Equité et explicabilité des algorithmes d'apprentissage automatique : un défi technique et juridique

Thierry Kirat, Olivia Tambou, Virginie Do, Alexis Tsoukias

Equité et explicabilité des algorithmes d'apprentissage automatique : un défi technique et juridique[1]

Thierry Kirat (IRISSO, Université Paris Dauphine-PSL CNRS-INRAE, UMR 7170), **Tambou** (CR2D, Université Paris Dauphine-PSL), **Virginie Do** (LAMSADE, Université Paris Dauphine-PSL, CNRS, UMR 7243), **Alexis Tsoukiàs** (LAMSADE, Université Paris Dauphine-PSL, CNRS, UMR 7243)

Version 1.0 – 6 mai 2022

RESUME

L'article propose une contribution aux constructions interdisciplinaires de l'analyse des enjeux d'équité dans les décisions algorithmiques automatiques. La section 1 montre que les choix techniques en apprentissage supervisé ont des implications sociales dont il faut prendre la mesure. La section 2 propose une approche contextuelle de la question de la discrimination de groupe non intentionnelle, c'est-à-dire de règles de décision facialement neutres mais qui génèrent des impacts disproportionnés selon les groupes sociaux (selon les cas : genrés, raciaux ou ethniques). La contextualisation portera sur les systèmes juridiques des États-Unis d'un côté, de l'Europe d'un autre côté. En particulier, la législation et la jurisprudence tendent à promouvoir des critères d'équité différents de part et d'autre de l'Atlantique. La section 3 est consacrée à l'explicabilité des décisions algorithmiques ; elle confrontera et tentera de croiser les concepts juridiques (en droit européen et en droit français) avec les concepts techniques et mettra en exergue la pluralité, voire la polysémie, des textes juridiques européens et français relatifs à l'explicabilité des décisions algorithmiques. La conclusion propose des orientations pour la recherche.

ABSTRACT

The paper offers a contribution to the interdisciplinary constructs of analyzing fairness issues in automatic algorithmic decisions. Section 1 shows that technical choices in supervised learning have social implications that need to be considered. Section 2 proposes a contextual approach to the issue of unintended group discrimination, i.e. decision rules that are facially neutral but generate disproportionate impacts across social groups (e.g., gender, race or ethnicity). The contextualization will focus on the legal systems of the United States on the one hand and Europe on the other. In particular, legislation and case law tend to promote different standards of fairness on both sides of the Atlantic. Section 3 is devoted to the explainabilityof algorithmic decisions; it will confront and attempt to cross-reference legal concepts (in European and French law) with technical concepts and will highlight the plurality, even polysemy, of European and French legal texts relating to the explicability of algorithmic decisions. The conclusion proposes directions for further research.

Introduction

Ce qu'on appelle en anglais la *fairness* des décisions algorithmiques fondées sur l'apprentissage machine, est une question fondamentale de la recherche en *computer science* depuis plus de deux décennies ; plus récemment, elle a été l'objet de recherches de plus en plus pluridisciplinaires, auxquelles contribuent des spécialistes du droit et des *social scientists*. Nous développerons ce point dans le texte, mais pour l'instant il est important de souligner que ce qui était au départ un enjeu sociétal traité de manière technique, sans besoin de regards extérieurs, est devenu l'objet d'une collaboration que l'on peut qualifier de fructueuse, entre *computer scientists* et chercheurs en droit et en analyse de politiques publiques.

[1] Ce projet a obtenu le soutien financier du CNRS à travers les programmes interdisciplinaires de la MITI.

Dans les publications en langue française, le mot « *fairness* » est souvent, mais pas toujours, traduit par « loyauté » (par ex. Besse & al., 2018, ou encore l'article 5 du RGPD), mais l'adjectif « équitable » » n'est jamais très loin du mot « loyauté ». Le mot « *fairness* » est également souvent accolé au vocable des biais, de la discrimination, mais aussi de la problématique d'une IA « digne de confiance ». Or, dans les théories de la justice sociale, la« *fairness* » est traduite en français par « équité » ; par exemple, la théorie rawlsienne de la « *justice as fairness* » a pu être traduite, dans les versions françaises des écrits de John Rawls, par « La justice comme équité ». Afin de ne pas nous écarter des usages courants et du fait que les dictionnaires de référence français-anglais proposent « équité » comme traduction de la « *fairness* », nous en tiendrons ici à la « fairness comme équité ». Au-delà d'une convention de traduction, des considérations de fond renforcent notre choix. Ainsi, par exemple, dans le domaine de la santé, l'accès aux soins sera considéré comme équitable, c'est-à-dire socialement juste horizontalement, donc sans inégalités entre individus selon leur position sociale, si cet accès est indépendant d'autres facteurs que l'état de santé de l'individu (Rochaix & Tubeuf, 2009).

Cet article s'intéresse aux conditions d'une équité horizontale entre les individus concernés par des décisions automatisées basées sur des algorithmes d'apprentissage. Sans méconnaître les possibilités techniques de réalisation d'algorithmes d'apprentissage-machine équitables (Barocas & al, 2019 ; Chouldechova, 2017 ; Corbett-Davies & Goel, 2018, Kleinberg et al, 2016, nous soutenons que l'objectif d'équité :

a) Ne peut être limité à une question exclusivement technique, c'est-à-dire de design des algorithmes et des modèles sans référence à d'autres dimensions,
b) Suppose que des choix sociaux soient faits pour définir la forme d'équité souhaitable parmi un ensemble de métriques possibles,
c) Doit être considéré au regard des dispositifs juridiques en vigueur, qui incorporent ces choix sociaux, et varient dans le temps et dans l'espace,
d) Peut supposer la réalisation de compromis entre contraintes techniques et contraintes juridiques.

L'article se veut une contribution aux constructions interdisciplinaires de l'analyse des enjeux d'équité dans les décisions algorithmiques. Il est clair que la conception d'algorithmes équitables n'est pas une opération aisée, pour les raisons évoquées plus haut. La section 1 montre que les choix techniques en apprentissage supervisé ont des implications sociales dont il faut prendre la mesure. La section 2 propose une approche contextuelle de la question de la discrimination de groupe non intentionnelle, c'est-à-dire de règles de décision en apparence neutres mais qui génèrent des impacts disproportionnés selon les groupes sociaux (selon les cas : genrés, raciaux ou ethniques). La contextualisation portera sur les systèmes juridiques des Etats-Unis d'un côté, de l'Europe d'un autre côté. En particulier, la législation et la jurisprudence tendent à promouvoir des critères d'équité différents de part et d'autre de l'Atlantique. La section 3 sera consacrée à l'explicabilité des décisions algorithmiques ; elle confrontera et tentera de croiser les concepts juridiques (en droit européen et en droit français) avec les concepts techniques et mettra en exergue la pluralité, voire la polysémie, des textes juridiques européens et français relatifs à l'explicabilité des décisions algorithmiques. La conclusion proposera des orientations pour la recherche.

1. L'équité en apprentissage machine: les implications sociales des choix techniques

Les algorithmes d'apprentissage machine (Machine Learning, ML) sont de plus en plus utilisés pour prendre ou aider des décisions qui affectent la vie des gens, dans des applications telles que les prêts bancaires, la tarification, le recrutement, la justice pénale et le diagnostic médical. Or, les modèles d'algorithmes reposent sur des choix techniques qui ont des implications sociales. Nous démontrerons cet aspect dans l'analyse de l'apprentissage supervisé.

1.1 Apprentissage supervisé

Dans cet article, nous nous concentrons sur la tâche dominante de l'apprentissage supervisé, dans lequel les algorithmes apprennent à prédire une variable de résultat y à partir de variables d'entrée x. Ces variables de résultat peuvent être quantitatives (par exemple, prédire le prix d'un logement en fonction de son emplacement et de sa surface) ou qualitatives (par exemple, prédire si un demandeur de prêt sera solvable ou non, en fonction de caractéristiques telles que l'historique des dettes et la profession). L'apprentissage supervisé consiste à apprendre une règle de prédiction h, à partir d'un échantillon de données étiquetées $S = \{(x_i, y_i)\}_{i=1}^{n}$ appelé données d'entraînement, de manière à prédire pour chaque x, un résultat y.

La relation entre les entrées x et les sorties y est caractérisée par une fonction inconnue. L'algorithme doit choisir h, parmi une famille donnée de fonctions (la classe d'hypothèses $\mathcal{H}$), qui approxime le mieux la fonction inconnue sur les exemples d'entraînement S. La principale approche pour choisir h est la minimisation du risque empirique (ou *empirical risk minimization*, ERM). L'algorithme est doté d'une fonction de perte $l(h(x), y)$qui quantifie la différence entre la prédiction $h(x)$ et le vrai résultat y. Il est courant d'utiliser l'ERM pour déterminer h en minimisant le risque empirique $\hat{R}(h)$, c'est-à-dire la perte moyenne sur l'ensemble de données d'apprentissage S:

$$\min_{h \in \mathcal{H}} \left\{ \hat{R}(h) := \sum_{i=1}^{n} l(h(x_i), y_i) \right\},$$

où $\mathcal{H}$ est la classe d'hypothèses, i.e., la classe de modèles d'apprentissage automatique considérée, qui peut inclure des modèles plus ou moins complexes tels que les arbres de décision ou les SVM.

1.2 L'impact des choix sous-jacents en apprentissage supervisé

La pratique courante de l'ERM pour l'apprentissage machine supervisé décrite plus haut implique de nombreux choix qui ont un impact sur les prédictions et les décisions qui en découlent :

- Collecte et préparation des données d'entraînement S,
- Choix d'une classe de modèle d'apprentissage machine $h \in \mathcal{H}$,
- Minimisation d'un risque moyen $\hat{R}(h)$ (plutôt; par exemple, que du risque en pire cas - *worst-case risk*),

- Choix d'une fonction de perte $l(h(x), y)$ (certaines fonctions de perte sont plus résistantes aux "valeurs aberrantes" que d'autres),
- Choix d'un benchmark d'évaluation.

Dans ce qui suit, nous décrivons les différentes formes de biais qui peuvent survenir à chaque étape et nous les classons en deux catégories : les biais liés aux données et les biais algorithmiques. En pratique, les pipelines ML sont beaucoup plus complexes et les biais sont susceptibles d'apparaître à de nombreux autres niveaux. Nous renvoyons à Suresh & Guttag (2021) et à Mitchell & al. (2021) pour une taxonomie plus complète des sources de problèmes potentiels dans les décisions basées sur l'apprentissage automatique.

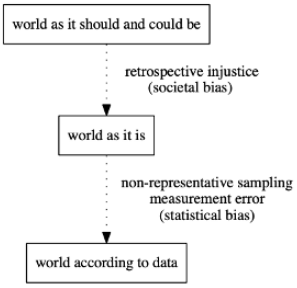

Figure 1: Illustration de biais dans les données [Mitchell et al., 2021].

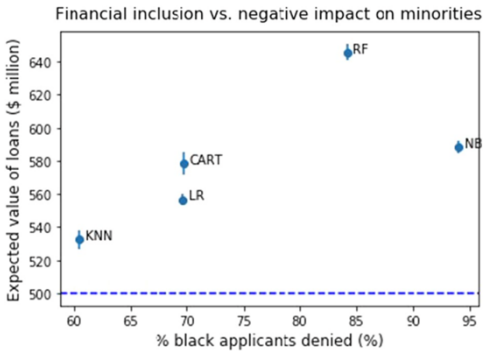

Figure 2: Inclusion financière vs. impact négatif sur les minorités [Lee & Floridi, 2021]

Biais dans les données. La présence de biais dans les décisions basées sur l'apprentissage machine est généralement attribuée à un biais dans les données. Plus précisément, les biais surviennent dans le processus de génération des données utilisées pour former et évaluer les modèles d'apprentissage machine, et il peut apparaître pour diverses raisons (voir figure 1). Tout d'abord, en statistiques, un biais d'échantillonnage ou de sélection se produit lorsque l'ensemble de données collectées sous-représente certaines parties de la population. Par exemple, la plupart des technologies de reconnaissance faciale sont entraînées sur des ensembles de données qui sont biaisées en faveur des hommes caucasiens à la peau claire (Buolamwini & Gebru, 2018), ce qui se traduit par de moins bonnes performances de prédiction pour les femmes à la peau plus foncée. De même, les ensembles de données de reconnaissance d'objets sont "biaisés" en faveur des pays développés et ne parviennent pas à détecter avec précision les objets situés dans des

ménages à faible revenu (de Vries et al., 2019). Deuxièmement, les biais de mesure se produisent lorsque les étiquettes y ne sont que des proxies des véritables résultats que nous voulons prédire. Par exemple, lorsque l'apprentissage machine est utilisé pour prédire la criminalité dans les problèmes de détention provisoire, la criminalité n'est en fait jamais mesurée et l'algorithme est plutôt entraîné à prédire les arrestations (Corbett-Davies & Goel, 2018, Lum & Isaac, 2016). Cependant, les taux d'arrestation diffèrent fortement selon les quartiers et les races, ce qui fait des arrestations un indicateur biaisé de la criminalité réelle. Troisièmement, les données peuvent refléter un biais sociétal ou historique même si elles sont parfaitement représentatives ou mesurées. Il y a biais sociétal lorsque les inégalités du monde réel sont structurelles, et lorsque leur reproduction ou leur exacerbation nuit de manière indésirable à un groupe défavorisé. Par exemple, même si nous étions en mesure de collecter parfaitement des données sur la qualification des hommes par rapport aux femmes dans les emplois technologiques (Dastin, 2018), il faut être conscient du préjudice potentiel ou de l'acceptabilité sociale de prédire systématiquement que les hommes sont plus compétents que les femmes pour des nouvelles candidatures. Étant donné que les modèles d'apprentissage machine sont formés pour s'adapter à la distribution des données, ils sont susceptibles de refléter les biais dans les données provenant de l'échantillonnage, des mesures et des processus historiques. ("*bias in*, *bias out*" [Kallus & Zhou, 2018]).

La collecte d'un plus grand nombre de données est souvent considérée comme une solution simple pour atténuer les biais des données, et les biais en apprentissage machine dans leur ensemble (Hooker, 2021). Bien qu'étant une stratégie pertinente pour compenser les biais de sélection (Chen et al., 2018), elle ne peut pas circonscrire les biais de mesure et les biais historiques. Pour remédier à ces biais, des auteurs ont proposé plusieurs techniques permettant d'apprendre des représentations équitables des données (Zemel & al., 2013, Louizos et al., 2015) ou de "dé-biaiser" les représentations standard des données (Bolukbasi & al., 2016, Zhang & al., 2018).

Biais algorithmique. Les choix algorithmiques, tels que (2-5) dans la liste ci-dessus, ont également un impact sur les résultats prédits par l'algorithme, d'une manière qui peut affecter différemment différents groupes sociaux. Dans la littérature, lorsque les algorithmes d'apprentissage machine exacerbent les biais existants dans les données, ce problème est appelé amplification des biais. (Zhao & al., 2017, Lloyd, 2018, Hall & al., 2022).

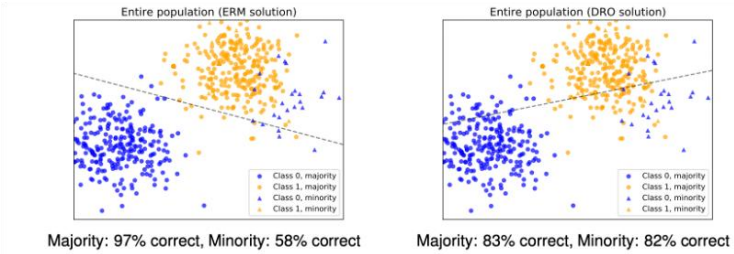

Figure 3: Illustration de l'ERM vs. DRO sur des données synthétiques [Słowik et Bottou, 2021]. Dans cet exemple, la minimisation du risque moyen est préjudiciable à la population minoritaire, tandis que la minimisation du risque pire cas égalise les taux d'erreur entre les groupes.

En premier lieu, il est établi que le choix d'un modèle d'apprentissage machine impacte le comportement prédictif, en particulier la qualité des prédictions sur des données inobservées. Toute considération de *fairness* mise de côté, un exemple largement enseigné est que la variation du degré d'un modèle polynomial conduit à différents niveaux d'overfitting. Comme le note Hooker (2021) – *"our modeling choices [...] express a preference for final model behaviour."* Dans le contexte de l'apprentissage machine équitable, le choix du modèle ML est un choix implicite pour le comportement du modèle sur différents groupes sociaux. La figure 2 illustre un exemple de prêt bancaire tiré de [Lee & al., 2020], avec différents modèles ML conduisant à différents compromis entre la proportion de demandeurs noirs refusés et l'espérance de la valeur des prêts accordés, montrant que certains modèles conduisent à une inclusion financière élevée uniquement à un coût élevé pour le groupe défavorisé.

En second lieu, le paradigme de l'ERM, qui consiste à minimiser des moyennes, est en soi un choix qui a un impact sur les prédictions. Des paradigmes alternatifs peuvent avoir un impact différent sur les groupes défavorisés. Une telle alternative utilise le cadre de l'optimisation distributionnellement robuste (*distributionally robust optimization*, DRO) [Ben-Tal & al., 2009], et consiste à minimiser le risque en pire cas, plutôt que le risque moyen. La figure 3 (tirée de Słowik & Bottou [2021]) montre un exemple synthétique où l'on constate que l'utilisation de l'ERM génère de forts taux d'erreurs pour les groupes minoritaires, alors que la DRO égalise les taux d'erreur entre les groupes, majoritaire et minoritaire, tout en assurant un taux d'erreur global raisonnablement faible. D'autres auteurs défendent des approches "min-max" similaires de la *fairness* (Hashimoto et al., 2018, Levy et al., 2020, Wang et al., 2020, Diana et al., 2021).

