du droit européen. Pour rappel, la proposition de règlement veut éviter *« une fragmentation du marché intérieur »*. Dans le contexte de la lutte contre les biais algorithmiques, accepter les deux indicateurs malgré leur incompatibilité irait à l'opposé de cet objectif. Ce choix peut dépendre de plusieurs facteurs, notamment la destination du système d'IA à haut risque[68]. Dans ce sens, il a été proposé qu'utiliser un indicateur conservant les biais dans un domaine d'activité notoirement inégalitaire pourrait se voir qualifier de discrimination indirecte[69]. Proposer une liste de critères d'aide au choix entre les différents indicateurs existants serait un premier pas judicieux de la part de la Commission européenne. L'Union européenne doit se saisir de cette question sous peine d'offrir une certification dépourvue de cohérence et sens quant à la protection des libertés fondamentales des personnes affectées par les décisions algorithmiques.

5.3 L'adéquation de la métrique avec le droit de la non-discrimination.

Comme énoncé précédemment, le besoin d'indicateur apporté par la proposition de règlement trouve des pistes de solutions dans les indicateurs de *fairness* de la littérature scientifique. Ils ont été conçus pour détecter des biais et non des discriminations au sens juridique du terme, ce qui en fait des outils certes utiles mais à manier avec précaution.

---

[64] Pour pluse d'informations sur l'utilisation du *disparate impact* pour corriger les biais dans les données, voir : Philippe Besse et al, "A Survey of Bias in Machine Learning Through the Prism of Statistical Parity" (2021) 0:0 The American Statistician 1–11.
[65] Wachter, Mittelstadt & Russell, *supra* note 63.
[66] Chouldechova, "Fair prediction with disparate impact", *supra* note 56.
[67] *Ibid*.
[68] Cette condition est exprimée à l'article 9.7 de la proposition de règlement. Voir partie 4.5.
[69] Wachter, Mittelstadt & Russell, *supra* note 63.

Jusqu'à présent, les notions de biais, *fairness* et discrimination ont pu sembler similaires pour les non juristes. Il est vrai que ces notions sont liées entre elles mais elles ne sont pas pour autant identiques. Pour qualifier une situation de discrimination au sens du droit, il existe plusieurs conditions listées préalablement[70]. Un biais est un phénomène plus général qu'une discrimination juridique. Premièrement, un biais peut exister sans pour autant concerner des individus. Tout système informatique peut se retrouver biaisé, aussi bien un système de détections d'anomalies électroniques qu'un système dans un prêt bancaire. Seulement dans le premier cas, la présence de biais n'impacte aucune personne, il ne peut donc pas y avoir de discrimination juridique. Deuxièmement un biais peut impacter une personne sans pour autant être lié, directement ou indirectement, à un critère prohibé par la loi. Cette liste de critères est établie par le droit. Un système d'IA à haut risque biaisé sur un critère n'ayant aucun lien avec cette liste ne constituera donc pas une discrimination. Enfin une discrimination nécessite des conséquences sur les individus, un *traitement défavorable* ou un *désavantage particulier*. Ainsi si un système d'IA est biaisé, ce sont les décisions finales qui seront qualifiées de discriminatoires. Par conséquent, un résultat biaisé n'aura pas automatiquement pour conséquence de discriminer la personne faisant l'objet du processus décisionnel. Par exemple, si le système d'IA produit un résultat biaisé qui vient s'ajouter à une multitude d'autres élément de décision qu'un opérateur humain prend en compte, le risque de rendre des décisions systématiquement discriminantes existe mais n'est pas certain. Toutefois, si la décision est prise sur le fondement seul ou exclusivement par le système d'IA à haut risque, le biais dans le résultat se retrouvera dans la décision finale. Encadrer les biais techniques des systèmes d'IA à haut risque permet de prévenir partiellement la production de décisions discriminatoires.

