La pollinisation croisée entre droit de la protection des données et droit de la non-discrimination

Le rôle des chercheurs pour garantir une intelligence artificielle non-discriminatoire

FABIAN LÜTZ
Doctorant en droit
Faculté de droit, Université de Lausanne

Table des matières

I. Introduction 212

II. Interdépendance, conflictualité, imitation et complémentarité 213
A. Interdépendance 215
B. La conflictualité entre protection des données et non-discrimination 215
C. Imitation 217
D. Complémentarité 217

III. Droit de savoir et transparence 218
A. Connaître l'existence et le contenu d'une décision automatisée en matière d'IA 219
B. Expliquer les décisions de l'IA 220
C. Enregistrement du processus décisionnel automatisé 221
D. Rôle des chercheurs 222

IV. Confiance et droit au « *human-in-the-loop* » 224
A. Art. 22 RGPD 225
B. Art. 21 LPD 226
C. Art. 14 *AI Act* et Art. 20 CdE *Zero Draft Convention on AI* 227
D. Rôle des chercheurs 227

V. Sphère privée et non-discrimination dès la conception et par défaut 228
A. Sphère privée dès la conception et par défaut 228
B. Non-discrimination dès la conception and par défaut 229
C. Quel modèle à suivre pour la discrimination algorithmique ? 230
D. Le rôle des chercheurs 230

VI. Analyse d'impact pour les données et les algorithmes..... 231
A. Analyse d'impact dans la protection des données 232
B. Analyse d'impact afin d'éviter les discriminations algorithmiques 233
C. *Monitoring* 234
D. Rôle des chercheurs 235

VII. Conclusion..... 237

VIII. Bibliographie 238
A. Littérature..... 238
B. Documents officiels..... 241

I. Introduction

Depuis une demi-décennie, la protection des données est régie par le Règlement général sur la protection des données (RGPD)[1] tant dans l'Union européenne[2] qu'en-dehors de la juridiction de l'UE, compte tenu de son effet extraterritorial[3] et à travers l'imitation, le fameux *Brussels effect*[4]. En Suisse, la nouvelle loi de la protection de données entrera en vigueur en septembre 2023, reprenant largement le contenu du RGPD, mais favorisant quelques spécificités suisses. Dans ce contexte, le *AI Act* proposé par la Commission européenne en 2021 et actuellement débattu par les colégislateurs, prévoit un cadre législatif pour réguler les algorithmes[5]. En Suisse, il n'y a pas encore de règles juridiques qui gouvernent spécifiquement l'intelligence artificielle (IA)[6] et la discrimination algorithmique sur la base d'une caractéristique protégée par le droit, par exemple le genre ou la religion, même si en principe les règles générales et plus

1 Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE), JO L 119, 4.5.2016, p. 1-88.

2 Voir MÉTILLE/DI TRIA.

3 Voir MÉTILLE/ACKERMANN.

4 BRADFORD, The Future of the Brussels Effect ; BRADFORD, Brussels Effect ; concernant l'égalité des genres, voir LÜTZ, Brussels effect, p. 142-163.

5 La dernière version disponible est le texte compromis du Conseil, <https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/fr/pdf> et la position du Parlement européen, <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf>.

6 La Suisse s'engage néanmoins internationalement dans l'élaboration des règles, notamment en ce qui concerne le Conseil de l'Europe, voir par exemple, RICART *et. al*., p. 78.

particulièrement l'interdiction de la discrimination[7] s'appliquent également dans le contexte de l'IA[8]. C'est dans ce cadre législatif et politique que s'inscrit la présente comparaison et analyse du droit de la protection des données et du droit de la non-discrimination dans l'ère algorithmique[9]. Cette contribution vise à éclairer le(s) rôle(s) des chercheurs en sciences informatiques et juridiques au cours de l'élaboration des algorithmes ainsi qu'à faire la lumière sur leur implication dans le processus législatif.

S'il est souvent question d'équité (*fairness*) dans l'IA, ce concept est loin de faire l'unanimité parmi les théoriciens et la définition de cette *fairness* reste débattue.[10] Par conséquent, l'article privilégiera le concept plus concret et spécifique de l'IA non discriminatoire.

En guise d'introduction, l'article détaillera de manière générale les similarités, l'interdépendance, le potentiel pour l'imitation et la complémentarité entre les deux domaines du droit que sont la protection des données et de la non-discrimination, en se référant principalement au cadre juridique de l'Union européenne et de la Suisse (II.). Ensuite, seront examinés le droit à l'information et à la transparence (III.), la confiance et le droit à la participation d'un humain dans le processus (IV.), les questions de sphère privée et de non-discrimination dès la conception et par défaut (V.) et les analyses d'impact ainsi que le *monitoring* (VI.). L'article conclut en soulignant le rôle des chercheurs et en identifiant les pistes de participations des chercheurs afin d'assurer une meilleure protection des citoyens et de leurs droits fondamentaux.