Il est important de noter que l'évaluation des modèles est également une source de biais, conduisant à des boucles de rétroaction biaisées qui renforcent les stéréotypes existants. Les critères de benchmark standards peuvent mal représenter certaines sous-populations et les mesures de performances moyennes peuvent cacher des performances médiocres sur un groupe minoritaire (Suresh & Guttag, 2021). Les audits existants pour la *fairness* évaluent les modèles en quantifiant les disparités de performance entre les groupes (Buolamwini et Gebru, 2018, Datta & al., 2015, Sweeney, 2013), ce qui peut être considéré comme des mesures supplémentaires pour éviter une évaluation biaisée. En outre, les model cards ont été introduites comme une norme pour rendre compte de plusieurs métriques pour l'évaluation des dommages potentiels des modèles pour les différents groupes culturels et démographiques, ainsi que leurs intersections (Mitchell & al., 2019). Ces *model cards* ont été de plus en plus adoptées par les praticiens de l'apprentissage machine (e.g., Menon & al., 2020).

1.3 Equité en apprentissage machine

Face aux exemples réels de biais dans les prédictions des algorithmes d'apprentissage machine, les chercheurs ont activement conçu des méthodes pour les atténuer, ce qui a donné lieu à de nombreux travaux sur la *fairness* dans l'apprentissage machine (qui font l'objet d'un examen approfondi dans Barocas et al. (2019). De nombreuses mesures de *fairness* ont été proposées. Il a été montré que les critères de *fairness* les plus utilisés sont incompatibles (Kleinberg et al., 2016, Chouldechova, 2017), et qu'ils présentent des risques

pour les populations qu'ils sont censés protéger (Corbett-Davies & Goel, 2018). Il est désormais reconnu que les approches purement techniques sont des correctifs superficiels pour lutter contre les biais dans l'apprentissage machine (Selbst & al., 2019), et que des approches plus réfléchies pour un apprentissage machine équitable devraient tenir compte du contexte social et juridique, que nous allons envisager sur le problème de la discrimination indirecte.

2. Problématiser la question éthique, entre droit et technique algorithmique : la discrimination indirecte

Le problème de l'équité est un problème classique en économie, en philosophie morale et politique, d'autant plus complexe qu'il est intimement lié à la question de la justice sociale et que celle-ci fait l'objet de systèmes théoriques hétérogènes (utilitarisme, libertarianisme, égalitarisme, marxisme) (Arnsperger & Van Parijs, 2003). Dans la doctrine juridique, la formule d'Aristote de l'égalité proportionnelle selon laquelle "*things [and persons] that are alike should be treated alike, while things that are unalike should be treated unalike in proportion to their unalikeness*" est bien connue, même si la doctrine et la pratique juridique en font apparaître les limites (Schiek & al., 2007, p. 27). La doctrine et la pratique juridiques retiennent plutôt d'autres principes aristotéliciens : la justice corrective et la justice distributive, dont la distinction est au fondement de la séparation du droit privé (un préjudice contractuel doit être compensé – corrigé) et du droit public (qui organise les droits d'accès à des ressources – d'où l'aspect distributif). Dans le prolongement de ces débats, la doctrine et la pratique juridiques défendent que l'égalité doit être substantielle plutôt que simplement procédurale, ce qui signifie que le droit anti-discrimination vise à transformer les réalités sociales pour y faire advenir davantage d'égalité. Toutefois, "*Substantive equality is not a uniform concept. It comprises equality of results, equality of opportunities, equality in relation to substantive rights such as freedom of profession or capabilities, and equal respect*…" (Schiek & al., 2007, p. 28) Il n'existe donc pas, ni en économie, ni en droit, ni en philosophie morale, de définition et de critère uniques de la *fairness*.

Cela étant dit, il faut être conscient du fait que la recherche sur la *fairness* en apprentissage machine est essentiellement américaine ou sous son influence ; elle prend comme arrière-plan, le plus souvent de manière sous-jacente, le contexte juridique et institutionnel des Etats-Unis. Les groupes ethno-raciaux y sont reconnus statistiquement ; il existe de grandes lois fédérales sur les droits civiques et la non-discrimination de groupes protégés ; la notion d'impact disproportionné (*disparate impact*) est une notion établie par le droit. Nous questionnons ici l'application des approches de la *fairness* à d'autres contextes juridiques et institutionnels, en développant le cas du droit européen anti-discrimination (2.1.). Le croisement entre les approches d'apprentissage machine d'un côté, juridique de l'autre, est une nécessité pour créer les conditions de modèles algorithmiques de décision cohérents avec le droit anti-discrimination, qui promeut certaines formes d'équité. Cela suppose de définir les termes dans lesquels les relations entre le droit et les modèles algorithmiques peuvent être posés, dans le but ultime d'algorithmes conformes au droit et aux valeurs qu'il porte (Koulu, 2021) (2.2.).

2.1. Gouverner la discrimination indirecte

Nous analysons la question de la *fairness* sous un double angle : a) la discrimination indirecte et les dispositifs juridiques concernant le *disparate impact* en menant une comparaison entre les Etats-Unis et l'Union européenne (2.1.1.) ; b) l'écart entre un droit anti-discrimination dont la mise en œuvre et les interprétations par les tribunaux est dynamique, et une formalisation idéale-typique en apprentissage machine (2.1.2.).

2.1.1. Le droit anti-discrimination aux Etats-Unis et dans l'Union européenne : des régimes différents

L'impact disproportionné est une question commune à la technique et au droit. Ainsi que mentionné plus haut, le fait que l'immense majorité des travaux de recherche en apprentissage automatique prenne pour arrière-plan le droit anti-discrimination des Etats-Unis conduit à caractériser également le contexte juridique européen. Nous nous proposons ici, via une comparaison des contextes juridiques américain et européen, de : a) évaluer la pertinence des travaux en Machine Learning éthique dans le contexte européen b) caractériser les normes éthiques et les mesures de la discrimination dans l'UE.

Concernant les Etats-Unis, il faut souligner que la notion de *disparate impact* y a émergé, sur base du Civil Rights Act de 1964, dont le Titre VII précise que les pratiques qui, sous l'apparence de règles neutres, ont un impact disproportionné sur la classe protégée comparativement à la classe non protégée sont prohibées – à moins qu'un intérêt légitime (une nécessité des affaires -*business necessity*) puisse être valablement invoqué.

Outre le titre VII, le titre VI du Civil Rights Act donne au gouvernement fédéral le droit d'agir en justice en cas d'impact disproportionné. De plus, deux lois de 1967 - l'Age Discrimination in Employment Act et le Fair Housing Act – donnent aux individus un droit d'action au titre du *disparate impact*.

Ces dispositions du droit fédéral se sont concrétisées par la mise en œuvre d'une règle de calcul visant à déterminer si une situation d'impact disproportionné peut être observée en pratique, connue comme la « règle des 80% » (ou du 4/5).

Elaborée initialement dans l'Etat de Californie, elle y fut formalisée dans le *State of California Guideline Selection Procedures* en 1972. Elle fut ensuite codifiée au niveau fédéral dans les *Uniform Guidelines on Employee Selection Procedures* (1978) utilisées par l'*US Equal Employment Opportunity Commission* (EEOC), mais aussi par le DoJ et le DoL dans leurs actions devant les tribunaux au titre du titre VII. La règle des 80% consiste à faire le ratio du taux d'embauche d'individus de la classe non protégée et de celui d'individus de la classe protégée : si une entreprise embauche 50% des hommes candidats à un emploi et 20% des femmes candidates, le taux d'embauche est égal à 50/20, soit 0,4 : le taux d'embauche des femmes représente 40% de celui des hommes. Selon le guide de l'EEOC, si le ratio des taux de sélection est inférieur à 80%, cela est l'indice d'une situation discriminatoire[2].

[2] La pertinence de la règle des 80% est discutée, à la fois sur le plan juridique et sur celui de la statistique (Peresie, 2009). Nous y reviendrons plus loin.

Sur le plan législatif, la doctrine du *disparate impact* a été importée au Royaume-Uni du début des années 1970, l'arrêt *Griggs* avait séduit des dirigeants politiques britanniques lors d'une visite aux Etats-Unis, à un point tel qu'ils ont intégré la discrimination indirecte dans la première loi sur la discrimination sexuelle (le *Sex Discrimination* Act de 1975) et dans la révision du *Race Relations* Act en 1976 (Suk, 2014, p. 287). Comme on le verra plus en détail plus bas, la mise en œuvre du Titre VII du Civil Rights Act par les tribunaux américains a progressivement introduit des restrictions de plus en plus importantes à son invocation par les victimes de discrimination indirecte : le plaignant doit établir avec précision laquelle des règles de gestion de l'emploi dans les entreprises est la cause d'un impact disproportionné ; apporter une preuve statistique probante ; démontrer que l'employeur a refusé de mettre en œuvre des règles de gestion qui auraient pu réduire l'impact disproportionné. Même si la victime parvient à surmonter ces épreuves, l'employeur peut démontrer que la règle ou la pratique contestée répond à une *business necessity*.

L'Union européenne a promulgué des directives de lutte contre la discrimination, directe et indirecte, fondée sur la race ou l'origine ethnique, la religion ou les convictions, un handicap, l'âge ou l'orientation sexuelle : deux directives respectivement sur l'égalité en matière d'emploi[3] et sur l'égalité entre les races[4] ont été adoptées en 2000. En 2009, lors du Traité de Lisbonne, une clause horizontale fut introduite afin d'intégrer la lutte contre les discriminations dans toutes les politiques et actions de l'UE (article 10 du TFUE[5]). La CJUE a traité de nombreuses affaires portant sur des discriminations, soit en réponse à des questions préjudicielles de juridictions nationales, soit sur le fond.

Suk (2014) souligne combien le droit américain et le droit européen anti-discrimination sont différents.

- La jurisprudence européenne est abondante sur l'égalité salariale entre les hommes et les femmes, alors que les tribunaux fédéraux américains l'ont exclue du système de *disparate impact*. A cet égard, Selmi (2006), au terme d'une analyse empirique du contentieux concernant des cas de *disparate impact* aux Etats-Unis, souligne qu'il a eu un impact nettement limité en dehors du contexte des tests écrits utilisés dans le recrutement d'employés.
- Les décisions de la CJUE sont plus protectrices des salariées qu'aux Etats-Unis, notamment parce que contrairement à la Cour européenne, les tribunaux américains ont refusé d'appliquer la théorie du *disparate impact* aux situations de travail à temps partiel (bien plus important pour les femmes que pour les hommes).
- Pour les demandeurs, il est relativement plus facile de plaider une discrimination *prima facie* en Europe qu'aux Etats-Unis : l'obligation américaine de déterminer quelle disposition, règle ou pratique précise serait la cause d'un impact disproportionné n'a pas d'équivalent en Europe.

[3] Directive 2000/78/CE du Conseil du 27 novembre 2000, portant création d'un cadre général en faveur de l'égalité de traitement en matière d'emploi et de travail, Journal officiel n° L 303 du 02/12/2000 p. 0016 - 0022

[4] Directive 2000/43/CE du Conseil du 29 juin 2000 relative à la mise en œuvre du principe de l'égalité de traitement entre les personnes sans distinction de race ou d'origine ethnique, Journal officiel n° L 180 du 19/07/2000 p. 0022 - 0026

[5] Traité sur le fonctionnement de l'Union européenne.

	Etats-Unis	**Union européenne.**
Principal objet	Inégalités raciales Embauche et déroulement des carrières	Egalité salariale entre femmes et hommes
Travail à temps partiel	Non pris en compte	Pris en compte
Charge de la preuve (pour le demandeur)	Exigeante et restrictive	Pas très contraignante
Justification des règles ou pratiques ayant un impact disproportionné sur les employés	L'argument de la nécessité des affaires bénéficie aux employeurs	Approche équilibrée de la nécessité des affaires dans la jurisprudence de la CJUE

Tableau 1. Comparaison entre les Etats-Unis et l'U.E. (Source : auteurs).

Ainsi, la CJUE privilégie l'égalité salariale entre hommes et femmes ; elle est relativement peu diserte sur les inégalités raciales alors qu'elles sont centrales aux Etats-Unis ; aux Etats-Unis le *disparate impact* est appliqué à l'emploi, hors salaire, c'est-à-dire aux situations de recrutement et de promotion. Les situations de travail à temps partiel n'entrent pas dans son champ d'application (tableau 1).

2.1.2. Impact disproportionné : un concept juridique dynamique

On s'interroge ici sur la correspondance entre approche informatique (idéale-typique) et pratique judicaire de l'impact disproportionné. La formule élémentaire du *disparate impact* est la suivante :

$$\mathrm{DI} = \frac{P(Y=1 \mid S=0)}{P(Y=1 \mid S=1)}$$

Elle consiste à comparer la probabilité d'obtenir un résultat (Y=1) selon que le groupe (S) est protégé (S=1) ou non protégé (S=0).

Comme envisagé ci-dessus, la mise en œuvre de ces dispositions juridiques s'est traduite par l'application d'une règle de calcul visant à déterminer si une situation d'impact disproportionné peut être observée en pratique (la « règle des 80% »).

L'Uniform Guideline a été codifié dans le Code of federal regulation dont l'article 1607 du Titre 29 prévoit que: "*A selection rate for any race, sex, or ethnic group which is less than four-fifths (⅘) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group*" (29 CFR 1607).

Par ailleurs l'EEOC est attentive à l'impact disproportionné que certaines pratiques de recrutement peuvent susciter, comme la consultation du casier judiciaire (*criminal record*) et les antécédents bancaires (défauts de remboursement de crédit), dans la mesure où les afro-américains sont sur-représentés dans la population carcérale et dans celle des débiteurs défaillants. Dans la décision de la Cour d'appel du 4ème circuit *EEOC v. Freeman* (2015), le juge a invalidé la demande de l'EEOC d'enjoindre l'entreprise à mettre un terme à la consultation du casier judiciaire et des antécédents bancaires. Le juge Titus est allé jusqu'à écrire que la demande de l'EEOC est une « théorie qui cherche sa pratique ». Il ajoute que les employeurs sont soumis à une tension entre, d'un côté s'exposer à une responsabilité potentielle si un employé recruté en aveugle s'avère commettre des actes criminels ou frauduleux, et de l'autre de s'exposer à des poursuites par l'EEOC.

La règle des 80% promue par l'EEOC a des limites, reconnues comme telles dans la loi. Selon le Code of Federal Regulation (section, article 1607), de fortes différences dans les taux de sélection ne signifient pas un impact disproportionné lorsqu'elles portent sur de petits effectifs ou ne sont pas statistiquement significatives. Dans ce cas, l'EEOC admet une preuve statistique sur une plus longue durée et/ou à étudier l'impact une pratique similaire mise en œuvre dans des circonstances similaires. La sensibilité aux petits effectifs rend la règle inéquitable lorsqu'elle conduit à exposer davantage les très petites entreprises à la responsabilité pour un impact discriminatoire que les grandes entreprises regroupant des effectifs beaucoup plus nombreux de candidat.e.s à l'embauche et de candidatures retenues pour examen à l'embauche (Peresie, 2009)[6]. Enfin, la règle des 80% a des limites en termes d'établissement de la causalité, au point que les tribunaux préfèrent recourir à des tests de significativité statistique (Peresie, 2009, p. 785).

Dans le système de common law, tel le système juridique des Etats-Unis, la jurisprudence a une importante particulière : étant reconnue comme une source du droit, elle interprète les lois et leur donne leur sens pratique. La décision *Griggs v. Duke Power* rendue en 1971 par la Cour suprême a marqué une avancée notable dans la garantie des droits civiques au profit des afro-américains. L'entreprise concernée pratiquait des tests d'intelligence et exigeait que les employés aient atteint le niveau Bac (*College*) pour décider de promotions à des postes plus rémunérés. Or, les salariés d'origine afro-américaine étaient peu promus, alors que les salariés caucasiens l'étaient fréquemment. Dans sa décision, la Cour suprême a posé comme règle que si une pratique dont le fonctionnement exclut les membres d'un groupe protégé n'est pas basée exclusivement sur les performances au travail, alors elle est prohibée. La pratique des promotions par l'entreprise incriminée n'était pas, en l'occurrence, basée sur les performances au travail (Vinik, 2010).

Après cette décision d'ouverture aux droits civiques dans des situations de discrimination indirecte, la Cour suprême des Etats-Unis a suivi une jurisprudence plus restrictive. Dans la seconde moitié des années 1980, les restrictions ont porté sur les protections

[6] Un exemple numérique permet de l'établir. Soient un grande entreprise (GE) et une petite (PE). La GE reçoit 20 000 candidatures réparties à égalité entre hommes et femmes. Le recruteur sélectionne le CV de 3200 femmes et 4000 hommes, soient 32% et 40%. Le sélection est alors égal à 0,32/0,4 = 0,8, conforme à la règle de l'EEOC. La PE reçoit 20 candidatures, réparties à égalité entre hommes et femmes. L'employeur sélectionne le CV de 4 hommes (40% des candidats) et 3 femmes (30% des candidates),d'où un ratio de sélection inférieur au seuil de 80% : 0,3/0,4 = 0,75.

constitutionnelles ; puis, dans les années 1990, les restrictions ont été introduites autour des règles de preuve pesant sur les plaignants. Enfin, dans les années 2000, la jurisprudence a limitée les moyens d'action des employeurs pour éviter des résultats racialement disproportionnés (cf. *Ricci v. DeStefano*) (Suk, 2014).