Les outils développés par la communauté scientifique n'ont pas été développés pour les besoins du droit à la non-discrimination. Les différentes conditions contenues dans la définition juridique de la discrimination ne sont pas prises en compte dans ces indicateurs. Comme nous l'avons vu, les indicateurs peuvent représenter différentes conceptions de *fairness*. Il est donc important de réfléchir également aux métriques qui représentent le mieux la conception juridique de non-discrimination. En effet l'objectif de détection et de correction des biais algorithmiques a pour objectif d'éviter au maximum le rendu de décisions discriminantes. Il est donc essentiel de créer un réel échange entre les techniciens et les juristes afin de déterminer le ou les indicateurs les plus proches de notre conception de la non-discrimination. Sinon, la future règlementation se voit courir le risque de certifier des systèmes d'IA à haut comme non-biaisé alors qu'ils rendront des décisions discriminatoires aux yeux du droit. Par exemple, la plupart des indicateurs de *fairness* existants ne sont pas compatibles avec la conception de non-discrimination défendue par la Cour de justice de l'Union européenne car ils conservent les biais existants au lieu de tenter de les réduire[71].Au contraire, l'indicateur nommée *Conditional Demographic Disparity* (une variante de la parité statistique) semble être la métrique la plus adaptée[72]. Toutefois, la conception de non-discrimination peut varier entre les législations. Si de nombreux critères prohibés sont

[70] Voir partie 3.1.
[71] Wachter, Mittelstadt & Russell, *supra* note 63.
[72] *Ibid*.

similaires au sein de l'Union européenne, la liste complète peut différer d'un pays à l'autre. *In fine*, le choix de l'indicateur de biais comme élément de certification des systèmes d'IA à haut risque revient à mettre en avant une conception de la non-discrimination plutôt qu'une autre. Le choix d'indicateur devrait donc être réalisé par des institutions légitimes pour réaliser une telle tâche. Laisser librement les fournisseurs de systèmes d'IA faire ce choix sans aide ou contrainte revient à leur déléguer la tâche d'interprétation du principe de la non-discrimination lorsqu'appliqué aux systèmes d'IA. Cet enjeu concerne tous les systèmes d'IA qui ne sont pas à haut risques, car même les systèmes d'IA exclus du règlement sont soumis au droit de la non-discrimination.

CONCLUSION

La problématique de la *fairness* dans les technologies d'IA a fait du chemin depuis les premières chartes éthiques. La lutte contre les biais algorithmiques est présente dans de nombreuses disciplines. En droit, la question est principalement abordée au travers de la notion juridique moins équivoque de non-discrimination. Cet enjeu se retrouve désormais dans les dispositions de la proposition de règlement européen sur les systèmes d'IA d'avril 2021. Le texte s'intéresse notamment à la gestion des jeux de données utilisés pour la conception des systèmes d'apprentissage automatique à haut risque. Il impose premièrement aux fournisseurs de ces outils d'examiner la présence de biais dans ces données. Secondement, la proposition européenne met la lumière l'importance des étapes de conception des systèmes d'IA, sources bien connues de biais. L'objectif est clair, les jeux de données doivent être « *pertinents, représentatifs, exempts d'erreurs et complets.* ». Pour cela, le texte ouvre la possibilité de recourir à des données sensibles afin de faciliter la détection et la correction des biais présents. Afin de vérifier le respect à ces obligations, le règlement fait référence à plusieurs reprises à des tests statistiques. La certification, indispensable à la commercialisation des systèmes d'IA à haut risque sur le marché européen, sera obtenue notamment si les systèmes réussissent ces tests. Par cette approche résolument technique, la Commission européenne pose la question des éléments nécessaires à la certification sans toutefois y répondre. Du côté des obligations relatives aux biais, des indicateurs existent déjà dans la communauté scientifique. Leur utilisation en tant que mode de preuve du respect aux exigences européennes n'est pas anodine. Ces métriques statistiques recèlent deux enjeux juridiques importants qui doivent être réglés avant d'être exploitées par droit. En effet, certains de ces indicateurs de biais sont incompatibles entre eux, ce qui les empêchent d'être utilisés simultanément. Il n'est donc pas possible d'accepter ces indicateurs pour un même système d'IA sous peine de rendre le système de certification totalement incohérent. Enfin, ces indicateurs ont pour objectif indirect de réduire au maximum le risque de discrimination algorithmique. Ils doivent donc au moins mesurer les biais discriminatoires, ce qui n'est pas le cas de tous actuellement car biais et discrimination sont deux concepts proches mais différents. Sans cette adéquation entre les biais mesurés et la conception juridique de non-discrimination, le marché européen risquera d'ouvrir ses portes à des systèmes d'IA certifiés non-biaisés mais qui en réalité produiront des résultats discriminants.