II. Interdépendance, conflictualité, imitation et complémentarité

Dans l'ère algorithmique[11], le RGPD est sans aucun doute le produit d'exportation phare de l'Union européenne. Il est par conséquent devenu le

7 Par l'interdiction de discrimination on entend toutes les règles de droit qui visent à empêcher la discrimination sur la base d'une caractéristique protégée, tel que le sexe, la religion ou la race.

8 Voir par exemple l'explication du CONSEIL FÉDÉRAL, « Intelligence artificielle » – lignes directrices pour la Confédération, Cadre d'orientation en matière d'IA dans l'administration fédérale (2020), p. 9 concernant l'applicabilité des normes générales comme les droits fondamentaux, droits de l'homme et l'interdiction de la discrimination.

9 Pour un aperçu du *AI Act*, voir LÜTZ, Gender equality and AI in Europe, p. 33-52.

10 Il existe au moins 21 conceptions fréquemment discutées en sciences informatiques, voir BAROCAS/HARDT/NARAYANAN ; voir aussi sur les risques de qualification incorrecte par les algorithmes qui prennent des décisions « juste », CREEL/HELLMANN, p. 2.

11 ABITEBOUL/DOWEK, p. 1 et not. le titre de l'ouvrage.

centre de gravité global visant une convergence régulatrice en termes de protection des données[12]. Les données se situent au cœur de tous les projets récents de l'Union européenne en matière digitale dans cette *décennie numérique*[13], tels que le *Digital Markets Act* (DMA)[14], *Digital Services Act* (DSA)[15], *Data Governance Act* (DGA)[16], *Artificial Intelligence Act* (AI Act)[17]. Ce sont en effet principalement les données[18] qui font fonctionner les algorithmes, l'IA et les données étant interdépendantes.[19] Cependant, ce sont aussi ces données qui doivent être protégées au profit des individus. Le RGPD reconnaît que :

> « *Des risques pour les droits et libertés des personnes physiques* […] *peuvent résulter du traitement de données à caractère personnel qui est susceptible d'entraîner des dommages physiques, matériels ou un préjudice moral, en particulier : lorsque le traitement peut donner lieu à une discrimination* […] ».[20].

---

12 Voir p. ex., SMUHA ; BENBOUZID/MENECEUR/SMUHA, p. 29-64 ; La nouvelle LPD en Suisse par exemple est très similaire, voir DI TRIA.

13 Voir not. EC, European Declaration on Digital Rights and Principles for the Digital Decade, para. 9 : « *Toute personne devrait être en mesure de bénéficier des avantages qu'offrent les systèmes algorithmiques et d'intelligence artificielle* […] *tout en étant protégée contre les risques et les atteintes à sa santé, à sa sécurité et à ses droits fondamentaux.* »

14 Règlement (UE) 2022/1925 du Parlement européen et du Conseil du 14 septembre 2022 relatif aux marchés contestables et équitables dans le secteur numérique et modifiant les directives (UE) 2019/1937 et (UE) 2020/1828 (règlement sur les marchés numériques) (Texte présentant de l'intérêt pour l'EEE), JO L 265 du 12.10.2022, p. 1-66.

15 Règlement (UE) 2022/2065 du Parlement européen et du Conseil du 19 octobre 2022 relatif à un marché unique des services numériques et modifiant la directive 2000/31/CE (règlement sur les services numériques) (Texte présentant de l'intérêt pour l'EEE), JO L 277 du 27.10.2022, p. 1-102.

16 Règlement (UE) 2022/868 du Parlement européen et du Conseil du 30 mai 2022 portant sur la gouvernance européenne des données et modifiant le règlement (UE) 2018/1724 (règlement sur la gouvernance des données) (Texte présentant de l'intérêt pour l'EEE), JO L 152 du 3.6.2022, p. 1-44.

17 Proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle (Législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'Union, COM/2021/206 final.

18 Les données utilisées par les algorithmes pour l'entraînement et la décision peuvent être affectées par des biais et stéréotypes ce qui facilite les discriminations, voir EU FUNDAMENTAL RIGHTS AGENCY, Bias in Algorithms, p. 24. SUSSKIND, p. 257-258 (« *The real world contains patterns of injustice. These patterns are reflected in data. Algorithms reproduce and amplify them.* »).