La Cour suprême a barré les actions en impact disproportionné lorsqu'elles invoquaient l'Equal Protection Clause contenue dans le 14ème amendement à la Constitution, considérant que cette clause ne porte que sur les seules discriminations intentionnelles (Suk, 2014). Dans sa décision *Arlington Heights v. Metropolitan Housing Corp*. (1977), la Cour suprême a considéré que "*Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause*." Il en de même avec le 5ème amendement (« *due process clause* ») : dans la décision *Washington v. Davis* rendue en 1976, la Cour suprême a considéré que l'argument d'un *disparate impact* ne peut pas être utilisé dans le cadre d'une demande au titre du 5ème amendement à moins que le plaignant démontre que des standards racialement neutres ont été utilisés avec une intention discriminatoire[7]. Un an plus tard, dans *Pothard v. Rawlinson* (1977), la Cour a considéré que le Titre VII du Civil Rights Act ne rend pas illégales les exigences physiques imposées à l'entrée dans la carrière de gardien de prison, même si elles excluent 40% des femmes candidates.

C'est dans les années 1980 que la jurisprudence a pris un virage restrictif, en durcissant les standards de preuve imposés aux plaignants. Ces restrictions ont culminé avec *Wards Cove Packing v. Atonis* (1989) : la Cour apporté une restriction très importante aux actions fondées sur un *disparate impact* en posant la règle de preuve selon laquelle le plaignant doit établir a) quelle pratique ou règle précisément définie est à l'origine d'un impact indirectement discriminatoire, et b) que l'employeur a refusé de mettre en œuvre des pratiques ou des règles qui auraient pu satisfaire les griefs du plaignant. De plus, l'entreprise incriminée a la faculté d'argumenter en faveur de la règle ou la pratique à l'origine d'un impact disproportionné, au regard des nécessités des affaires (*necessity of business*). Il faut noter que l'obligation de définir précisément la pratique ou la règle à l'origine d'un impact disproportionné a été étendue à d'autres domaines, et est devenue une contrainte sur la possibilité d'avancer des preuves statistiques : dans la décision *Texas Department of Housing & Community Affairs v. Inclusive Community Project* rendue en 2015, la Cour suprême a considéré qu'une demande fondée sur la parité statistique doit être rejetée si elle n'établit pas quelle mesure de politique publique précisément définie est censée générer un impact disproportionné. La Cour estime également que les politiques du logement doivent pouvoir bénéficier de marges de manœuvre, lesquelles sont nécessairement pour réaliser leurs intérêts légitimes (Vinik, 2010).

Une troisième phase des restrictions apportées par la Cour suprême à la théorie du *disparate impact* a été amorcée dans les années 2000. La décision la plus significative est *Ricci v, DeStefano*, rendue en 2009 à la suite de l'annulation par la ville de New Haven (Connecticut) d'un concours interne organisé pour la promotion des pompiers de la ville, le taux de succès des pompiers blancs étant le double de celui des afro-américains. Le Maire (DeStefano) a considéré qu'un impact disproportionné justifiait l'annulation du concours,

[7] Phénomène que Barocas et Selbst (2016) appellent le « masking » : «… any form of discrimination that happens unintentionally can also be orchestrated intentionally" (p. 692).

laquelle a été contestée par les pompiers blancs et un hispanique dont Ricci qui avaient passé avec succès les épreuves et auraient dû être promus. La Cour suprême a jugé que la décision d'annulation était contraire au Titre VII du Civil Rights Act, dans la mesure où la ville de New Haven ne disposait pas d'un "fondement solide en matière de preuve" du fait qu'elle se serait exposée à une responsabilité au titre du *disparate impact* si elle avait promu les pompiers blancs et hispaniques au lieu des pompiers afro-américains. En somme, la Cour suprême a limité les moyens d'action utilisables, pour les employeurs, d'éviter des résultats racialement disproportionnés.

Ainsi, alors que les termes de la loi ne donnent pas un contenu fixe à la doctrine du *disparate impact* et leur sa portée empirique est façonnée (aux Etats-Unis de manière de plus en restrictive) par les tribunaux, sa modélisation dans les travaux menés en *computer science* est basée sur une formulation mathématique forcément insensible à sa mise en œuvre dans le monde réel. Cependant, certains (en particulier Barocas & Selbst, 2016) prennent acte de ces restrictions jurisprudentielles et insistent sur le fait que si les données d'apprentissage reflètent des biais sociaux, il n'y a pas de solution technique en la matière, d'autant plus que cela pourrait poser des problèmes juridiques voire politiques (ibid.). Feldman et al. (2015) ont proposé une procédure pour certifier l'absence de disparate impact dans les données, et faire en sorte que l'algorithme d'apprentissage automatique ne puisse pas prédire l'attribut protégé. Feldman et al. s'attachent explicitement à la règle des 80% de l'EEOC et utilisent, dans la partie expérimentale, les données issues du concours de promotion des pompiers, qui a donné lieu à la décision Ricci v. DeStefano (Ricci dataset).

2.2. Faire dialoguer droit anti-discrimination et algorithmes : comment ? pour quelles finalités ?

Des auteurs de plus en plus nombreux s'attachent à mettre en relation les concepts juridiques et les concepts techniques liés à la question de la *fairness* et de la discrimination directe et indirecte. Des enjeux de politique publique sont posés, qui appellent un décloisonnement des disciplines. Le fait que la littérature soit de plus en plus abondante dans ce domaine est une bonne nouvelle. Mais il nous apparaît que les problématiques et les objectifs visés par les auteurs sont diverses et révèlent des postures différentes qu'il semble important d'identifier. Nous en identifions quatre types.

2.2.1. Comparer les concepts per se

Xiang et Raji (2019) comparent les concepts de l'apprentissage machine et du droit anti-discrimination des Etats-Unis sur plusieurs dimensions. Nous présentons ici leur comparaison assortie de commentaires.

Equité procédurale: en apprentissage machine, elle renvoie à l'identification des caractéristiques des inputs (en particulier des proxies utilisées) qui conduisent à un résultat particulier du modèle, la focale étant mise sur l'algorithme lui-même et ses prédictions. A l'inverse, pour le droit, il s'agît de principes de gouvernance du processus de prise de décision. L'accent est mis sur le « système entourant l'algorithme et ses usages ».

Discrimination : en apprentissage machine, elle est souvent présentée comme le produit d'une corrélation injuste entre des variables de classe protégée et une métrique d'intérêt telles que les résultats, les taux de faux positifs entre les classes, ou autre. Pour le droit, il

s'agît d'un enjeu social pour lequel des règles sont produites, qui visent à réprimer les intentions discriminatoires ou d'exclusion ou la causalité. Notons que Xiang et Raji n'évoquent pas la discrimination indirecte, qui ne suppose pas d'intention discriminatoire.

Classe protégée/attribut sensible: alors que c'est une question majeure pour le droit, la recherche en apprentissage machine en fait une question de caractéristiques qui ne devraient pas intervenir dans la décision algorithmique, sans prise en compte du droit.

Anti-classification et anti-subordination : la notion d'anti-classification en droit signifie simplement que l'Etat ne peut pas classifier les individus en fonction de leur race, sexe, âge, etc. Elle est cohérente avec le concept de *fairness through unawareness* dont la communauté en apprentissage machine travaillant sur la *fairness* est familière.

Le principe de l'anti-subordination renvoie à l'égalité des droits comme objectif, cependant inatteignable dans une société qui connait une forte stratification sociale à moins que le droit se donne pour objectif de renforcer la position des catégories sociales les moins favorisées. Xiang et Raji estiment que c'est une dimension rarement prise en considération dans la littérature en apprentissage machine. Il existe en effet des approches, minoritaires, comme les approches "min-max" qui visent à minimiser les erreurs de classification pour les groupes sociaux les plus désavantagés (cf. les approches de type DRO évoquées précédemment). Do et al. (2021), qui recourent à la dominance de Lorenz, considèrent la « fairness » au sens de l'amélioration de l'utilité des individus dont la situation est le plus défavorable, en suivant le principe du transfert en économie.

Affirmative action : la recherche en apprentissage machine est ouverte sur ce sujet, dans ses développements sur la parité démographique ; dans le droit, elle est sujette à des interprétations variables dans le temps.

Traitement différencié (*disparate treatment*) et impact disproportionné (*disparate impact*): Du point de vue juridique, le *disparate treatment* renvoie à un traitement intentionnellement discriminatoire ; pour l'apprentissage machine, l'intention n'est pas une dimension pertinente, puisqu'il apparaît lorsqu'un attribut protégé est utilisé dans le processus de décision, et qu'il est possible d'en éviter le recours. Dans le droit, Xiang et Raji estiment (à tort) que le *disparate impact* est illégal s'il est intentionnel. Alors que du point de vue technique le *disparate impact* apparaît lorsque les résultats entre sous-groupes diffèrent, même sans intentionnalité.

En définitive, dans la perspective de l'interdisciplinarité, il est très important de comprendre la manière dont un même concept est mis en forme dans les deux champs et quelle signification s'y attache. Ce n'est pas cependant une fin en soi. C'est un travail nécessaire mais qui ne doit être qu'un préalable à une ouverture réciproque des champs, dont verrons plus loin quelques possibilités.

2.2.2. Qualifier les normes juridiques à partir d'indicateurs techniques : mesures de la discrimination de groupe et droit anti-discrimination

Pedreschi, Ruggieri et Turini (2012) ont proposé une classification juridiquement fondée des indicateurs de mesure de la discrimination Ils partent de l'idée que l'interprétation de la législation conduit à des mesures différentes de la discrimination et à des *rankings* différents de contextes possiblement discriminatoires.

Les auteurs partent d'un table de contingence permettant de caractériser quatre situations, selon que le groupe est protégé ou non, et que la décision est positive ou négative, et de définir une série de mesures de la discrimination.

	benefit		
group	denied	granted	
protected	*a*	*b*	n_1
unprotected	*c*	*d*	n_2
	m_1	m_2	*n*

$p_1 = a / n_1$ $p_2 = a / n_2$ $p = m_1 / n$

$RD = p_1 - p_2$ $RR = \frac{\mathbf{p1}}{\mathbf{p2}}$ $RC = \frac{\mathbf{1-p1}}{\mathbf{1-p2}}$ $OR = \frac{\mathbf{RR}}{\mathbf{RC}} = \frac{\mathbf{a/b}}{\mathbf{c/d}}$

$ED = p_1 - p$ $ER = \frac{\mathbf{p1}}{\mathbf{p}}$ $EC = \frac{\mathbf{1-p1}}{\mathbf{1-p}}$

P1 est la proportion de décisions négatives pour les membres du groupe protégé

P2 est la proportion de décisions négatives pour les membres du groupe non protégé

P est la proportion globale de décisions négatives

RD (risk difference) est la différence (absolue) de risque : RD = P1 – P2

RR (relative risk) est le risque relatif : RR = P1/P2

RC (relative chance, ou selection rate) est la chance relative d'obtenir une décision positive :

RC= (1-P1)/(1-P2)

OR est l'odds ratio : OR = p1(1 - p2)/p2(1 - p1)

ED (extended difference) : différente entre la proportion décisions négatives du groupe protégé et le différence moyenne de proportion p : ED = P1 - P

ER (extended ratio) : ER = P1 / P

EC (extended chance) : EC = 1 – P1 / 1 - P

Figure 4- : Mesures de la discrimination [Predeschi, Ruggieri & Turini, 2012]

Pedreschi, Ruggieri et Turini (2012) estiment que "*From a legal point of view, several measures are adopted worldwide. UK law mentions risk difference, EU Court of Justice has given more emphasis on the risk ratio, and US laws and courts mainly refer to the selection rate*" (Pedreschi, Ruggieri et Turini, 2012, p. 2). Mais dans la conclusion de leur article le

diagnostic est différent : le système juridique européen est connecté au RC (*relative chance*) et le système américain au RR (*risk ratio*), ce qui jette un doute sérieux sur la cohérence de leur analyse des dispositifs juridiques.

Il est en effet un peu aventureux d'associer de manière univoque une mesure de la discrimination et un système juridique national ou régional. Ainsi, les directives européennes « égalité » des années 2000 raisonnent en dominante en terme de différence de risque (pour établir une discrimination *prima facie* il suffit d'établir qu'une disposition, critère ou pratique place des personnes d'origine raciale ou ethnique dans une position désavantageuse par rapport à d'autres personnes – Schiek & al, 2007, § 13.072) ; la jurisprudence de la CJUE retient plutôt le principe de preuve d'une position disproportionnellement désavantageuse d'un groupe pour qualifier une situation de discrimination indirecte, ce qui renvoie à un risque relatif (Schiek & al., 2007, § 12.125 – inégalités salariales entre hommes et femmes et § 12.659 – doctrine plus générale). On peut se référer à la décision *Kirshammer-Hack* de la CJCE (30 novembre 1993) : dans cette affaire, une employée à temps partiel dans une entreprise de moins de 5 salariés a été licenciée sans indemnité, le droit allemand ne garantissant le droit à indemnisation qu'aux entreprises de plus de 5 salariés ; la salariée licenciée s'estimait discriminée compte tenu du fait que les femmes sont plus nombreuses que les hommes dans les très petites entreprises et à travailler à temps partiel. La CJCE a estimée qu'il y aurait une discrimination entre hommes et femmes s'il était établi que les petites entreprises emploient un pourcentage considérablement plus élevé que d'hommes, ce qui n'est pas établi statistiquement par la demandeuse.

Dans la même lignée, l'arrêt de la CJCE *Krüger* (9 sept 1999), a été rendu en réponse à une question préjudicielle de la Munich Labour Court : l'exclusion des employés à temps partiel (prévue par la convention collective) du bénéfice d'une prime salariale est-elle discriminatoire ? La CJCE a considéré que si le tribunal national estime que la règle, même formellement indépendante du sexe, affecte en réalité un pourcentage considérablement plus élevé de femmes que d'hommes, alors il s'agit d'une discrimination sexuelle.

De nombreux autres travaux portent sur les métriques de la discrimination, directe et indirecte (Abu Elyounes, 2020 ; Besse & al. 2018 ; Chouldechova 2016 ; Dwork & al., 2011). Nous nous en tenons à la discrimination indirecte pour laquelle plusieurs critères de *fairness* ont été formalisés, et qui sont incompatibles les uns avec les autres (Chouldechova, 2016 ; Kleinberg & al., 2016) :

Parité statistique	$P(\hat{y}=1 \mid A=a) = P(\hat{y}=1 \mid A=a')\ \forall a,a'$
Egalité des chances	$\mathrm{P}(\hat{y}=1 \mid y=1, A=a) = (\hat{y}=1 \mid y=1, \mathrm{A}=a')\ \forall \mathrm{a}, a,'$
Odds égalisés	$\mathrm{P}(\hat{y}=1 \mid y=i, A=a) = (\hat{y}=1 \mid y=i, A=a')\ \forall i \in \{0,1\}, a, a'$
Calibration par groupe	$P(y=1 \mid S=s, A=a) = P(y=1 \mid S=s, A=a'), \forall \mathrm{s} \in \mathrm{R}, \forall \mathrm{a}, \mathrm{a}'$

$\hat{y}$ prédicteur du résultat y; A attributs protégés ; x inputs (hors attribut protégé)

Tableau 2 : Principales métriques de la discrimination indirecte [d'après Wachter & al ; 2021, p. 49-50]

Abu Elyounes estime que, au regard du droit, la *group fairness* vise à améliorer la position des groupes défavorisés et à atteindre une égalité réelle entre eux, sans s'en tenir à une simple égalité procédurale ou formelle. En effet, l'impact disproportionné survient en présence d'une règle ou d'une pratique formellement non discriminatoire et basée sur le principe procédural de traitement égal des personnes placées dans des situations similaires. Partant de l'idée que la *fairness* est contextuelle – qu'il n'y a pas de « *one size fits all* » dans ce domaine, la démarche d'Abu Elyounes consiste à déterminer les conditions dans lesquelles les métriques de la discrimination indirecte ont un domaine de pertinence pour les approches juridiques (tableau 2).

Sub-notion	**Corresponding Legal Mechanism**	**Example of implementable case**
Decoupling	Affirmative action (as separate but equal)	When the minority group is very small and has unique characteristics like women in the criminal justice system
Statistical or conditional parity	Affirmative action (preferably through critical diversity)	Cases where affirmative action was approved by law like hiring and school admission
Equal opportunity	Affirmative action (via equality of opportunity)	When fixing on the outcome is sufficient and does not require fixing the process that led to this outcome
Equalized odds	Achieving equality by equalizing the false positive and false negative errors	When it is possible to achieve the right balance between the two types of errors
Calibration	Achieving equality by statistical significance	High stake cases that society is willing to give up on equalizing the error rates
Multicalibration	Achieving equality by statistical significance, and accounting for intersectionality	Pretrial, but it should be applied cautiously since it is a new notion

Tableau 3 : Concepts de *Group Fairness* et mécanismes juridiques correspondants aux Etats-Unis [extrait de Abu Elyounes, 2020]

Un exemple concret des enjeux que posent le choix d'une métrique parmi d'autres est celui que la controverse sur l'application prédictive de récidive en matière pénale COMPAS, qui a été critiquée comme racialement discriminatoire par ProPublica. : l'algorithme classifie à tort les afro-américains comme futurs criminels à un taux deux fois plus élevé que pour les Blancs ; les Blancs sont classés à tort comme non récidivistes à un taux disproportionné par rapport aux Noirs (Angwin & al., 2016 ; Larson & al, 2016) ; d'autres auteurs (Chouldechova 2016 , Corbett-Davies & al.,2016 ; Rahman, 2020) estiment que le modèle de COMPAS est calibré, en montrant que les risques de récidive des afro-américains et des caucasiens sont égaux pour tous les scores de risque de récidive. A cet égard, Abu Elyounes (2020) a raison de considérer que Propublica retient le critère d'égalité des chances, alors que COMPAS soutient que son algorithme est équitable parce que calibré.

2.2.3. Des algorithmes conformes au droit ?

Wachter, Mittelstadt et Russell (2020) font le constat qu'il existe un écart, voire une incompatibilité, entre les notions juridiques européennes de discrimination et les travaux existants sur la *fairness* algorithmique et font des propositions pour réduire ce *gap*.

Du côté du droit anti-discrimination au sein de l'UE, les directives « égalité » sont rédigées avec un niveau élevé de généralité ; les Etats-membres doivent les transposer en droit interne, et ils le font avec un niveau de généralité ou de précision (notamment sur les

conditions de la preuve statistique) très variables ; les décisions de la CJUE ne sont pas constantes au cours du temps sur les conditions de la preuve d'une discrimination indirecte bien qu'un « gold standard » puisse y être trouvé. La régulation juridique donne donc de la souplesse (via la transposition des directives et la jurisprudence de la CJUE, souvent marquée par un raisonnement intuitif plutôt que sur une métrique définie et stable[8]) alors que la recherche sur la *fairness* algorithmique repose sur une recherche de précision et de cohérence. D'un côté, « *European conceptualisation of discrimination […] is contextual* » (Wachter & al. 2018, p. 5), alors que de l'autre les méthodes automatiques visant à détecter et combattre les décisions discriminatoires ont besoin de règles claires.

L'objectif de Wachter & al. est de clarifier comment construire des considérations de *fairness* dans la décision automatique, qui respectent autant que possible l'approche contextuelle du droit européen, notamment de la CJUE. Dans ce dernier, il est cependant possible de trouver un « gold standard », posé dans la décision *Seymour-Smith* de la CJCE, qui consiste à penser que la comparaison entre le groupe discriminé et le groupe non discriminé est la méthode la meilleure. Pour Wachter & al., la seule métrique de *fairness* compatible avec ce principe est la (dis)parité démographique (conditionnelle), formalisée ainsi (pour un attribut donné) :

$$A = \frac{No.of\ protected\ people\ in\ the\ advantaged\ group}{Total\ No.of\ people\ in\ the\ advantaged\ group}$$

$$D = \frac{No.of\ protected\ people\ in\ the\ disadvantaged\ group}{Total\ No.of\ people\ in\ the\ disadvantaged\ group}$$

(Wachter & al, 2020, p. 51)

Dès lors que D > A, il y a une disparité démographique au détriment du groupe désavantagé.

Wachter & al. ajoutent un autre test, dit de « Negative dominance », qui survient si D > 50% > A : ce test n'existe pas dans la littérature sur la *fairness* algorithmique. Le test se fait en deux temps: a) la majorité du groupe désavantagé ne doit pas appartenir à la classe protégée, b) seule une minorité de la classe protégée appartient au groupe désavantagé.

Dans une autre contribution, Wachter & al. (2021) approfondissent le rôle que le droit anti-discrimination peut jouer pour contribuer à définir une métrique conforme à ses visées, à savoir promouvoir l'égalité substantielle plutôt que procédurale et formelle. Ils examinent la compatibilité entre les métriques de *fairness* utilisées en apprentissage machine et les finalités du droit européen anti-discrimination. Ce dernier vise, au-delà de la prévention des discriminations, à changer la société, les politiques publiques et les pratiques pour « *level the playing field »* et réaliser l'égalité substantielle. Alors qu'en apprentissage machine il s'agit de penser des techniques de *fairness* qui règlent des problèmes de biais dans les

[8]CJUE 19 avril 2012, aff. C 415-10, Galina Meister contre Speech Design Carrier Systems GmbH:"indirect discrimination may be established by any means, and not only on the basis of statistical evidence".

données d'apprentissage, sachant que ces données reflètent des stratifications sociales historiques, ces techniques sont qualifiables de *« bias preserving* » ; à l'inverse, la finalité du droit est de transformer ces biais sociétaux historiques (« *bias transforming* »)[9]. Techniquement, les *equalized odds*, *equal opportunity* et calibration et d'autres relèvent d'une orientation « *bias preserving* » (Wachter & al, 2021, p. 26). Ils estiment que sur les 20 métriques de *fairness* existant dans la littérature, 13 d'entre elles sont *bias preserving* dans la mesure où elles sont satisfaites par des « *matching error rates between groups* » alors que les 7 métriques[10] '*bias transforming*' sont satisfaites par des « *matching decision rates between groups* » (2021, p. 29). Comme les *bias preserving metrics* portent le risque de consolider des situations d'injustice sociale et de discrimination, donc de maintenir un *statu quo*, elles ne sont pas en phase avec le cœur du droit européen anti-discrimination, qui vise à réaliser une égalité substantielle (2021, p. 31). Wachter & al. (2020, 2021) recommandent de recourir à la Conditional Demographic Disparity (CDD) qu'ils estiment compatible avec les finalités du droit européen et sont de ce fait '*bias transforming'* : « *CDD treats all people (groups) as equal, meaning they should be treated the same. The test flags up any disparity between groups that remains once an appropriate conditioning variable has been applied. This notion of fairness follows the Aristotelian postulate of treating 'like cases alike' and enables formal equality. At the same time, CDD enables substantive equality by flagging up … any relative disparity between groups in a given population over a set of decisions or other outcomes*" (Wachter & al, 2021, p. 43).

Xiang (2021) formule des propositions sur les moyens de réconcilier les approches techniques et juridiques du biais algorithmique, en partant du constat que la diffusion de modèles de décision automatique suscite un risque sérieux d'accentuation des inégalités sociales. Elle analyse les approches techniques susceptibles d'être conformes au droit américain anti-discrimination et de réduire le risque de voir des algorithmes qui accentuent les inégalités être considérés comme conformes au droit. Xiang argumente en faveur de l'inférence causale, conforme à la jurisprudence *Texas Dpt of Housing and Community Affairs v. Inclusive Communities Project, Inc*, qui pose le principe d'une « connection causale» entre le processus de prise de décision et l'existence d'un impact disproportionné. Pour Xiang l'inférence causale suppose le recours à un contrefactuel puisqu'il s'agit de « *comparing what happened in the real world with what would have happened in a counterfactual world with different conditions* » (Xiang, 2021, p. 61).

2.2.4. Le droit confronté à la discrimination algorithmique

La discrimination algorithmique suscite des interrogations quant à la capacité des victimes à recourir au tribunal pour contester le traitement différentiel dont elles sont l'objet. L'accès aux caractéristiques du traitement algorithmique, voire la connaissance de l'existence d'un

[9] 'Bias preserving' fairness metrics seek to reproduce historic performance in the outputs of the target model with equivalent error rates for each group as reflected in the training data (or status quo). In contrast, 'bias transforming' metrics do not blindly accept social bias as given or neutral starting point that should be preserved, but instead require people to make an explicit decision as to which biases the system should exhibit (Wachter & al., 2021, p. 25).

[10] A savoir : statistical parity, conditional statistical parity, fairness through unawareness, fairness through awareness, counterfactual fairness, no unresolved discrimination, no proxy discrimination, path base causal reasoning.

processus de décision automatique, n'est pas assuré pour les victimes. Hacker (2018) pose le problème de l'effectivité limitée du droit anti-discrimination en Europe en présence d'une discrimination algorithmique. Selon lui, le problème peut être résolu au prix d'un décloisonnement du droit anti-discrimination et de la protection des données personnelles telle que prévue dans le RGPD ; ce décloisonnement est propice à une coopération fructueuse entre juristes et informaticiens susceptible de faire émerger des algorithmes assurant une « *equal protection by design* » (Hacker, 2018, p. 25). La mobilisation du RGPD peut déplacer le curseur en amont du contentieux, c'est-à-dire permettre dans une certaine mesure de susciter la conception de règles de décision qui préviennent les situations de discrimination. Ainsi, le levier de la protection des données personnelles peut avoir des effets bénéfiques en termes antidiscriminatoires, sans qu'il soit besoin de renforcer le droit relatif à la protection des données. Les effets bénéfiques attendus peuvent être réalisés à droit constant, c'est-à-dire sans qu'il soit besoin de modifier le RGPD (Hacker, 2018, p. 25).

Le RGPD contient des considérants qui posent des principes essentiels, dont le n° 39 qui pose que tout traitement de données à caractère personnel devrait être licite et loyal. ; et surtout le n° 71 qui pose que « le responsable du traitement devrait utiliser des procédures mathématiques ou statistiques et appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte… que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le risques d'erreur soit réduit au minimum…et qui prévienne, entre autres, les effets discriminatoires… » (nous soulignons).

Le RGPD contient en effet plusieurs instruments qui peuvent susciter ces effets bénéfiques : le droit d'accès individuel aux données ; les *data protection impact assessments* (DPIAs) et audits qui avaient été proposés par l'article 29 des *Working Party*[11] *Guidelines on Automated Individual Decision Making and Profiling*. Les guidelines proposaient, entre autres : un dispositif de "algorithmic auditing", consistant à "*testing the algorithms used and developed by machine learning systems to prove that they are actually performing as intended, and not producing discriminatory, erroneous or unjustified results*" ; ces propositions ont été en partie intégrées dans le considérant n° 71 et dans l'article 22 du RGPD.

Hacker estime que le RGPD incorpore des principes essentiels – tels que les données doivent être licites, loyales, précises et la prévention des effets discriminatoires – qui sont une base importante d'une *algorithmic fairness*. Les autres éléments que recèle le RGPD sont les dispositions de l'article 15 (droits d'accès aux données) et de l'article 22 (décision individuelle automatisée) sur le fond ; mais Hacker considère que ces dispositions, notamment l'article 15 (1)(h)[12] doivent être complétées par un « *public enforcement that aims at uncovering the right metrics and exact causes of discrimination* » (Hacker, 2018, p. 27). Le *public enforcement* renvoie aux articles 83 (conditions générales pour imposer des amendes administratives) et 58 (pouvoirs des autorités de contrôle) ainsi qu'à l'article 35 (analyse d'impact relative à la protection des données). Ainsi, « *National data protection*

[11] The Article 29 Working Party is an independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018 (entry into application of the GDPR)

[12] RGPD, article 15(1)(h) : L'individu a droit à être informé de l'existence d'une prise de décision automatisée, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée.

authorities should moreover make use of algorithmic audits and data protection impact assessments, according to Article 58(1)(b) and Article 35 et seq. GDPR, to uncover the causes of bias and to enforce adequate metrics of algorithmic fairness » (Hacker, 2018, p. 35).

Les développements qui suivent mettent l'accent sur l'explicabilité, en confrontant les modèles juridiques et d'apprentissage-machine, autour de l'idée d'un gap entre les perspectives qu'ils poursuivent.

3. L'explicabilité des décisions algorithmiques : la nécessité d'une clarification juridico-technique

Appliquée dans le domaine des algorithmes la notion d'explicabilité est une notion juridique et technique encore en construction. Cette notion est au cœur d'un champ important de l'IA connu sous le vocable *Explainable AI* (XAI), initié il y a quarante ans[13] mais elle prend une plus grande importance avec le développement des décisions algorithmiques.

Dans le langage courant le terme explicabilité désigne la capacité à expliquer, à rendre intelligible ou compréhensible. D'emblée ce terme comporte plusieurs dimensions. Il s'agit d'expliquer le pourquoi et/ou le comment de quelque chose. Le résultat de cette explicabilité sera de fournir des explications. Parler d'explicabilité implique nécessairement d'en délimiter l'objet, de prendre en compte ses destinataires, et de déterminer qui devra livrer ces explications, voire sous quelles formes. Le droit européen et le droit français insistent sur le fait que les décisions algorithmiques doivent être explicables, cependant sans en donner clairement une définition ni préciser ses modalités. En IA, l'explicabilité est différenciée de l'interprétabilité, ce qui introduit des troubles dans la compréhension de la signification des notions dans les différents champs.

L'objet de cette partie est de dresser un inventaire croisé de l'état du droit et de la technique en matière d'explicabilité. Une des premières sources de difficultés pour mener un tel dialogue réside dans la nécessité de s'entendre sur un vocabulaire commun, sachant en outre qu'à l'intérieur de chaque discipline, il n'existe pas d'unanimité sur la signification de telles ou telles terminologies. Aussi, il apparaît nécessaire de confronter l'approche juridique à celle de la technique en dressant l'inventaire de l'état de la recherche en IA sur l'explicabilité. L'objectif est de contribuer à mieux saisir comment l'explicabilité pourrait prévenir et lutter contre les biais, donc être un moyen de la loyauté des algorithmes.

3.1. L'explicabilité : une notion sollicitée mais pas définie par le législateur

À notre connaissance, il n'existe pas de définition générale et abstraite de la notion d'explicabilité par le législateur que cela soit en droit européen ou français. Bien plus, ce terme n'a été que très récemment consacré explicitement en droit français à l'article art. 17

[13] L'apparition de ce terme est souvent associée à deux articles : Scott, Clancey, Davis, Shortliffe (1977) et Xu, Uszkoreit, Du, Fan, Zhao, and Zhu (1981).

de la Loi sur la bioéthique du 2 août 2021. Ni la Loi République numérique (LRN), ni la Loi Informatique et Libertés (LIL) ne comportent une telle référence.[14]

Cela étant, il existe en droit européen et en droit américain des obligations juridiques d'explicabilité essentiellement liées aux décisions algorithmiques. Ainsi, aux États-Unis, un droit constitutionnel d'explicabilité a été consacré par la jurisprudence dans deux affaires en particulier. La première concerne un algorithme établissant un classement des enseignants. Le juge a considéré que "*without access to value-added equations, computer source codes, decision rules, and assumptions, teachers could not exercise their constitutionally-protected rights to due process*"[15]. La seconde affaire concerne l'algorithme de justice prédictive COMPAS. Si la Cour du Wisconsin[16] n'a pas exigé la diffusion du code source de l'algorithme de COMPAS elle a consacré une obligation d'explicabilité globale pour permettre aux juges de mieux évaluer l'exactitude et le poids à accorder au score de risque[17]. Il existe également une obligation d'explicabilité dans le domaine bancaire pour l'attribution des prêts. La banque doit ainsi donner les raisons spécifiques du refus de prêts[18]. Enfin, certains États fédérés ont aussi introduit des obligations d'explicabilité ; tel est le cas de la loi de l'État de Washington sur la reconnaissance faciale[19].

Dans l'UE, il existe également plusieurs obligations d'explicabilité, la plus connue ayant été introduite à l'article 22 du RGPD[20]. Il est important de souligner que le champ d'application de cette obligation est doublement limité : d'abord aux systèmes qui utilisent des données personnelles – les données non personnelles ne sont pas régies par le RGPD ; ensuite aux décisions individuelles automatisées, y compris le profilage.

3.2. Regards croisés sur les notions : décision, profilage, explicabilité, interprétabilité

Du point de vue juridique, l'explicabilité se rapporte essentiellement à une décision dite algorithmique. Il faut néanmoins d'emblée souligner que les notions de décision et de prise de décisions ont un sens différent en informatique. Il en est de même avec le terme « d'explicabilité ». Il est alors utile de confronter les notions en droit et en informatique.

---

[14] Pour autant, en droit français, ce terme est communément utilisé notamment en lien avec : les articles. L-311-3-1 et R311-3-1-1 du code des relations entre le public et l'administration ainsi que l'article 47 de la Loi Informatique et Liberté introduit pour mettre en application l'article 22 du RGPD.

[15] *Local 2415 v. Houston Independent School District*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017), p. 17.

[16] Cf. *State of Wisconsin v. Loomis*, Supreme Court of Wisconsin, n° 2015AP157-CR, July 13, 2016.

[17] En ce sens cf. Beaudouin & al (2020a,b).

[18] Cette obligation est comprise dans le Fair Credit Reporting Act (2018).

[19] Voir https://www.dlapiper.com/en/us/insights/publications/2020/04/in-washington-states-landmark-facial-recognition-law-public-sector-practices-come-under-scrutiny/

[20] Sur cette disposition et les différences de formulation avec l'article de la Convention 109+ du Conseil de l'Europe, Olivia Tambou : (2020, p. 209 et s.).

3.2.1. Décision et processus de décision

En informatique une décision est la réponse d'un algorithme (implémenté sous forme d'un logiciel) à une requête.

Du point de vue du formalisme mathématique, une décision est la partition d'un ensemble donné. Une partition consiste dans la séparation de l'ensemble en sous-ensembles avec une intersection vide et dont l'union constitue l'ensemble. Chaque sous-ensemble est appelé une "classe d'équivalence". Si ces classes sont ordonnées et définies en référence à une norme nous appelons la décision un "rating". Si les classes sont ordonnées mais non définies par rapport à une norme nous appelons la décision un "*ranking*". Si les classes ne sont pas ordonnées, mais sont définies en référence à une norme nous appelons la décision un "*assignment*" ; enfin, si les classes ne sont pas ordonnées et ne sont pas définies par rapport à une norme nous appelons la décision un "*clustering*".

Les activités d'une entité (humaine ou artificielle) qui mènent à "une décision" sont un processus de décision. De ce point de vue l'activité d'un algorithme qui doit calculer pour un ensemble d'objets un score (de quelque chose) est un processus de décision. In fine, l'algorithme prend une décision (le résultat de ces calculs).

Cependant, dans la plupart des cas, cette activité fait partie d'un processus plus large qui consiste dans l'aide à une autre entité qui, à son tour, doit prendre une décision. Nous appelons cet ensemble d'activités un processus d'aide à la décision. Les décisions prises dans le cadre d'un processus d'aide à la décision sont des "recommandations".

Il est possible que les processus d'aide à la décision soient utilisés dans un processus de décision d'une autre entité. On peut aller jusqu'à un entrelacement de processus de décision et de processus d'aide à la décision qui produisent une cascade de recommandations ... jusqu'à une prise de décision que nous appellerons "finale" ... La notion de responsabilité s'applique en général à cette "décision finale" ... même si en réalité toute décision (mathématiquement définie) peut être considérée comme porteuse de responsabilités. Par exemple, lorsqu'un employé de banque reçoit une demande de crédit, un algorithme calcule le "*credit score*" du demandeur. Dans un tel processus de décision, l'algorithme "décide" le "*credit score*" (décision) et le transmet à l'employé (recommandation) qui à son tour décide d'accorder ou de refuser le crédit (décision finale). C'est essentiellement à cette « décision finale » que s'attache le droit relatif à l'explicabilité des algorithmes.