19 Voir, EC, European Declaration on Digital Rights and Principles for the Digital Decade, para. 9 (c) : *« veiller à ce que les systèmes algorithmiques reposent sur des ensembles de données appropriés, afin d'éviter toute discrimination* […]. *»*.

20 RGPD, consid. 75. Même si le RGPD mentionne des risques de discrimination dans les considérants du règlement, son objectif premier n'est pas la protection contre les discriminations basées par exemple sur le sexe ou la religion, mais bien *« la protection des*

A. Interdépendance

Le fait que la doctrine a souvent combiné la protection des données et la non-discrimination afin d'illuminer la discrimination algorithmique témoigne de l'interdépendance et de la complémentarité de ces deux domaines juridiques pour adresser la problématique[21]. Ceci s'explique également par le fait qu'initialement dans la doctrine émergente, le seul outil adapté au monde numérique dans le droit européen était le RGPD. À la suite de l'adoption de la proposition du règlement *AI Act*, la doctrine s'est progressivement tournée vers d'autres (futurs) instruments.

B. La conflictualité entre protection des données et non-discrimination

La doctrine admet que le RGPD ne vise pas uniquement la protection des données, mais également d'autres droits, comme le droit à la non-discrimination[22]. Il n'en reste pas moins que les objectifs poursuivis par ces deux domaines juridiques peuvent entrer en conflit (conflictualité)[23]. Plus précisément, le RGPD crée une dichotomie entre les données relatives aux caractéristiques protégées qui peuvent être récoltées et celles qui ne le peuvent pas. Ces dernières comprennent les « catégories particulières de données à caractère personnel », notamment la race, la religion et l'orientation sexuelle, à l'exception des caractéristiques « sexe » ou « âge »[24]. Par conséquent, le principe d'interdiction ne permet pas de collecter des données relatives à des caractéristiques protégées par le droit de la non-discrimination (sauf sexe et âge), même dans l'objectif d'empêcher, de diminuer ou de détecter des discriminations. C'est une des raisons pour lesquelles le *AI Act* rappelle :

personnes physiques à l'égard du traitement des données à caractère personnel » (Art. 1 par. 1 RGPD) tout en protégeant les libertés et droits fondamentaux (Art. 1 par. 2 RGPD).

21 À titre d'exemple, voir HACKER, p. 55 ; KULLMANN, p. 61 ainsi que l'étude préparée pour le Conseil de l'Europe par BORGHESIUS, p. 36-46 (droit de la protection des données) et p. 32-36 (droit de la non-discrimination).

22 VAN BEKKUM/BORGESIUS, p. 5.

23 Voir aussi sur le manque de données qui peut mener à une discrimination algorithmique, WILLIAMS/BROOKS/SHMARGAD, p. 78-115.

24 Art. 9 par. 1 RGPD.

> « *Afin de protéger le droit d'autres personnes contre la discrimination qui pourrait résulter des biais dans les systèmes d'IA, les fournisseurs devraient être en mesure de traiter également des catégories spéciales de données à caractère personnel, pour des motifs d'intérêt public important au sens de l'article 9, paragraphe 2, point g), du règlement (UE) 2016/679 et de l'article 10, paragraphe 2, point g), du règlement (UE) 2018/1725, afin d'assurer la surveillance, la détection et la correction des biais liés aux systèmes d'IA à haut risque.* »[25].

Vu que la plupart des algorithmes et *business models* des entreprises nécessitent une large quantité de données (de qualité), cela peut créer un conflit entre les intérêts des personnes concernées et ceux des entreprises. La question se complique encore lorsque la protection des données, notamment à caractère personnel, impacte la lutte efficace contre la discrimination. Afin d'éviter ou diminuer le risque d'une discrimination, il peut en effet être nécessaire d'avoir connaissance du sexe ou de la religion, justement afin d'empêcher toute discrimination[26]. Ceci en raison de la corrélation effectuée par les algorithmes qui, même sans identification spécifique en tant qu'homme ou femme, trouvent des indices dans les données fournies et par conséquent peuvent mener à une discrimination.

Dans ce cas, l'obstacle que représente l'interdiction générale de collecter des données relatives à des caractéristiques protégées comme la race ou l'orientation sexuelle peut être levé, soit par une exception créée par la loi – ce qui est proposé dans le *AI Act*, soit par le consentement de l'individu concerné. Concernant le consentement, celui-ci est prévu par exemple dans le RGPD, à l'art. 9 par. 2 lit. a. S'il est donc en principe envisageable, il s'avère impossible et irréaliste dans les situations d'inégalité de pouvoir comme dans la relation employeur/employée. Même en dehors de ces cas, l'obtention du consentement libre et valable reste complexe et fait l'objet de nombreux débats en doctrine[27].