3.2.2. Profilage

Du point de vue informatique, le profilage est un problème de décision de classification (*clustering decision*) dans lequel les individus, caractérisés par un ensemble d'attributs "externes" (par exemple démographiques) et un ensemble d'attributs comportementaux (tels que les préférences de consommation) sont regroupés dans des classes similaires. Dans le cas où les classes de comportement potentiel sont prédéfinies, nous avons un problème d'"affectation". Dans un premier cas, nous ne savons pas a priori quelles sont les classes d'équivalence qui définissent la correspondance entre les attributs externes et comportementaux et le *clustering* permet de les "découvrir". Dans un second cas, nous

avons une hypothèse sur le nombre de modèles comportementaux possibles au sein de la population observée et nous la testons. Le résultat d'un exercice de profilage est que tous les individus regroupés ou affectés au même profil (modèle de comportement) sont supposés se comporter de manière similaire et, dans cette perspective, peuvent être la cible d'actions visant à influencer ou à modifier ce modèle de comportement.

La notion de décision individuelle automatisée n'a pas été définie dans le RGPD[21] alors que la notion de profilage est définie comme une forme de traitement automatisé visant à évaluer une personne physique, que cela soit pour analyser ou prédire son rendement au travail, sa situation économique son comportement etc.[22]. Dès lors, la notion juridique de décision automatisée ne recoupe pas la distinction faite en IA entre les décisions automatisées (« *automated* »), les décisions autonomes (« *autonomous* ») et les décisions algorithmiques (« *algorithmic decision-making* »[23]. En IA la notion de décision automatisée signifie que la décision a été prise sur la base d'une série d'actions prédéfinies précises sans intervention ultérieure d'un humain. Ces décisions seraient facilement prédictibles. La notion de décision autonome implique que seuls les objectifs généraux ont été établis par un humain, laissant à la machine le soin de déterminer comment y parvenir. Or, bien que se référant au terme de décision automatisée, l'article 22 du RGPD concerne du point de vue de l'IA les décisions algorithmiques. Cette notion plus générale signifie simplement qu'une décision a été prise avec l'aide d'un algorithme. Il s'agit donc d'une notion qui englobe les décisions autonomes et automatisées.

Il faut souligner que certaines décisions individuelles algorithmiques ne sont pas couvertes par l'article 22 : seules sont concernées celles qui sont exclusivement fondée sur un traitement automatisé et qui produisent des effets juridiques concernant la personne ou l'affectent de manière significative de façon similaire. Cette dernière expression fait débat chez les juristes. En effet, l'interdiction de principe des traitements exclusivement automatisés ne concerne que des décisions qui ont une incidence grave. Cela va de soi lorsque la décision a des effets juridiques comme l'annulation d'un contrat, le droit ou le refus d'un avantage social accordée par la loi comme une allocation familiale, ou une allocation logement, le refus d'admission dans un pays etc. En revanche, il n'est pas toujours aisé de déterminer ce qu'est une décision automatisée qui sans avoir d'effet juridique affecte la personne concernée de « manière significative de façon similaire ». Le considérant n° 71 du RGPD évoque simplement « le rejet automatique d'une demande de crédit en ligne ou des pratiques de recrutement en ligne sans aucune intervention humaine».[24] Cela couvre

---

[21] Les autorités de protection des données européennes considèrent que « La prise de décision exclusivement automatisée est la capacité de prendre des décisions par des moyens technologiques sans intervention humaine. », cf. Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, version révisée du 6 février 2018, WP251rev.01, p. 8.

[22] Cf. article 4 du RGPD.

[23] Sur l'ensemble de cette question cf. Brkan, M., & Bonnet, G. (2020).

[24] Sur les difficultés de l'interprétation de cette notion d'affectation de manière significative de façon similaire à un effet juridique cf. Lignes directrices relatives à la prise de décision individuelle automatisée et au profilage aux fins du règlement (UE) 2016/679, version révisée du 6 février 2018 du précitées p. 23 et s. Notons que la France n'a pas repris une formule

donc la perte d'une chance, d'une opportunité. Cela concerne ainsi des décisions concernant l'accès à un service dans le domaine de la santé ou de l'éducation.

3.2.3. Explicabilité et interprétabilité

Pour certains auteurs « explicabilité » et « interprétabilité » sont des notions interchangeables (Beaudouin et al., 2020a, p. 8) alors que d'autres établissent une distinction. L'existence d'une terminologie duale se retrouve dans la doctrine anglo-saxonne à travers tantôt l'opposition, tantôt l'assimilation entre « *explainability* » et « *interpretability* ». Le point commun entre ces deux appellations est que dans les deux cas il s'agit bien de rendre compréhensible les décisions prises par un algorithme. Néanmoins, dans le domaine de l'IA, des auteurs considèrent que l'interprétabilité vise à évaluer globalement le processus d'une décision, c'est-à-dire, en réalité, à rendre compréhensible le modèle utilisé. Autrement dit, l'interprétation répond à la question de : comment un algorithme prend une décision d'une manière générale.

Ainsi, nous pouvons distinguer trois niveaux qui peuvent techniquement être considérés comme des explications pour un algorithme donné et son exécution.

1. Le premier (**descriptif**) consiste à **reconstruire et à retracer chaque étape de l'algorithme**, depuis la réception d'une entrée jusqu'à la réalisation de la sortie prévue (les informaticiens recueillent ces informations dans ce qu'ils appellent un fichier .log). Il s'agit essentiellement d'une explication descriptive de ce que l'algorithme a fait ou fait et, la plupart du temps, elle ne peut être interprétée que par des spécialistes du codage.
2. Le second (**logique**) consiste à **reconstruire les raisons pour lesquelles un algorithme effectue une étape donnée, pendant son exécution**. Il s'agit essentiellement d'une explication logique par laquelle chaque étape peut être considérée comme une relation causale. De telles explications sont utiles pour montrer que la sortie de l'algorithme est logiquement liée à l'entrée. La vérification formelle des algorithmes et la vérification de la sécurité utilisent ce type d'explication afin de démontrer qu'un algorithme donné fait effectivement ce que les spécifications attendaient et que l'exécution est robuste, sûre, fiable, etc.
3. Le troisième (**argumentatif**) consiste à **fournir les raisons ultimes pour lesquelles un algorithme calcule une certaine sortie, étant donné une certaine entrée**. Ces raisons sont essentiellement de deux types : les données fournies et la procédure utilisée. Un exemple typique est le résultat d'une procédure de vote : le résultat dépend des bulletins déposés par les votants, mais aussi de la façon dont la majorité est calculée avec cette procédure particulière. Une explication argumentée doit toujours fournir ces deux éléments. Ce type d'explication devient problématique lorsque l'algorithme modifie l'exécution à chaque fois en fonction des connaissances accumulées lors des exécutions précédentes ou lorsque l'algorithme comprend des composants pour lesquels il est pratiquement impossible d'établir un lien logique entre l'entrée et la sortie (boîtes noires).

---

complètement identique à celle de l'article 22 laissant de côté le terme « similaire » Pour autant, seule une interprétation conforme au RGPD serait valide.

Pour Besse & al. (2018, p. 25), une règle de décision est interprétable si on comprend comment elle associe une réponse à des observations (par exemple : un arbre de décision) ; elle est explicable si on comprend sur quels éléments est basée la décision. Pour Stoica & al. (2017), l'interprétabilité "*means that the output of the AI algorithm is understandable to a subject matter expert in terms of concepts from the domain from which the data are drawn*" alors que l'explicabilité "*means that one can identify the properties of the input to the AI algorithm that are responsible for the particular output and can answer counterfactual or 'what-if' questions*".

En revanche, l'explicabilité va plus loin car il s'agit de préciser dans un cas concret quelles sont les variables précises qui ont été déterminantes pour prendre une décision particulière. Substituant la distinction sémantique entre interprétabilité et explicabilité, certains auteurs en apprentissage machine (Guidotti & al, 2018) opposent l'explicabilité globale à l'explicabilité locale, appelée aussi « post hoc ». L'explicabilité globale vise à expliquer l'algorithme dans son entier, alors que l'explicabilité locale est la capacité à expliquer une décision algorithmique particulière. On peut également distinguer entre « *post-hoc explanation* » et « *build-interpretable models* » (Barale, 2021).

Les explications *post-hoc* sont construites par l'analyse statistique après que les données ont été révélées (et l'hypothèse testée). Elles peuvent être utilisées positivement pour approfondir la compréhension d'un résultat donné, mais peuvent aussi contribuer négativement à la construction de corrélations sans signification. Les modèles interprétables par construction sont des modèles construits de manière que les explications soient calculées (ou collectées) pendant leur utilisation. Cela peut améliorer l'efficacité du calcul des explications, mais présente deux faiblesses : cela peut produire des interprétations hors contexte sans signification et cela peut cacher d'autres explications possibles pour lesquelles aucune disposition n'a été prise dans les spécifications.

Ces variations sémantiques ne se retrouvent pas réellement dans la doctrine juridique. La plupart des documents institutionnels et la doctrine utilisent le terme d'explicabilité parfois d'intelligibilité[25]. Pour autant, il est admis qu'il existe également deux niveaux d'exigences d'explicabilité. D'une part, il existe une explicabilité « ex ante » qui a pour but d'informer l'individu sur la logique de l'algorithme et qui est étroitement reliée à un droit à l'information. D'autre, part, il existe une explicabilité « ex post », qui a pour but d'expliquer à un individu pourquoi un algorithme a pris une décision précise. Le débat dans la sphère juridique est de savoir dans quelle mesure la seconde forme d'explicabilité constitue une obligation légale à l'échelle européenne, comme nous le verrons ultérieurement.

Concept	**Informatique**	**Droit**
Décision	Partition d'un ensemble donné	Le RGPD ne définit pas la décision individuelle automatisée
	Sous-ensemble : classe	Profilage art. 4 RGPD : finalité du traitement qui

[25] Cf. CNIL, Comment permettre à l'homme de garder la main ? rapport précité notamment p. 53

	d'équivalence	caractérise le profilage
Décision automatisée	Suit une série prédéfinie d'actions sans action humaine	RGPD : interdiction des décisions individuelles automatisées ayant une incidence grave - Inclut aussi la décision autonome au sens des informaticiens
Décision autonome	Définition des objectifs généraux par un humain. Apprentissage	Concept non utilisé
Explicabilité	Interprétabilité Logique interne de l'algorithme : association observation/résultat	Explicabilité ex ante : droit à l'information sur l'existence d'une décision algorithmique Explicabilité ex post : pourquoi l'algorithme a pris une décision précise

Tableau 4. Comparaison des concepts [source : auteurs]

3.3. Une notion récente concrétisant l'État de droit à l'ère numérique

Les préoccupations ou la raison d'être juridique du droit de l'explicabilité des décisions algorithmiques est pour partie ancrée dans ce que l'on pourrait appeler le méta-principe juridique de l'État de droit. A ce stade, il suffit de souligner que l'État de droit implique que toute décision juridique doit être en principe, prévisible, fondée sur un processus transparent, motivée pour pouvoir être ensuite éventuellement contestée notamment devant le juge.

L'État de droit vise à garantir que « toutes les autorités publiques agissent toujours dans les limites fixées par la loi, conformément aux valeurs de la démocratie et aux droits fondamentaux, et sous le contrôle de juridictions indépendantes et impartiales. L'État de droit est une notion qui recouvre des principes tels que la légalité, qui suppose l'existence d'une procédure d'adoption des textes de loi transparente, responsable, démocratique et pluraliste; la sécurité juridique; l'interdiction de l'arbitraire du pouvoir exécutif; une protection juridictionnelle effective assurée par des juridictions indépendantes et impartiales, un contrôle juridictionnel effectif y compris le respect des droits fondamentaux; la séparation des pouvoirs et l'égalité devant la loi »[26]. L'État de droit est une notion juridique cardinale. Aujourd'hui, le respect de l'État de droit est évalué sur la base de critères permettant d'attester de la prééminence du droit dans les États européens tant au sein du Conseil de l'Europe[27] que de l'Union européenne.

Ce lien entre État de droit et droit algorithmique a été abordé par la doctrine juridique. Certains auteurs évoquent l'apparition d'une nouvelle forme de « normativité algorithmique » (Barraud, 2018), voire de « gouvernance algorithmique » (Rouvroy et Berns, 2013). Ces expressions sont utilisées tout autant pour affirmer la nécessaire régulation des algorithmes en particulier ceux qui sont utilisés par les autorités publiques que pour

[26] Définition tirée du Premier Rapport de la Commission Européenne sur l'État de droit, 20 septembre 2020, COM(2020) 580 final, p. 1.

[27] Au sein du Conseil de l'Europe, la Commission de Venise qui joue un rôle fondamental dans l'affirmation et la mise en œuvre de l'État. Cf. sa grille de critères adoptée en 2016 https://www.venice.coe.int/webforms/documents/?pdf=CDL-AD(2016)007-f

constater la dilution du pouvoir de régulation des États face aux acteurs privés essentiellement américains pour l'instant voire chinois. Le droit de l'explicabilité est alors présenté comme un viatique vers la reprise du contrôle de l'homme sur la Machine. L'explicabilité a ainsi une résonnance avec le concept plus large de souveraineté numérique susceptible de se décliner tant à l'échelle individuelle[28] que collective.

Du point de vue juridique ce qui importe c'est d'aligner les décisions algorithmiques sur les standards de l'État de droit relatif aux décisions juridiques classiques afin de maintenir le modèle démocratique. Du point de vue de l'IA, la quête va au-delà d'expliquer comment répondre aux exigences légales de fournir des explications aux décisions algorithmiques. Il s'agit de mener des recherches sur le « pourquoi faire » de l'usage des potentialités techniques de l'intelligence artificielle. Autrement dit, la recherche en apprentissage machine comporte une dimension éthique : "*Being able to explain an AI-based system may help to make algorithmic decisions more satisfying and acceptable, to better control and update AI-based systems in case of failure, to build more accurate models, and to discover new knowledge directly or indirectly*" (Brkan,& Bonnet, 2020). Ainsi, Adadi et Berrada (2018), identifient quatre raisons d'être principales pour le développement de la recherche en XIA : Justifier des résultats obtenus par des décisions algorithmiques, contrôler le comportement des systèmes, améliorer les modèles et gagner en connaissance.

3.4. Des approches normatives instables/floues de l'explicabilité

3.4.1. Approches juridiques : droit européen et droit français

Nous avons déjà évoqué le RGPD précédemment. Il importe de souligner ici qu'il fait l'objet d'interprétations diverses et contradictoires quant à la contrainte d'explicabilité. Du point de vue formel, seul le droit à obtenir des explication pour une décision individuelle algorithmique figure dans le considérant (n° 71), et non dans le règlement lui-même. Aussi certains auteurs affirment qu'il n'existe pas un droit à l'explicabilité dans le RGPD. (Wacher, Mittelsadt & Floridi, 2017). D'autres soutiennent que c'est dans une lecture d'ensemble, qui articule différentes dispositions, que l'on peut trouver une contrainte d'explicabilité dans le RGPD (Brkan & Bonnet, 2020 ; Hacker, 2018). Pour d'autres enfin, l'explicabilité prévue dans le RGPD n'est pas la solution. (Edwards & Veale, 2017).

Outre le RGPD, le règlement européen 2019/1150[29] impose une obligation d'explicabilité aux plateformes en particulier les intermédiaires en ligne et les moteurs en ce qui concerne leurs classements (Article 5). Ces obligations ont vocation à être complétées par d'autres textes en cours d'élaboration. Ainsi, une série de nouvelles obligations de loyauté et de transparence des algorithmes utilisés par les plateformes numériques sont au cœur de l'adoption des règlements sur la législation des services numériques[30] et sur la législation

[28] Il existe d'ailleurs un courant doctrinal qui utilise le terme de "souveraineté cognitive" Cf. Lee A. Bygrave, 2020, p. 9.

[29] Règlement (UE) 2019/1150 du Parlement européen et du Conseil du 20 juin 2019 promouvant l'équité et la transparence pour les entreprises utilisatrices de services d'intermédiation en ligne (Texte présentant de l'intérêt pour l'EEE), JOUE L 186 du 11 juillet 2019

[30] Proposition de règlement UE relatif à un marché intérieur des services numériques (Législation sur les services numériques) et modifiant la directive 2000/31/CE, 15 décembre 2020, COM(2020), 825 final

des marchés numériques[31]. Ces deux textes comportent en germe de nouvelles formes d'explicabilité algorithmique autour de l'utilisation d'outils de modération algorithmiques de lutte contre la haine en ligne, ou encore de systèmes de recommandations.

Enfin, le règlement établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle)[32] et modifiant certains actes législatifs de l'union, va imposer des obligations d'explicabilité des systèmes d'IA modulées en fonction du risque, avec des obligations renforcées pour les systèmes d'IA à haut risque. La plupart de ces textes européens en cours d'adoption semblent surtout envisager une explication *ex ante* portant à la connaissance des personnes concernées l'utilisation d'algorithmes pour la prise de décision et la logique générale des algorithmes utilisés.

En France, la Loi sur la République Numérique (LRN) de 2016 constitue le point de départ de la réglementation française sur la transparence et l'explicabilité des décisions administratives algorithmiques. En effet, l'article 4 de la LRN a créé un nouvel article L-311-1-1 dans le CRPA selon lequel « *une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande* ». Il s'agit surtout d'une obligation de communication à la suite d'une demande. L'article R. 311-3-1-2 du CRPA en précise l'étendue :« *L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes : le degré et le mode de contribution du traitement algorithmique à la prise de décision ; les données traitées et leurs sources ; les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ; les opérations effectuées par le traitement* ».

Ici l'explicabilité est essentiellement donnée *a posteriori,* elle est non seulement globale mais aussi locale. Au-delà de l'explicabilité proprement dite, cet article comporte une obligation de transparence constituant une forme de droit à l'information avec l'obligation de mention explicite de l'existence d'une décision algorithmique. D'ailleurs, **une décision prise sur le seul fondement d'un traitement algorithmique ne comportant pas la mention explicite est considérée comme nulle**. Etalab, département de la direction interministérielle du numérique, chargé de mettre en œuvre la stratégie de l'État dans le domaine de la donnée a élaboré un guide pour accompagner les administrations dans leurs obligations d'explicabilité des algorithmes publics[33], au-delà de guides sur l'ouverture des codes sources[34]. Cette obligation générale de mention explicite inclus les traitements

[31] Proposition de règlement UE relatif aux marchés contestables et équitables dans le secteur numérique (législation sur les marchés numériques), 15 décembre 2020, COM (2020) 842 final

[32] Proposition de règlement UE établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union, 21 avril 2021, COM(2021) 206 final

[33] Cf. Expliquer les algorithmes publics, https://guides.etalab.gouv.fr/algorithmes/, dernière version consultée 13/07/2021

[34] Cf. Ouvrir les codes sources, https://guides.etalab.gouv.fr/pdf/guide-logiciels.pdf, dernière version consultée 13/07/2021

exclusivement automatisés et ceux qui constituent simplement une aide à la décision. Cela oblige les administrations à dresser un inventaire de leurs principaux traitements algorithmiques. Cette obligation de transparence s'inscrit plus largement dans l'idée que l'administration doit rendre compte de l'usage de ses algorithmes publics. Enfin, ces obligations concernent toutes les décisions administratives individuelles, qu'elles concernent les personnes physiques ou les personnes morales.

Ces obligations introduites dans la LRN viennent se compléter avec celles plus spécifiques liées aux décisions individuelles fondées sur des traitements automatisés comportant un traitement de données à caractère personnel issues du RGPD. Ainsi l'article 13 §2 f) RGPD inclus dans le droit à l'information dès lors qu'il existe une prise de décision individuelle automatisée. Il s'agit de donner « des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée ». Cette disposition semble, techniquement, correspondre à un principe d'interprétabilité (Besse & al., 2018). La doctrine juridique y voit au moins la reconnaissance d'une obligation globale d'explicabilité. De son côté, l'article 22 RGPD consacre le droit à ne pas faire l'objet d'une décision individuelle exclusivement automatisée tout en laissant aux États membres le soin sous réserve de certaines garanties d'autoriser de telles décisions algorithmiques. **C'est dans le cadre de l'exercice de sa marge de manœuvre que la France a adopté une base juridique spécifique afin d'autoriser des décisions administratives individuelles exclusivement automatisées**. La formulation choisie aboutit à un droit d'explicabilité étendu. En effet, selon l'article 47§ 2 de la LIL le responsable de traitement doit « *s'assurer de la maîtrise du traitement algorithmique et de ses évolutions afin de pouvoir l'expliquer, en détail et sous une forme intelligible, à la personne concernée la manière dont le traitement a été mis en œuvre à son égard* »[35]. Il s'agit bien là de ce que la doctrine appelle un droit d'explicabilité locale. Aussi, le débat sur l'existence ou non d'un droit d'explicabilité des décisions concrètes pris à l'égard d'un individu exclusivement par le biais d'un algorithme (Wachter, Mittelstadt, Floridi, 2017) n'a eu que peu d'écho dans la doctrine française. Tout au plus certains auteurs se sont-ils interrogés sur la capacité pour la France de pouvoir imposer un tel droit d'explicabilité locale, tout en restant dans les limites de sa marge de manœuvre au regard du RGPD[36].

En outre, d'utiles précisions ont été apportées par le **Conseil Constitutionnel.** D'une part, **il considère que « *ne peuvent être utilisés, comme fondement exclusif d'une décision administrative individuelle, des algorithmes susceptibles de réviser eux-mêmes les règles qu'ils appliquent, sans le contrôle et la validation du responsable du traitement* »[37] . Cette interprétation semble limiter la possibilité pour l'administration française d'utiliser des algorithmes dits d'auto-apprentissage** (« deep learning Machine »). D'autre part, le Conseil Constitutionnel rappelle que le devoir d'explication de l'administration à la demande de la personne concernée restreint aussi sa marge de manœuvre dans le choix de l'outil algorithmique. En effet, « *lorsque les principes de fonctionnement d'un algorithme ne peuvent être communiqués sans porter atteinte à l'un des secrets ou intérêts énoncés au 2°*

[35] Cf. art. 47§2 LIL.

[36] Cf. par exemple Castets-Renard (2018).

[37] Conseil Constitutionnel français, Décision n° 2018-765 DC du 12 juin 2018, point 71.

de l'article L. 311-5 du code des relations entre le public et l'administration, aucune décision individuelle ne peut être prise sur le fondement exclusif de cet algorithme »[38]. **Autrement dit, le recours à des algorithmes protégés par de droits de propriété intellectuelle semble exclu afin que l'administration puisse s'acquitter de son obligation de transparence**. Ces deux réserves d'interprétation tentent de lutter contre le phénomène de « *black box* ». C'est-à-dire l'impossibilité de comprendre les raisons exactes qui ont conduit un algorithme à prendre une décision algorithmique particulière. Cette approche illustre la volonté de la France de construire un modèle transparent de décisions purement algorithmiques qui puisse aussi inspirer les acteurs privés en les incitant indirectement à mettre en œuvre des garanties similaires.

Enfin, il convient de souligner que l'explicabilité dans le cadre de l'article 22 du RGPD ne peut être donnée qu'à l'individu concerné par la décision individuelle automatisée. En France, cette limite a été contournée par le Conseil Constitutionnel dans une décision opposant une Université à un syndicat qui souhaitait comprendre la logique des algorithmes utilisés par des Universités pour mettre en œuvre Parcoursup et permettre l'accès d'étudiants à leurs diplômes. Le Conseil Constitutionnel a considéré qu'un principe constitutionnel de droit d'accès à des documents administratifs permettait aux tiers d'obtenir « sous la forme d'un rapport, les critères en fonction desquels les candidatures ont été examinées et précisant, le cas échéant, dans quelle mesure des traitements algorithmiques ont été utilisés pour procéder à cet examen ». [39] On peut y voir selon notre terminologie la consécration d'un droit français d'explicabilité globale élargi à des tiers, dans le cas de décisions administratives fondées sur algorithmes publics. Cette possibilité ouvre la voie vers une auditabilité par la société civile.

L'article art. 17 de la Loi sur la bioéthique du 2 août 2021 introduit à l'article L-4001-3 du code de la santé publique consacre explicitement pour la première fois une obligation d'explicabilité des concepteurs de certains traitements algorithmiques pour les professionnels de santé qui les utilisent, pour un acte de prévention, de diagnostic ou de soin. Cet article prévoit également l'obligation d'informer le patient et de l'avertir, le cas échéant, de l'interprétation qui résulte du traitement de données algorithmiques.

Comme le rappelle le rapport n° 2243 de l'Assemblée nationale « premier article d'une loi de bioéthique à introduire la notion d'algorithme en lien avec l'intelligence artificielle dans le code de la santé publique, l'article 11 du projet de loi propose de compléter le chapitre I[er] du titre préliminaire du livre préliminaire de la quatrième partie du code de la santé publique, relatif aux professionnels de santé, avec un nouvel article L. 4001-3 destiné à clarifier les places respectives du patient, du professionnel de santé et de l'algorithme dans la relation de soin. »

[38] Cf. *Ibid*., point 70.

[39] Cf. Conseil Constitutionnel, DC- 2020-834 Union nationale des étudiants de France, QPC du 3 avril 2020, point 17. Ce principe constitutionnel a été consacré sur le fondement de l'article 15 de la Déclaration des Droits de l'Homme de 1789 selon lequel « La société a le droit de demander des compte à tout agent public de son administration ».

Pour autant, le projet de loi ne prévoyait qu'un alinéa 3 concernant la traçabilité des actions d'un traitement algorithmique[40]. Et le rapport de souligner « qu'il serait utile de poser, comme le proposent la CNIL[41] et le Conseil d'État[42)], une exigence d'« explicabilité », qui vise à permettre aux utilisateurs de ces systèmes d'intelligence artificielle d'en comprendre la logique générale de fonctionnement. Cette exigence suppose que leurs concepteurs mettent à disposition des utilisateurs les informations nécessaires à cette compréhension et que les professionnels de santé puissent apporter leurs compétences dès l'élaboration des algorithmes et des stratégies de collecte de données de santé qui les alimentent. Elle implique également que les professionnels de santé bénéficient d'une formation leur permettant de comprendre le fonctionnement de ces dispositifs afin d'en cerner les limites et de pouvoir expliciter au patient les fondements sur lesquels reposent les décisions médicales les concernant.

La formulation finale avec le terme d'explicabilité est issue de l'amendement n°1503 adopté l'Assemblée nationale en dernière lecture, le 9 juin 2021[43]. Dans son exposé sommaire, les auteurs, députés LRM justifient cet amendement de la manière suivante : « Par anticipation de la législation européenne sur l'intelligence artificielle, l'amendement propose de renforcer le principe de transparence ainsi que le devoir d'information en imposant aux concepteurs de traitements algorithmiques de garantir l'explicabilité de ces derniers. »[44]

Cette motivation atteste encore une fois la stratégie française qui va au-delà de la « sur transposition »[45] et pourrait s'appeler la préemption. Il s'agit d'anticiper dans le droit national des obligations futures européennes dans le but aussi d'influencer la conception du droit européen. La difficulté réside alors dans la détermination de l'ampleur légale des obligations au regard de la primauté du droit de l'UE. Autrement dit, dans quelle mesure l'État français en anticipant ou renforçant des exigences du droit de l'UE reste néanmoins en conformité avec celui-ci.

---

[40] Art. 11, III du projet de loi déposé le 24 juillet 2019 : « La traçabilité des actions d'un traitement mentionné au I et des données ayant été utilisées par celui- ci est assurée et les informations qui en résultent sont accessibles aux professionnels de santé concernés. »

[41] CNIL, Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, décembre 2017.

[42] Conseil d'État, Révision de la loi de bioéthique : quelles options pour demain ? 28 juin 2018

[43] Cf. art. 11 du projet de loi n°4281enregistré à l'Assemblée nationale le 24 juin 2021, https://www.assemblee-nationale.fr/dyn/15/textes/l15b4281_projet-loi# et le tableau de l'évolution de la loi accessible à http://www.senat.fr/tableau-historique/pjl19-063.html

[44] Cf. https://www.assemblee-nationale.fr/dyn/15/amendements/4222/AN/1503

[45] La « surtransposition » des directives européennes en droit interne. Un concept équivoque et instrumentalisé, Dalloz 2019 p. 2360-2361. Cf. également Projet de loi déposé en 2018 sur la suppression de sur-transposition de directives européennes, qui définit la surtransposition comme « toute mesure nationale de transposition instaurant une norme plus contraignante que celle qui résulterait de la stricte application de la directive, sans que cela ne soit justifié par un objectif national identifié".

3.4.2. Pourquoi le législateur n'a pas défini l'explicabilité

La notion de d'explicabilité est très répandue dans le monde académique tout comme dans le discours institutionnel tant à l'échelle européenne[46] que nationale[47]. L'absence de définition de l'expression d'explicabilité par le législateur se justifie car cette notion est trop floue et n'a été introduite en droit que sous forme contextualisée. Winston Maxwell (2020) précise à juste titre que l'explicabilité dépend de quatre facteurs importants :

« − Le destinataire de l'explicabilité, c'est-à-dire le public visé par l'explication. Son niveau sera différent selon qu'il soit utilisateur ou régulateur par exemple.

− Le niveau d'importance et d'impact de l'algorithme. L'explicabilité d'un accident d'une voiture autonome n'a pas le même degré d'importance que celle d'un algorithme de publicités ou de recommandations de vidéos.

− Le cadre légal et réglementaire, qui est différent selon les zones géographiques, comme en Europe avec le règlement général sur la protection des données (RGPD).

− L'environnement opérationnel de l'explicabilité, comme son caractère obligatoire pour certaines applications critiques, le besoin de certification avant le déploiement ou la facilitation d'utilisation par les usagers » (Maxwell, 2020, p. 14).

Ainsi, il n'existe non pas un droit général à l'explicabilité en droit du numérique, mais bien plus des obligations d'explicabilité introduites dans différents textes, comme nous l'avons précisé précédemment. Rappelons que la plupart nécessitent une intervention du pouvoir règlementaire complétée par des autorités de régulation qui tentent d'expliquer comment mettre en œuvre ce droit d'explicabilité.

Pour autant, cette notion est le plus souvent rattachée à une exigence de transparence.

3.4.3. Une notion récente rattachée le plus souvent au principe de la transparence

Le groupe d'experts de l'Union européenne[48] (ci-après GHEN) consacre l'explicabilité comme l'un des quatre principes éthiques, que doit respecter tout système d'IA, aux côtés du respect de l'autonomie humaine, de la prévention de toute atteinte aux êtres humains et

[46] Le terme explicabilité est employé dans de nombreux documents du Parlement européen et la Commission européenne y compris dans les lignes directrices du Groupe d'experts de haut niveau sur l'intelligence artificielle institué par la Commission européenne, Lignes directrices en matière d'éthique pour une IA digne de confiance, avril 2019.

[47] Cf. par exemple CNIL, Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle les enjeux éthiques, décembre 2017, p. 30, 50 ou 53. ou encore Défenseur des Droits en partenariat avec la CNIL, Algorithmes : prévenir l'automatisation des discriminations, 2020 p. 8 ; Commission Nationale consultative des droits de l'Homme, Avis sur la lutte contre la haine en ligne, 8 juillet 2021, p. 27

[48] Cf. Groupe d'experts de haut niveau sur l'intelligence artificielle institué par la Commission européenne, Lignes directrices en matière d'éthique pour une IA digne de confiance, avril 2019.. , accessible à https://op.europa.eu/o/opportal-service/download-handler?identifier=d3988569-0434-11ea-8c1f-01aa75ed71a1&format=pdf&language=fr&productionSystem=cellar&part=.

de l'équité.[49] D'une manière générale le GEHN rattache l'explicabilité à la transparence[50], à côté des principes de traçabilité, et de communication. Le GEHN considère que « L'explicabilité est essentielle pour renforcer et conserver la confiance des utilisateurs envers les systèmes d'IA. Cela signifie que les processus doivent être transparents, que les capacités et la finalité des systèmes d'IA doivent être communiquées ouvertement, et que les décisions – dans la mesure du possible – doivent pouvoir être expliquées aux personnes directement et indirectement concernées.

De même pour **l'OCDE,** transparence et explicabilité constituent le troisième des cinq principes qui fondent une approche responsable d'une IA susceptible de générer de la confiance. Les autres principes étant i) croissance inclusive, développement durable et bien-être ; ii) valeurs centrées sur l'humain et équité ; iv) robustesse, sûreté et sécurité ; et v) responsabilité. Dans le paragraphe 1.3. consacré à la Transparence et explicabilité[51], l'OCDE considère que :

« Les acteurs de l'IA devraient s'engager à assurer la transparence et une divulgation responsable des informations liées aux systèmes d'IA. À cet effet, ils devraient fournir des informations pertinentes, adaptées au contexte et à l'état de l'art, afin:

i. de favoriser une compréhension générale des systèmes d'IA,
ii. d'informer les parties prenantes de leurs interactions avec les systèmes d'IA, y compris dans la sphère professionnelle,
iii. de permettre aux personnes concernées par un système d'IA d'en appréhender le résultat, et,
iv. de permettre aux personnes subissant les effets néfastes d'un système d'IA de contester les résultats sur la base d'informations claires et facilement compréhensibles sur les facteurs, et sur la logique ayant servi à la formulation de prévisions, recommandations ou décisions. »

La transparence est conçue comme un principe plus vaste qui peut se satisfaire de rendre accessible des informations brutes ou rendues intelligibles, alors que l'explicabilité semble induire la nécessité de rendre compréhensibles les informations données[52]. Ainsi, l'ouverture des codes sources d'un algorithme est le plus souvent rattachée par les juristes au principe de transparence et plus indirectement à l'explicabilité car ce code est rarement compréhensible pour un profane. D'emblée, le principe d'explicabilité soulève une difficulté de délimitation au regard d'autres concepts ce qui explique que son champ d'application est débattu.

---

[49] Cf. ces Lignes directrices, p.14. Le GEHN considère que « Le respect de l'autonomie humaine est fortement associé au droit à la dignité humaine et à la liberté (reflété aux articles 1 et 6 de la charte). La prévention de toute atteinte est fortement liée à la protection de l'intégrité physique ou mentale (reflétée à l'article 3). L'équité est étroitement liée aux droits à la non-discrimination, à la solidarité et à la justice (reflétés aux articles 21 et s.). L'explicabilité et la responsabilité sont étroitement liées aux droits relatifs à la justice, (tels que reflétés à l'article 47 de la Charte des Droits fondamentaux de l'UE)Le GEHN considère que « L'explicabilité et la responsabilité sont étroitement liées aux droits relatifs à la justice, (tels que reflétés à l'article 47 de la Charte des Droits fondamentaux de l'UE)

[50] Cf. p. 18 ; point 75 p. 22.

[51] p. 9 des lignes directrices.

[52] Cf. En ce sens Valérie Beaudouin et al. (2020a, p. 10)) qui constatent d'autres auteurs considèrent que la transparence est un sous principe de l'explicabilité.