Outre l'*AI Act*, pour la Suisse les travaux du Conseil de l'Europe en matière de protection des données[28] et d'intelligence artificielle[29] sont très pertinents. Un rapport a été publié sur l'impact de l'IA sur l'égalité de genre et les risques de discriminations algorithmiques dans le cadre du développement d'un cadre législatif général relatif à l'IA à l'échelle du Conseil de l'Europe prévu pour

25 *AI Act*, consid. 44.
26 *AI Act*, consid. 44.
27 VAN BEKKUM/BORGESIUS, p. 6.
28 Voir not. CONSEIL DE L'EUROPE, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, N° 108.
29 Voir CONSEIL DE L'EUROPE, Recommendation on the human rights impacts of algorithmic systems.

2023 et du cadre spécifique pour l'égalité de genre et non-discrimination prévu pour 2025[30].

C. Imitation

Les règles spécifiques, comme l'*AI Act* ou tout acte législatif suisse à venir, s'inspireront très certainement à plus ou moins grande échelle des règles européennes existantes, phénomène décrit comme le *Brussels Effect*[31]. Cela s'explique par la force d'attraction du marché intérieur, les soucis de compliance des entreprises ainsi que la rapidité d'adoption de ces règles en comparaison avec d'autres juridictions dans le monde[32]. Ainsi, les règles sur la régulation de l'IA s'inspirent largement du RGPD en prenant comme outil majeur les *algorithmic impact assessments*, connus du RGPD. En matière de protection de données, les observateurs ont fait état du fait que l'UE exerçait une influence afin que le contenu de la Convention 108+ du Conseil de l'Europe (CdE) soit aligné sur le RGPD. Un alignement semblable paraît possible concernant le AI Act et le projet législatif du CdE, qui impactera directement et indirectement la Suisse.

D. Complémentarité

Au niveau du droit européen, le RGPD reste un cadre juridique incontournable, à présent complété par le DSA, DMA et DGA en ce qui concerne la régulation des plateformes algorithmiques qui peuvent éventuellement mener à des discriminations. Dans un proche avenir, le *AI Act* s'ajoutera aux règles spécifiques et concrètes concernant quelques scénarios de discrimination algorithmique.

Nous avons analysé ci-dessus une série d'interdépendances et des opportunités pour une *imitation* des concepts du droit de la protection des données dans le droit de la non-discrimination, et évoqué quelques questions de conflictualité et complémentarité.

Ensuite, il incombe à l'individu de savoir s'il est soumis à une décision automatisée (si un algorithme est utilisé), car cette connaissance conditionne tout

30 CONSEIL DE L'EUROPE, Preliminary draft Council of Europe study on the impact of artificial intelligence, its potential for promoting equality, including gender equality, and the risks to non-discrimination, GEC(2022)9 CDADI(2022)21.

31 BRADFORD, The Future of the Brussels effect.

32 La ville de New York City a toutefois adopté une loi spécifique pour les systèmes de recrutement automatisés en 2021 qui est entrée en vigueur le 1er janvier 2023. Pour plus de détails, voir LÜTZ, Discrimination algorithmique.

autre droit ultérieur et c'est aussi ce qui mène au sujet de la transparence. Mis en place depuis quelques années, les dispositifs du droit de la protection des données peuvent servir de modèle, mais aussi être combinés afin de créer des synergies, par exemple dans le cas des analyses d'impact.

La présente section examinera le droit de savoir, car sans information il est impossible d'introduire une action en raison d'une possible discrimination.

III. Droit de savoir et transparence

Le droit de savoir et l'information qui en résulte sont des conditions préalables et essentielles pour permettre la transparence[33] et l'exercice effectif des droits fondamentaux[34]. En effet, sans cette connaissance, ni les victimes de discrimination ni les chercheurs ne sont en mesure de vérifier si les systèmes d'IA ne sont pas discriminatoires, une condition vitale de toute démocratie[35]. De plus, l'information concernant l'utilisation d'un algorithme et les éléments clés du fonctionnement constituent la base pour permettre une compréhension[36] et une explicabilité d'une décision algorithmique[37].

33 Voir Art. 13 *AI Act* sur la transparence et fourniture d'informations aux utilisateurs ainsi que EC, European Declaration on Digital Rights and Principles for the Digital Decade, par. 9 (b) : « *assurer un niveau de transparence adéquat quant à l'utilisation des algorithmes et de l'intelligence artificielle, et à faire en sorte que les citoyens soient formés à les utiliser et qu'ils soient informés lorsqu'ils interagissent avec ces technologies* » ; voir aussi CoE, Zero Draft Convention on AI, Art. 15 (*Principle of Transparency and Oversight*).