3.4.4. Une notion dont le champ d'application est débattu

Le principe d'explicabilité soulève quatre principaux types de débat relatif à son champ d'application, qui sont étroitement imbriqués les uns aux autres.

Premièrement, il est important de noter que la construction d'explications, de justifications ou d'interprétations n'est pas une activité simple et neutre basée sur des spécifications exclusivement techniques. Il s'agit d'une activité délibérée, qui renvoie à trois dimensions (Barale, 2021).

1. Des explications pour qui ? Les algorithmes, plateformes et autres artefacts autonomes sont conçus, codés, utilisés par et produisent des impacts pour plusieurs parties prenantes différentes : ingénieurs logiciels, utilisateurs indirects (qui spécifient l'outil, mais ne l'utilisent pas directement), utilisateurs finaux (ceux qui l'utilisent effectivement dans un but spécifique), citoyens/clients touchés, la société dans son ensemble, etc. Chacun d'entre eux a des attentes différentes vis-à-vis de ces artefacts et attend par conséquent différents types d'explications.
2. Des explications pour quel besoin ? Étant donné la diversité des parties prenantes impliquées, il est normal de s'attendre à une diversité de motivations et de spécifications pour les explications. Parmi les motivations génériques, citons le test, la vérification, le contrôle, la compréhension et l'appel/la révision/la mise à jour.
3. Des explications pour faire quoi ? Outre les explications conçues pour un objectif générique, il existe des utilisations spécifiques des explications qui doivent être prises en compte et anticipées (dans la mesure du possible). Dans une telle perspective, les explications construites pour un objectif spécifique ne répondent généralement pas aux exigences d'autres objectifs et d'autres parties prenantes spécifiques qui pourraient vouloir se servir des explications.

Deuxièmement, **ce débat sémantique relève en réalité un débat relatif à l'objet de l'explicabilité**. La dualité de l'objet de l'explicabilité a été rappelée par le GEHN qui distingue bien deux types d'explicabilité. D'une part, une explicabilité technique et d'autre part une explicabilité liée aux décisions humaines prises en s'appuyant sur cette technique. « L'explicabilité technique suppose que les décisions prises par un système d'IA puissent être comprises et retracées par des êtres humains. Par ailleurs, des arbitrages peuvent s'avérer nécessaires entre le renforcement de l'explicabilité d'un système (qui pourrait réduire sa précision) et l'amélioration de sa précision (au détriment de l'explicabilité). Dès qu'un système d'IA a une incidence importante sur la vie des personnes, il devrait être possible d'exiger une explication appropriée du processus de décision du système d'IA. Ces explications devraient être présentées en temps opportun et adaptées à l'expertise de la partie prenante concernée (par exemple, non-spécialiste, autorité de réglementation ou chercheur). Des explications devraient également être fournies sur la mesure dans laquelle un système d'IA influence et façonne le processus de prise de décisions organisationnel, les choix opérés dans la conception du système, et la justification de son déploiement (de manière à assurer la transparence du modèle économique) ».[53]

[53] Point 77 de ses lignes directrices.

Ainsi, le caractère contextuel de l'explicabilité implique nécessairement de déterminer avec précision son objet. Schématiquement, l'explicabilité peut porter sur le choix d'un jeu de données, sur l'algorithme qui va utiliser les données, sur le modèle qui va ainsi être créé et enfin sur le contenu, la décision ou la prédiction qui sera prise sur la base de ce modèle. Même lorsque l'explicabilité ne porte que sur l'algorithme, certains auteurs tentent de classifier les formes d'explicabilité possibles en fonction du type d'algorithme en cause (Brkan & Bonnet, 2020).

Troisièmement, il existe un débat sur **l'ampleur de l'explicabilité**. Ce débat comporte plusieurs aspects. D'une part, il existe un débat sur la nécessité, la pertinence ou pas de prévoir systématiquement une forme d'explicabilité. Certains auteurs proposent une classification des systèmes d'intelligence artificielle en fonction du risque, avec pour objectif de limiter l'explicabilité à certains types d'intelligence artificielle à haut risque (Robbins, 2019). C'est d'ailleurs dans cette voie que le législateur européen semble vouloir aller. L'article 13 de sa récente proposition de règlement limite la transparence et la fourniture d'informations aux utilisateurs au système d'IA à haut risque. Au-delà de l'approche par le risque, certains auteurs envisagent l'identification du degré d'explicabilité dans une situation donnée par le biais d'une analyse des coûts et avantages pour la société (Beaudoin et al., 2020a). Ils ont identifié sept catégories de coûts à prendre en compte dans cette évaluation : conception, réduction de la précision de prédictions, création et stockage des journaux, violation du secret d'affaires, conflit avec la sécurité et d'autres objectifs politiques, réduction de la flexibilité décisionnelle dans le futur, ralentissement de l'innovation. Quant aux avantages opérationnels, ils évoquent la capacité à favoriser la confiance des utilisateurs dans le système, et la capacité à rendre les algorithmes plus robustes et certifiables. Ces différents critères sont à prendre en considération de manière contextuelle.

L'ampleur de l'explicabilité est aussi débattue en lien avec le destinataire de l'explicabilité : expert, régulateur ou individu. En effet, l'intelligibilité des explications données varie nécessairement en fonction des connaissances du destinataire. Une information très détaillée et complexe ne sera pas très utile pour le profane qui cherchera surtout à savoir en des termes simples quel modèle algorithmique a été utilisé et les critères qui ont été utilisés par la décision prise à son encontre.

L'ampleur de l'explicabilité est au cœur d'autres notions connexes, notamment celle de **traçabilité et d'audibilité**. Comme le souligne le GEHN, « Il n'est pas toujours possible d'expliquer pour quelle raison un modèle a généré un résultat ou une décision en particulier (et quelle combinaison de facteurs d'entrée y a contribué). On parle d'algorithmes à effet « boîte noire ». Ceux-ci doivent faire l'objet d'une attention particulière. Dans de telles circonstances, d'autres mesures d'explicabilité (par exemple la traçabilité, l'auditabilité et la communication transparente concernant les capacités du système) pourraient être requises, pour autant que le système dans son ensemble respecte les droits fondamentaux. » Ainsi, la **traçabilité** peut être une exigence minimale lorsqu'une explicabilité complète n'est pas possible. Dans ce contexte le GEHN définit la traçabilité comme « Les ensembles de données et les processus permettant au système d'IA de rendre une décision, y compris les processus de collecte et d'étiquetage de données, ainsi que les algorithmes utilisés, devraient être documentés selon les normes les plus strictes afin de permettre la traçabilité ainsi qu'une

amélioration de la transparence. Ce principe s'applique également aux décisions rendues par le système d'IA. Cela permet de déterminer les raisons pour lesquelles une décision d'IA était erronée ce qui, en retour, pourrait contribuer à éviter de futures erreurs ». La traçabilité peut reposer sur des obligations de journalisation, c'est à dire de mémorisation des différentes actions menées par un modèle ou un algorithme. **L'auditabilité** va plus loin puisqu'il s'agit de permettre à un tiers d'inspecter *ex ante* et/ou *ex post* le modèle et/ou les décisions algorithmiques prises sur la base d'un modèle. L'auditabilité est donc un moyen de vérifier la pertinence des explications fournies tant dans le cadre d'une explicabilité globale que locale. Bien que distinctes les notions de traçabilité et d'auditabilité sont donc étroitement liées à l'explicabilité

Quatrièmement, il existe un débat sur les **limites au droit d'explicabilité.** Ce débat est très présent chez les juristes qui tentent de délimiter dans quelles mesures, l'explicabilité pourrait être remise en cause pour des raisons de droit de propriété intellectuelle, de secret d'affaires, ou d'autres secrets liés à la défense ou à la sécurité nationale (Brkan,& Bonnet, 2020 ; Maggiolino, 2019). Dans la recherche en apprentissage machine, ce sont surtout les limites techniques de l'explicabilité qui sont analysées. Il s'agit d'ailleurs parfois d'expliquer que dans certaines situations les explications ne pourront être que partielles, voire ne devraient se concentrer que sur les raisons les plus importantes qui ont conduit à la décision dans une situation particulière. À ce stade, il convient de mentionner l'existence d'une recherche bidisciplinaire (conjuguant apprentissage machine et droit) ayant eu pour objet d'aboutir à une grille d'évaluation de la faisabilité technique de l'obligation d'explicabilité inscrite dans l'article 22 du RGPD.

Type of explanation	Meaning of explanation	Time of issue	Best fitted systems to explain	Best fitted technical method	Overall technical feasibility	Feasibility to overcome legal obstacles	Degree of correspondence with GDPR explanation
Property checking	Proving a given decision will always or never happen in a given situation	Ex ante	Unsupervised	External white-box methods	Moderate	Difficult	Low
Interpretable traces	Translating execution traces into natural language	Ex post	Supervised	Reflexive built on top methods	Difficult	Moderate	High
Local explanation	Identifying the main factors involved	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Counter-factual faithfulness	Evaluating the influence of different factors	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Decision decompilation	Approximating the logics involved	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Providing arguments	Providing pro and con arguments to support a decision	Ex post	Supervised / Unsupervised	Reflexive built within methods	Difficult	Moderate	Medium

Tableau 5 – Faisabilité pratique des explications des décisions algorithmiques [Brkan & Bonnet, 2020 p. 47]

L'évaluation de la faisabilité technique de l'obligation d'explicabilité est intéressante mais si l'article 22 du RGPD est important, ce n'est pas la seule disposition juridique qui porte sur le sujet. De plus, les fondements et la méthode de la qualification de la faisabilité technique de différents types d'explications ne sont pas détaillés par les auteurs. Nous en venons

maintenant au constat d'une certaine diversité, pour ne pas dire hétérogénéité de la finalité des dispositions juridiques portant sur l'explicabilité, en droit de l'Union européenne et en droit français.

3.4.5. Hétérogénéité de la finalité des dispositions juridiques en droit européen et en droit français.

Le RGPD comprend, en plus de l'article 22, deux dispositions importantes : l'article 15 et le considérant n° 71. Les deux articles visent des objectifs différents : l'obligation d'information (art. 15) et le droit à une décision humaine finale (art. 22) ; ils donnent des droits aux individus, cependant avec des restrictions pour les bénéficiaires du droit à une décision humaine au titre de l'article 22 (voir supra § 3.4.1. et tableau 6, ci-dessous). En revanche, le considérant n° 71, ainsi que l'article 29 des *Working Party Guidelines* (qui ont inspiré le RGPD), visent les experts des algorithmes, concepteurs et programmateurs. Ces deux textes poursuivent l'objectif d'améliorer les modèles, alors que les précédents ont d'autres finalités. Enfin, le Règlement européen 2019/1150 déjà évoqué, vise les acteurs économiques (fournisseurs de services d'intermédiation en ligne ou de moteurs de recherche) et leur impose d'être en état d'expliquer (au régulateur ou au juge) les paramètres de classement utilisés dans leurs algorithmes.

En droit français, deux lois (la loi informatique et liberté et la loi bioéthique) ainsi que le Code des relations du public avec l'administration et deux décisions du Conseil constitutionnel dessinent un paysage contrasté des règles relatives à l'explicabilité des décisions algorithmiques (tableau 7, ci-dessous). Par ailleurs, le CRPA et la loi sur la bioéthique définissent l'explicabilité en termes d'information sur la structure des modèles, mais la première concerne les individus alors que la seconde concerne également les professionnels de santé et les concepteurs d'algorithmes utilisés en santé.

Conclusion

L'argument central avancé dans cet article est que la réalisation de l'équité des algorithmes d'apprentissage machine ne peut être prise en charge par des disciplines prises séparément : la science juridique et la technique algorithmique, dans ses composantes mathématiques et informatiques. Comme d'autres (Abu Elyounes, 2020 ; Besse et al., 2018 ;Hacker, 2019 ; Wachter et al., 2020, 2021 ; Xiang, 2021) nous soutenons que les deux disciplines doivent croiser et enrichir leurs perspectives. D'un côté, le design des modèles et les tests utilisés doivent être compris par les *social scientists* et les juristes ; de l'autre, les regards proposés par ces derniers, joints à la compréhension des mécanismes juridiques et de leur contexte culturel et institutionnel, peuvent être utilement intégrés par les chercheurs en IA et les concepteurs d'algorithmes d'apprentissage automatique.

Même si la coopération interdisciplinaire n'est pas simple, il est souhaitable qu'elle se développe à la fois pour produire des connaissances et constituer une aide à la décision publique. Les chercheurs pourraient promouvoir des travaux sur l'équité fondés sur les valeurs et principes en vigueur dans la société, ce qui revient à instrumenter la modélisation sur la base des préférences sociales incorporées dans la loi et la jurisprudence ; il est également souhaitable que des regards critiques soient jetés sur ces dernières afin, le cas échéant, d'en améliorer la pertinence. Ainsi, des recommandations pourraient être

formulées pour le législateur, le régulateur ou le juge. Nous avons soutenu qu'il serait utile de mieux légiférer sur l'explicabilité des décisions algorithmiques, en lui imposant des objectifs clairs et réalistes.

Des voies de recherche interdisciplinaire peuvent être esquissées.

a) Une division du travail entre *computer scientists* et *social scientists* a été évoquée précédemment, lorsque Wachter & al. (2021) soutenaient que le droit et les politiques publiques ont seuls la capacité à faire évoluer les biais sociaux historiques vers davantage d'équité (notion de « *bias transforming* »), alors que la recherche de « fair machine learning » serait fatalement condamnée à utiliser des *« bias preserving metrics*». On peut estimer que cette division du travail ne rend pas grâce à la vigueur de la recherche en apprentissage-machine qui vise à faire advenir un monde meilleur via des algorithmes non discriminatoires (par ex. Feldman & al., 2015), mais qui peut gagner en efficacité avec la coopération de juristes et autres *social scientists*.

b) L'étude des préférences encapsulées dans les algorithmes mériterait d'être développée. Comme le soutient Binns (2018), l'équité en apprentissage-machine peut renvoyer à une variété de considérations égalitaristes, donc de critères et de théories de la justice sociale, qu'il est légitime d'expliciter et de justifier. Dans un même ordre d'idées, Tsoukiàs soutient qu'il y a, dans tout système d'aide à la décision, des préférences et des valeurs ; par conséquent, "*If an autonomous artefact is able to make a decision or to compute a recommendation, it means that somebody embedded within the artefact his/her preferences. And these are independent from how the artefact turns to learn out from the data feeding it. It turns out that is of paramount importance to know how values are actually embedded in any of such systems and/or how these are learned*"(Tsoukiàs, 2021, p. 158). Friedler et al. (2016) analysent les variables, inobservables mais significatives pour la prédiction, qui doivent être prises en compte comme espace intermédiaire entre l'espace des inputs et celui des résultats. Ils montrent alors que les axiomatiques de choix du décideur pour la sélection des candidats à l'entrée à l'université ont des implications sur les résultats et leur (in)équité.

Cela nous mène à une troisième proposition.

c) En généralisant les considérations précédentes nous aboutissons au problème du choix social d'une norme d'équité. La société préfère-t-elle chercher l'égalité des chances, ou la parité démographique, la (multi)calibration, la causalité, ou toute autre forme de *fairness* ? Une question fondamentale se pose aussi : l'équité est-elle la seule valeur sociale recherchée ? Est-elle compatible avec d'autres objectifs sociaux ? A cet égard, par exemple Corbett-Davies & al. (2017) montrent, dans le cas du risque de récidive de COMPAS et dans des termes utilitaristes, que la poursuite de l'objectif de maximisation de la sécurité publique n'est pas nécessairement compatible avec l'égalité de traitement des individus selon la race ; "*Since the optimal constrained and unconstrained algorithms in general differ, there is tension between reducing racial disparities and improving public safety*" (Corbett-Davies & al., 2017). De plus, les politiques publiques et le droit peuvent véhiculer plusieurs logiques compte-tenu du fait que les choix sociaux peuvent être différenciés selon les domaines (par exemple entre les politiques pénales et de sécurité publique et la protection

sociale ou l'accès à l'université). La modélisation de choix sociaux composites pourrait éclairer les débats publics et le législateur, qui pourrait être mieux éclairé sur les implications des choix faits[54]. De plus, la possibilité d'un écart entre les valeurs sociales portées par le décideur public et celles de la société civile peut-elle donner lieu à une modélisation en termes *de bi-sided fairness* ?

d) L'explicabilité des décisions algorithmiques automatiques mérite d'être davantage étudié. Est-il possible de concevoir une ou des méthode(s) d'évaluation technique de la faisabilité de l'explicabilité telle que le droit l'impose, avec ses zones de flou ou d'indétermination ? Est-il envisageable de démontrer que l'explicabilité peut être une voie de réalisation de l'équité des décisions algorithmiques ?

e) Si des travaux sur le droit anti-discrimination existent sur les cas des Etats-Unis et de l'Union européenne, les régimes juridiques et institutionnels nationaux sont très peu analysés. Une meilleure connaissance des enjeux juridique et politiques de systèmes nationaux (en Europe, en Asie, Océanie, Amérique latine) pourrait alimenter les voies de recherche pluridisciplinaire auxquelles cet article s'est attaché.

[54] Par exemple, les computer scientists peuvent montrer les impossibilités, telle la conciliation entre non-discrimination individuelle et non-discrimination de groupe (Friedler, Scheidegger & Venkatasubramanian, 2016).