34 UNESCO, Guidelines for regulating digital platforms, par. 5 : « *Information empowers citizens to exercise their fundamental rights, supports gender equality, and allows for participation and trust in democratic governance and sustainable development, leaving no one behind.* ».

35 DJEFFAL, p. 255-284 ; BERSINI, p. 131.

36 En Italie, la *Corte di Cassazione* a récemment rendu un jugement dans le contexte du RGPD qui impose une connaissance des citoyens du fonctionnement des algorithmes, LA CORTE SUPREMA DI CASSAZIONE, Cassazione civile sez. I, 25/05/2021, (ud. 24/03/2021, dep. 25/05/2021), sentenza n. 14381, <https://i2.res.24o.it/pdf2010/Editrice/ILSOLE24ORE/QUOTIDIANI_VERTICALI/Online/_Oggetti_Embedded/Documenti/2021/05/26/14381.pdf>.

37 Voir le projet de recherche suisse « *Nachvollziehbare Algorithmen* », qui utilise le terme « *Nachvollziehbarkeit* » plutôt que « *Erklärbarkeit* », qui est plus réaliste parce qu'il se focalise plutôt sur la traçabilité d'une décision au lieu d'une explication (souvent difficile), <https://ius.unibas.ch/de/e-piaf/nachvollziehbare-algorithmen/> ; BINDER *et. al.* ; THOUVENIN/BRAUN BINDER/LOHMANN ; THOUVENIN *et. al.*, Positionspapier : Ein Rechtsrahmen für KI.

Les recherches empiriques ont montré que les individus perçoivent différemment les décisions des algorithmes et des humains[38], d'où la nécessité et la valeur ajoutée d'accompagner le processus décisionnel par algorithme d'informations et d'explications.

Ces dernières années, de nombreuses méthodes ont été développées afin d'ouvrir la *black box* des algorithmes et d'essayer d'en expliquer le fonctionnement[39], au moins pour un public spécifique[40]. Néanmoins, la question clé reste de savoir à qui doivent profiter la transparence et la compréhension. Est-il suffisant que les experts en IA et les chercheurs comprennent le fonctionnement de l'IA en question, ou est-ce que l'utilisateur standard doit également pouvoir comprendre l'IA dans les grandes lignes ? Peut-être qu'il faudrait une sorte de *label* pour les consommateurs, qui réduit les explications techniques destinées aux experts à l'essentiel.

Des recommandations en la matière, comme celles du *European Law Institute* (ELI)[41] ou du récent rapport des Nations Unies[42], ont souligné l'importance de la transparence et de l'information et peuvent inspirer des solutions.

A. Connaître l'existence et le contenu d'une décision automatisée en matière d'IA

La première étape pour une administration ou une entreprise est de décider de l'utilisation d'un algorithme pour soutenir ou substituer un processus décisionnel. L'utilisation de l'IA doit alors être rendue publique ou communiquée aux personnes ou autorités concernées, d'une part afin qu'elles sachent quelles règles juridiques seront applicables et d'autre part afin de leur

38 Voir les recherches empiriques sur cette question, HERMSTRÜWER/LANGENBACH, Fair governance with humans and machines.

39 Les chercheurs académiques sont aussi importants que des organisations telles que *AlgorithmWatch* (<www.algorithmwatch.ch>), le *AI Now Institute* avec leurs rapports annuels (<https://ainowinstitute.org/our-work.html>), the *Ada Lovelace Institute* (<https://www.adalovelaceinstitute.org>), le *Weizenbaum Institute* (<https://www.weizenbaum-institut.de>) ou le *Alan Turing Institut*e (<https://www.turing.ac.uk>) qui contribuent massivement aux discours académiques et scientifiques ainsi qu'à la compréhension de l'IA.

40 Voir HACKER/PASSOTH, p. 343-373.

41 Voir ELI, Guiding Principles for Automated Decision-Making.

42 Voir UNITED NATIONS, A/77/196, Principles underpinning privacy and the protection of personal data, par. 45 (« *One of the principles relating to the processing of personal data is that controllers must process data* [...]. *In accordance with this principle, controllers must inform subjects of the processing conditions to which their personal information will be subject from the time of collection, so that subjects are in a position to exercise due control over the data.* »).

permettre de faire valoir leurs droits dans des cas de violations des droits individuels (protection des données et non-discrimination)[43].

Dans le contexte administratif, si les autorités publiques utilisent des algorithmes, le droit d'accès devrait inclure l'accès aux informations sur l'utilisation d'un algorithme ainsi que sur son fonctionnement, et ce afin de permettre un éventuel recours contre un acte administratif devant les tribunaux[44]. Dans le contexte civil, ceci impliquerait par exemple d'accéder aux éléments de preuves générés par l'algorithme qui sont nécessaires afin d'établir les éléments constitutifs d'une discrimination ou une discrimination *prima facie*.

B. Expliquer les décisions de l'IA

Dans la doctrine, plusieurs points de vue[45] ont été exposés concernant l'existence[46] ou non[47] d'un droit à l'explication dans le RGPD. Ceux-ci pourront servir d'exemple (à imiter ou à corriger) dans le contexte de la régulation des algorithmes afin d'éviter les discriminations.

Ce qui est important pour la présente discussion, c'est l'accès aux informations utiles et utilisables pour les humains[48]. Pour cela, il est impératif de connaître les grandes lignes du fonctionnement, le type d'algorithme, les données utilisées. Ces éléments pourraient être présentés de façon succincte par des *labels*, sans rentrer dans les détails.

43 Voir par exemple, CoE, Zero Draft Convention on AI, Art. 20 par. 2 : « *(..) any person has the right to know that one is interacting with an artificial intelligence system rather than with a human and, where appropriate, shall provide for the option of interacting with a human in addition to or instead of such system.* ».

44 Voir BVGer, C-5007/2019 du 1er juin 2022 qui traitait de la question de savoir si le droit d'accès incluait bien les algorithmes ; BVGer, B-626/2016, par. 7.2.1 (sur le droit d'accès et les algorithmes). Le droit général d'accès découle de l'art. 29 de la Constitution fédérale du 18 avril 1999 de la Confédération Suisse (RS 101) en combinaison avec les art. 61 par. 2 BBG, 26 VwVG (le principe) et 27 VwVG (les exceptions).

45 Begley/Schwedes/Frye/Feige ; Hacker/Krestel/Grundmann/Naumann ; Kaminski, The right to explanation, explained ; Malgieri/Comande ; Wachter/Mittledstadt/Floridi, p. 76-99.

46 Voir not. Kaminski, The right to explanation, explained, p. 209-217.

47 Contre l'existence d'un tel droit, Wachter/Mittledstadt/Floridi, p. 76-99.

48 Selbst/Powles, p. 48.

En Europe[49] et en Suisse[50], il y a plusieurs projets de recherche en cours essayant d'éclairer le sujet de la compréhension et de l'explicabilité des algorithmes. Afin de faciliter la transparence et l'explicabilité, une publication, notamment des tests, des analyses d'impact et des audits augmenterait probablement la confiance des citoyens dans les algorithmes[51].

C. Enregistrement du processus décisionnel automatisé

L'utilisation des algorithmes n'est pas seulement basée sur les données, elle produit également une large quantité de données. Afin d'atteindre plus de transparence et de permettre éventuellement d'expliquer le processus de prise de décision de l'algorithme, l'enregistrement du processus décisionnel s'impose pour créer des traces[52]. Ces traces peuvent être utilisées par des victimes potentielles, par exemple pour prouver une éventuelle discrimination dans le cadre d'une procédure de recrutement, et par les employeurs pour démontrer l'absence de discrimination. À ces fins, l'*AI Act* prévoit une documentation technique (art. 11) et un enregistrement des données qui doivent permettre de garantir

> « *un degré de traçabilité du fonctionnement du système d'IA tout au long de son cycle de vie qui soit adapté à la destination du système.* » (art. 12 par. 1 et 2)[53].

---

49 Voir pour un projet européen avec collaboration de *AlgorithmWatch CH*, qui a démarré en novembre 2022, FINDHR: an interdisciplinary project to prevent, detect, and mitigate discrimination in AI, <https://www.mpi-sp.org/43595/news_publication_18991677_transferred> ; un projet allemand de l'université de Hannovre "Bias and Discrimination in Algorithmic Decision-Making" explore la question *"How can we ensure that big data analysis and algorithm-based decision-making are unbiased and nondiscriminatory?"*, <https://www.bias-project.org>.

50 Voir notamment le projet "Nachvollziehbare Algorithmen: ein Rechtsrahmen für den Einsatz von Künstlicher Intelligenz", https://ius.unibas.ch/de/e-piaf/nachvollziehbare-algorithmen/ (consulté le 28 mars 2023).

51 SUSSKIND, p. 194 ; Voir COURT OF AUDITORS NL, Algorithm Audit, not. p. 38 sur les biais.

52 Voir pour une telle obligation du point de vue de la protection des données, MÉTILLE, Le traitement des données personnelles sous l'angle de la nLPD, p. 16 ; voir aussi COE, Zero Draft Convention on AI, Art. 19 lit. a. (« […] *the relevant usage of the system is recorded* […] »).