Texte	Disposition	Champ d'application/scope	Modalités pratiques	Qui est concerné	Objectif /Goals
Article 29 Working Party Guidelines on Automated individual decision-making and Profiling.[55]	"algorithmic auditing" : "testing the algorithms used and developed by machine learning systems to prove that they are actually performing as intended, and not producing discriminatory, erroneous or unjustified results"	Sans restriction	Vérification formelle des algorithmes + sécurité	Experts : Concepteurs/programmateurs + certificateurs	Améliorer modèle
RGPD	Article 15(1)(h) : droit à être informé d'une prise décision automatisée + logique sous-jacente + conséquences pour la personne	Sans restriction	Logique interne : modèle + causalité	Individus	Transparence + causalité
	Article 22 : droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé, y compris le profilage, produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire.	Non applicable si -pas d'effets juridiques ou personne non affectée -exécution d'un contrat, consentement, autorisation légale	Concerne aussi bien décision automatisée (explicable) mais aussi autonome (black box non explicable) Causalité	Individus	Décision humaine finale
	Considérant n° 71- utilisation des procédures mathématiques ou statistiques Recours à des mesures techniques et organisationnelles … faire en sorte… que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le risque d'erreur soit réduit au minimum…et qui prévienne les effets discriminatoires	Sans restriction	Pipel Vérification formelle Contrôle données Test	Experts : Concepteurs/programmateurs	Améliorer modèle
Règlement Européen 2019/1150	Article 5 – Paramètres des classements par les plateformes	Sans restriction	Vérification modèle de ranking	Industrie : (fournisseurs de services d'intermédiation en ligne + de moteurs de recherche)	Transparence + compliance

Tableau 6. L'explicabilité en droit européen (source : auteurs)

[55] En partie intégrées dans le considérant n° 71 et dans l'article 22 du RGPD.

Droit français - Texte juridique	**Disposition**	**Champ d'application**	**Modalités pratiques**	**Qui est concerné**	**Objectif/Goal**
Code des relations du Public avec l'Administration (intégration de la LRN)	Droit à l'information si demande de l'intéressé	Décisions algorithmiques de l'administration (organisation privées exclues)	Structure du modèle	Individus	Information + transparence
Loi Informatique et Liberté	Art 47§2 : explication en détail et sous une forme intelligible : -contribution à la prise de décision -données -paramètres de traitement opérations effectuées	Décisions individuelles algorithmiques (administration et organisations privées)	Pipe Causalité Données Structure du modèle	Individus	Justification de la décision
Conseil constitutionnel (Décision du 12 juin 2018 relative à la loi sur les données personnelles)	Limitation du recours deep learning (algo autonomes) et des algorithmes protégés par les DPI ou le secret	Décisions individuelles algorithmiques de l'administration (organisation privées exclues)	np	Individus	Limitation des décisions autonomes
Conseil constitutionnel (Décision du 3 avril 2020 - Communicabilité et publicité des algorithmes utilisés pour l'entrée à l'Université)	Droit des tiers à explication (UNEF, syndicat étudiants)	Algorithme de sélection à l'entrée de l'Université (Parcours Sup)	np	Tiers (non expert)	Transparence
Loi sur la bioéthique du 2 août 2021	Article 17 – Information sur l'usage d'un algorithme d'apprentissage (pour prévention, diagnostic ou soin) et l'interprétation qui en résulte. ; les concepteurs du traitement algorithmique s'assurent de l'explicabilité de son fonctionnement pour les utilisateurs	Décisions algorithmiques en santé	Structure du modèle	Individus Professionnels de la santé Experts : concepteurs	Information + transparence

Tableau 7. L'explicabilité en droit français (source : auteurs)

Références

Abu Elyounes, Doaa, "Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness", Journal of Law, Technology and Policy, 2020/1, p. 1-54. Available at SSRN: https://ssrn.com/abstract=3478296 or http://dx.doi.org/10.2139/ssrn.3478296

Adadi A and M Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence" (2018) 6 IEEE Access 52140, 52142–52143.

Angwin J., Larson J., Mattu S., Kirchner L. (2016, May), Machine bias, ProPublica - https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arnsperger Christian, Van Parijs Philippe, *Éthique économique et sociale.* Paris, La Découverte, « Repères », 2003.

ARTICLE 29 DATA PROTECTION WORKING PARTY, 17/EN, WP251rev.01, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, Adopted on 3 October 2017, As last Revised and Adopted on 6 February 2018

Barale, Claire, Explanations in decision support. Generating Fairness through explanations, Master Thesis, LAMSADE, Université ParisDauphine-PSL, sept. 2021.

Barocas Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Barocas, Solon & Selbst, Andrew D., Big Data's Disparate Impact, 104 *CALIF. L. REV*. 671 (2016).

Barraud Boris, Les algorithmes au cœur du droit et de l'état postmoderne, International Journal of Digital and Data Law [2018–Vol 4] http://ojs.imodev.org/index.php/RIDDN;

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J., Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach (2020a), , accessible à https://arxiv.org/pdf/2003.07703.pdf

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020b). « Identifying the right level of explanability in a given situation » https://hal.telecom-paris.fr/hal-02507316

Ben-Tal, Aharon, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

Besse Philippe, Castets-Renard, Céline, Garivier, Aurélien, Loubes, Jean-Michel, L'IA du quotidien peut-elle être éthique ? Loyauté des algorithmes d'apprentissage automatique, Statistique et Société, vol 6, n° 3, décembre 2018.

Binns, Reuben, Fairness in Machine Learning: Lessons from Political Philosophy, Proceedings of Machine Learning Research 81:1–11, 2018 Conference on Fairness, Accountability, and Transparency.

Bolukbasi Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Brkan, M., & Bonnet, G. (2020). "Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas". *European Journal of Risk Regulation*, 11(1), 18-50. doi:10.1017/err.2020.10, accessible à https://dx.doi.org/10.1017/err.2020.10

Brkan, Maja, "Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond" (August 1, 2017). A revised version of this paper has been published in *International Journal of Law and Information Technology*, 11 January 2019, DOI; 10.1093/ijlit/eay017, Available at SSRN: https://ssrn.com/abstract=3124901 or http://dx.doi.org/10.2139/ssrn.3124901

Buolamwini Joy and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

Castets-Renard C, "Accountability of Algorithms in the GDPR and Beyond: A European Legal Framework on Automated Decision-Making", 30 *Fordham Intell. Prop. Media & Ent. L.J.* 91 (2019). Available at: https://ir.lawnet.fordham.edu/iplj/vol30/iss1/3;

Castets-Renard, Céline, "Régulation des algorithmes et gouvernance du machine learning : vers une transparence et « explicabilité » des décisions algorithmiques ? (Algorithm Regulation and Machine Learning Governance: Towards Transparency and 'Explainability' of Algorithmic Decisions?) " (September 20, 2018). Revue *Droit & Affaires*, Revue Paris II Assas, 15ème édition, 2018., Available at SSRN: https://ssrn.com/abstract=3391282

Chen, Irene, Fredrik D Johansson, and David Sontag. "Why is my classifier discriminatory?" *Advances in Neural Information Processing Systems*, 31, 2018.

Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". *arXiv preprint arXiv:1703.00056*, 2017.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. "The variational fair auto encoder ». *arXiv preprint arXiv:1511.00830*, 2015.

CNIL, *Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, décembre 2017.

com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 2018.

Corbett-Davis, Sam, Pierson Emma, Feller Avi, Goel, Sharad, Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear", The Washington Post, 10/2016.

Corbett-Davies Sam and Goel Sharad. "The measure and mismeasure of fairness: A critical review of fair machine learning". *arXiv preprint arXiv:1808.00023*, 2018.

Corbett-Davis, Sam, Pierson Emma, Feller Avi, Goel, Sharad Huq Aziz, "Algorithmic decision making and the cost of fairness", 2017.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated experiments on ad privacy settings". *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

de Vries Terrance, Ishan Misra, Changhan Wang, and Laurens van der Maaten. "Does object recognition work for everyone?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

Diana Emily, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. "Minimax group fairness: Algorithms and experiments". In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Du Perron Simon et Karim Benyekhlef, “Les algorithmes et l’État de droit“, (2021) Document de travail # 27, Laboratoire de cyberjustice, Faculté de droit, Université de Montréal : https://cyberjustice.openum.ca/files/sites/102/Working-Paper_Simon-du-Perron_-vers-17-juin-2021-1.pdf

Do V., Corbett-Davies S, Atif J., & Usunier N, “Two-sided fairness in rankings via Lorenz dominance”, *https://arxiv.org/abs/2110.15781v1*, 2021

Dwork, Cynthia, Pitassi, Toniann, Reingold, Omer, Zemel, “Rich, Fairness Through Awareness“, *arXiv, 1104.3913*, 2011.

Edwards, L. and Veale, M. (2017). Slave to the algorithm: Why a “right to an explanation” is probably not the remedy you are looking for. Duke Law & Technology Review, 16(1), 18–84, https://ssrn.com/abstract=2972855

Edwards, Lilian and Veale, Michael, “Enslaving the Algorithm: From a ‘Right to an Explanation’ to a ‘Right to Better Decisions’? “*IEEE Security & Privacy* (2018) 16(3), pp. 46-54, DOI: 10.1109/MSP.2018.2701152, Available at SSRN: https://ssrn.com/abstract=3052831

Farhan Rahman, COMPAS Case Study: “Fairness of a Machine Learning Model”, Sep 7, 2020· https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751.

Feldman, Michael, Friedler, Sorelle,Moeller, John, Scheidegger, Carlos & Venkatasubramanian, Suresh, “Certifying and removing disparate impact”, https://arxiv.org/abs/1412.3756, 16 Jul 2015.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges“, Natural Language Processing and Chinese Computing (pp.563-574)

Friedler, Sorelle A. and Scheidegger, Carlos and Venkatasubramanian, Suresh, “On the (im)possibility of fairness”, https://arxiv.org/abs/1609.07236, 2016; doi = {10.48550/ARXIV.1609.07236},

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi,D. (2018). “A survey of methods for explaining black box models“, *ACM computing surveys* (CSUR), 51(5):93.

Groupe d’experts de haut niveau sur l’intelligence artificielle, *Lignes directrices en matière d’éthique pour une IA digne de confiance*, Commission européenne, Bruxelles, 8 avril 2019.

Hacker, Philipp, “Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law“ (April 18, 2018). 55 Common Market Law Review 1143-1186 (2018), Available at SSRN: https://ssrn.com/abstract=3164973

Hall Melissa, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. “A systematic study of bias amplification”. *arXiv preprint arXiv:2201.11706*, 2022.

Hashimoto Tatsunori, Megha Srivastava, Hongseok Namkoong, and Percy Liang. “Fairness without demographics in repeated loss minimization“. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Hildebrandt M. “Algorithmic regulation and the rule of law“. Philosophical transactions. Series A, Mathematical, Physical, and Engineering Sciences. 2018 Sep;376(2128). DOI: 10.1098/rsta.2017.0355. PMID: 30082301

Kallus, Nathan and Zhou. Angela, “Residual unfairness in fair machine learning from prejudiced data”. *International Conference on Machine Learning*, pages 2439–2448. PMLR, 2018.

Kleinberg, Jon Sendhil Mullainathan, and Manish Raghavan. “Inherent trade-offs in the fair determination of risk scores“. *arXiv preprint arXiv:1609.05807*, 2016.

Koulu, Riikka, "Crafting Digital Transparency: Implementing Legal Values into Algorithmic Design", *Critical Analysis of Law*, Special Issue: Transparency in the Digital Environment, Vol. 8 No. 1 , 2021.

Larson, Jeff, Mattu, Surya, Kirchner Lauren and Angwin Julia, "How We Analyzed the COMPAS Recidivism Algorithm", May 2016.

Lee A. Bygrave, "Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions", *University of Oslo Faculty of Law Research Paper No. 2020-35*, p. 9 accessible à https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3721118

Levy Daniel, Yair Carmon, John C Duchi, and Aaron Sidford. "Large-scale methods for distributionally robust optimization". *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

Lloyd Kirsten. "Bias amplification in artificial intelligence systems". *arXiv preprint arXiv:1809.07842*, 2018.

Lum Kristian and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016. Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.

Maggiolino, Mariateresa, "EU Trade Secrets Law and Algorithmic Transparency" (March 31, 2019). Bocconi Legal Studies Research Paper No. 3363178, Available at SSRN: https://ssrn.com/abstract=3363178

Malgieri G, "Automated decision-making in the EU Member States: The right to explanation and other 'suitable safeguards'" in the national legislations, *Computer Law and Security Review* 35, (2019), p.1-26, accessible à https://www.sciencedirect.com/science/article/pii/S0267364918303753

Maxwell Winston, "Comment améliorer l'explicabilité et la responsabilité des algorithmes ? " *Les cahiers Louis Bachelier*, avril 2020;

Menon Sachit, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. "Pulse: Self-supervised photo upsampling via latent space exploration of generative models". In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020.

Mitchell Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting". In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

Mitchell Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. "Algorithmic fairness: Choices, assumptions, and definitions". *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

OCDE, *Recommandation du Conseil sur l'intelligence artificielle*, OECD/LEGAL/0449, adoptée le 22/05/2019

Pedreschi, Dino and Ruggieri, Salvatore and Turini, Franco, "A Study of Top-K Measures for Discrimination Discover"y, *Proceedings of the ACM Symposium on Applied Computing*, May 2012, doi = 10.1145/2245276.2245303

Peresie, Jennifer L., " Toward a Coherent Test for Disparate Impact Discrimination", *Indiana Law Journal*, vol. 84, 2009, p. 773-802.

Robbins, Scott , "A Misdirected Principle with a Cath: Explicability for AI", *Minds & Machines* 29, 495–514 (2019). https://doi.org/10.1007/s11023-019-09509-3

Rochaix, Lise, et Sandy Tubeuf. "Mesures de l'équité en santé. Fondements éthiques et implications", *Revue économique*, vol. 60, no. 2, 2009, pp. 325-344.

Rouvroy Antoinette et Thomas Berns, "Gouvernementalité algorithmique et perspectives d'émancipation", *Réseaux*, 177, 2013, p. 163-196

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

Schiek, Dagmar, Waddington, Lisa & Bell, Mark (Eds), *Cases, Materials and Text on National, Supranational and International Non-Discrimination Law*, Oxford & Oregon, Hart Publishing, Ius Commune casebooks for the common law of Europe, 2007.

Scott, A.C., Clancey, W.J., Davis, R., Shortliffe, E.H.: "Explanation capabilities of production-based consultation systems". *American Journal of Computational Linguistic*s 62, 1977.

Selbst, Andrew D Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

Selmi, Michael, "Was the Disparate Impact Theory a Mistake? ", 53 *UCLA L. Rev*. 701, 2006.

Seng Ah, Lee Michelle, Luciano Floridi, and Jatinder Singh. "From fairness metrics to key ethics indicators (keis): a context-aware approach to algorithmic ethics in an unequal society". *Available at SSRN*, 2020.

Seng Ah, Michelle Lee and Luciano Floridi. "Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs". *Minds and Machines*, 31:165–191, 2021.

Słowik, Agnieszka and Léon Bottou. "Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation". *arXiv preprint arXiv:2106.09467*, 2021.

Stoica, Ion& al, "A Berkeley View of Systems Challenges for AI, Electrical Engineering and Computer Sciences", *University of California at Berkeley, Technical Report* No. UCB/EECS-2017-159, October 16, 2017, http://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.html

Suk, Julie C., "Disparate Impact Abroad", in : *A Nation of Widening Opportunities? The Civil Rights Act at 50*, Samuel Bagenstos and Ellen Katz, eds., University of Michigan Press, 2014.

Suresh Harini and John Guttag. "A framework for understanding sources of harm throughout the machine learning life cycle", In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9., 2021.

Sweeney Latanya. "Discrimination in online ad delivery". *Queue*, 11(3):10, 2013.

Tambou, Olivia, *Manuel de droit européen des données*, Bruxelles: Bruylant (2020).

Tsoukiàs, Alexis, "Social Responsibility of Algorithms: An Overview", in J. Papathanasiou, P. Zaraté, J. Freire de Sousa (eds.), *EURO Working Group on DSS*, Springer International Publishing, pp.153-166, 2021. https://arxiv.org/abs/2012.03319.

Vinik, D. Frank, *Disparate impact*. *Encyclopedia Britannica*. https://www.britannica.com/topic/disparate-impact, 2010.

Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, " Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law" (January 15, 2021). *West Virginia Law Review*, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3792772 or http://dx.doi.org/10.2139/ssrn.3792772

Wachter, Sandra and Mittelstadt, Brent and Russell, Chris, "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (March 3, 2020). *Computer Law & Security Review* (forthcoming), Available at SSRN: https://ssrn.com/abstract=3547922 or http://dx.doi.org/10.2139/ssrn.3547922

Wachter, Sandra, Affinity “Profiling and Discrimination by Association“ in Online Behavioural Advertising (May 15, 2019). *Berkeley Technology Law Journal*, Vol. 35, No. 2, 2020, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3388639 or http://dx.doi.org/10.2139/ssrn.3388639

Wachter/Mittelstadt/Floridi, “Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation“. International *Data Privacy Law* 2017, 7, n°2, p. 76;

Wang Serena, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. “Robust optimization for fairness with noisy protected groups“, *Advances in Neural Information Processing Systems*, 33: 5190–5203, 2020.

Watcher,Sandra Mittelstadt, “Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR“, *Harvard Journal of Law & Technology*, 2018, accessible à https://arxiv.org/abs/1711.00399

Xiang, Alice & Raji, Inioluwa Deborah, “On the Legal Compatibility of Fairness Definitions“, *arXiv: 1912.00761* [cs.CY] 25 Nov 2019.

Xiang, Alice, “Reconciling Legal and Technical Approaches to Algorithmic Bias“, *Tennessee Law Review*, Vol. 88, No. 3, 2021, Available at SSRN: https://ssrn.com/abstract=3650635

Zemel Richard S Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. *ICML (3)*, 28:325–333, 2013.

Zhao Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. “Men also like shopping: Reducing gender bias amplification using corpus-level constraints“. *arXiv preprint arXiv:1707.09457*, 2017.