53 Ces obligations sont spécifiées par l'art. 18 *AI Act* concernant l'obligation d'établir une documentation technique et l'art. 20 *AI Act* sur les journaux générés automatiquement par les systèmes d'IA à haut risque. L'annexe IV de l'*AI Act* finalement prévoit des spécifications à respecter.

Cela n'aidera pas seulement dans de tels cas spécifiques, mais contribuera plus globalement à une plus grande transparence des algorithmes[54].

Afin de protéger les données des employés et d'éviter des discriminations, les encadrements juridiques issus du *algorithmic work management*[55] deviennent de plus en plus importants, non seulement pour permettre aux employeurs de procéder à des tests et vérifications du bon fonctionnement des algorithmes, mais aussi et surtout pour permettre un accès ultérieur aux données afin d'établir une éventuelle discrimination.

D. Rôle des chercheurs

En ce qui concerne la transparence et l'explicabilité des algorithmes, les chercheurs, notamment en sciences informatiques, en éthique et en sciences juridiques, ont le rôle de diriger vers les différentes méthodes d'analyse des décisions des algorithmes et d'attirer l'attention des juristes et des *policy makers* afin que celles-ci soient intégrées dans les lois.

Afin de garantir la transparence, le code des algorithmes pourrait être publié dans son intégralité ou en partie. Cette publication devrait idéalement se faire en ligne (si cela est possible au vu des droits de la propriété intellectuelle) ou du moins mise à disposition des chercheurs afin de tester, vérifier ainsi qu'améliorer le code en vue du respect des obligations légales[56].

Même si les développeurs des entreprises de l'IA sont dotés de compétences comparables, l'avantage du recours aux chercheurs (indépendants et sans financement ni conflit d'intérêt) par rapport aux entreprises utilisant des algorithmes et aux administrations se trouve notamment dans leur expertise indépendante afin d'identifier des problèmes et solutions pour des algorithmes spécifiques. Ceci peut aider à la fois les entreprises et les autorités publiques[57].

Finalement, la transparence et l'explicabilité offrent aux chercheurs la possibilité de détecter des discriminations qui restent souvent cachées en l'absence d'une analyse automatisée approfondie[58]. Les chercheurs jouent ainsi un rôle crucial afin de faire respecter l'art. 21 du *AI Act* et entamer des mesures correctives :

54 Voir en général BURRELL, p. 10.
55 ALOISI.
56 Voir par COE, Zero Draft Convention on AI, Art. 17.
57 Dans ce sens, la remarque des autorités néerlandaises, voir COLLEGE VOOR DE RECHTEN VAN DEN MENS, pt. 2.
58 HEINRICHS, not. p. 143 et 151.

> « *Les fournisseurs de systèmes d'IA à haut risque qui considèrent ou ont des raisons de considérer qu'un système d'IA à haut risque qu'ils ont mis sur le marché ou mis en service n'est pas conforme au présent règlement prennent immédiatement les mesures correctives nécessaires pour le mettre en conformité, le retirer ou le rappeler, selon le cas. Ils informent les distributeurs du système d'IA à haut risque en question et, le cas échéant, le mandataire et les importateurs en conséquence.* ».

Finalement, l'UNESCO, qui a publié des recommandations sur l'IA[59], plaide également pour un accès des chercheurs aux données afin de comprendre les enjeux, les risques et les opportunités des algorithmes[60]. Concrètement, la suggestion est :

> « *Digital platforms should provide access to non-personal data and anonymised data for vetted researchers that is necessary for them to undertake research on content to understand the impact of digital platforms. This data should be made available through automated means, such as application programming interfaces (APIs), or other open and accessible technical solutions allowing the analysis of said data.* »[61].

Cela peut inclure notamment d'impliquer la communauté de chercheurs :

> « *Develop and launch inclusive structured community feedback mechanisms to eliminate gender bias in generative AI and generative algorithms producing content that perpetuates or creates gendered disinformation or harmful or stereotypical content.* »[62].

In fine, il incombe aux développeurs d'algorithmes d'en faire davantage, notamment en permettant aux chercheurs de différentes disciplines d'accéder aux modèles et données des algorithmes afin d'assurer plus de transparence, de compréhension et d'acceptation[63].

59 Voir UNESCO, Recommendation on the Ethics of AI, not. par. 87-93.
60 UNESCO, Guidelines for regulating Digital Platforms, par. 72-74.
61 UNESCO, Guidelines for regulating Digital Platforms, par. 72.
62 UNESCO, Guidelines for regulating Digital Platforms, par. 98 (d).
63 JATON, p. 7.

IV. Confiance et droit au « *human-in-the-loop* »

L'utilisation des algorithmes pour remplacer ou compléter les décisions humaines, nécessite une acceptation sociétale et une confiance des citoyens[64] dans les systèmes algorithmiques[65]. Dans ce contexte, la doctrine et des projets législatifs parlent souvent de confiance dans l'IA ou d'IA digne de confiance[66] et prévoient l'implication d'un humain dans le processus décisionnel qui pourra influencer, mettre fin ou intervenir dans la prise de décision de l'algorithme. Il est clair que l'humain est toujours impliqué « dans » les algorithmes au sens de *human-in-the-loop*[67], car tous les algorithmes, du moins au début, sont pensés, créés et mis en place par des humains[68]. Inclure l'humain dans la surveillance de l'algorithme devient plus compliqué lorsqu'il s'agit d'algorithmes de type *machine learning*, *deep learning etc*.[69]. Dans ces cas, la question de la décision deviendra essentielle ; si c'est un humain qui décide (ou peut décider) ou si la machine prend les décisions toute seule. Inclure l'humain au cours du processus de décision des algorithmes constitue alors la garantie d'une transparence et confiance[70]. Inclure les humains dans le processus prévient aussi les risques d'*automation bias*[71]. Tel qu'il a été démontré dans la section *supra* II., la question est étroitement liée à la connaissance par l'individu du fait que la décision est prise par un algorithme et si possible de quelle manière exactement ou à quel pourcentage l'algorithme a été impliqué dans le processus décisionnel.

---

64 La confiance est utilisée telle que définie par le Larousse comme « *sentiment d'assurance* » ou « *sentiment de quelqu'un qui se fie entièrement à quelqu'un d'autre, à quelque chose* », <www.larousse.fr>.

65 Voir 2023. European Declaration on Digital Rights and Principles for the Digital Decade 2023/C 23/01., par. 9 (a) : « *promouvoir des systèmes d'intelligence artificielle axés sur l'humain, fiables et éthiques tout au long de leur mise au point, de leur déploiement et de leur utilisation, conformément aux valeurs de l'UE.* » et para. 9 (c) : « […] *permettre une surveillance humaine de tous les résultats qui affectent la sécurité et les droits fondamentaux des citoyens* ».

66 Voir par exemple le projet du *AI Act* qui utilise 41 fois le terme « *trust* » ou « *trustworthy* » et qui explique dans l'exposé des motifs : « *La présente proposition vise à mettre en œuvre le deuxième objectif, relatif à la mise en place d'un écosystème de confiance, en proposant un cadre juridique pour une IA digne de confiance* ».

67 Voir en général sur ce sujet et l'expertise humaine, PASQUALE ; CROOTOF/KAMINSKI/PRICE II, not. p. 45-49 (sur le rôle des humains qui consisterait surtout dans la correction des erreurs et des biais).

68 ENARSSON/ENQVIST/NAARTTIJÄRVI, p. 123-153.

69 Voir pour l'apprentissage des algorithmes, notamment LE CUN, not. p. 298-299.

70 Voir comme l'illustration l'art. 14 par. 1 *AI Act* sur le contrôle humain qui le suggère « *au moyen d'interfaces homme-machine appropriées, un contrôle effectif par des personnes physiques pendant la période d'utilisation du système d'IA* ». L'art. 14 par. 2 spécifie que « *Le contrôle humain vise à prévenir ou à réduire au minimum les risques pour* [...] *les droits fondamentaux* ».

71 Art. 14 par. 4, let. b *AI Act*.

Un droit de ne pas être soumis à un processus de décision automatisé pourrait aussi être envisagé[72], étroitement lié au droit à la supervision par un humain du processus de l'IA et à l'intervention d'un humain si besoin ou si souhaité par un individu qui exerce ce droit. À cet égard, l'IA pourrait modifier la façon dont les humains prennent des décisions et la façon dont les humains se perçoivent[73]. Les *Guiding Principles* 9 et 10 du ELI sont importants dans ce contexte[74], car ils prévoient une supervision humaine de l'IA qui devrait être raisonnable et proportionnelle, notamment au vu des coûts et des efforts engendrés pour les entreprises utilisant des systèmes d'IA.

Le jugement de la CJUE C-817/19 *Ligue des Droits Humains*[75], qui est à ce jour la seule décision de la CJEU traitant des questions d'algorithmes et de *machine learning,* éclaire l'utilisation d'algorithmes (notamment le *machine learning*) et les limites imposées par le droit européen, par exemple dans le cas d'une directive qui requiert des critères définis et préétablis :

> « *Cette exigence s'oppose à l'utilisation de technologies d'intelligence artificielle dans le cadre de systèmes d'autoapprentissage (machine learning), susceptibles de modifier, sans intervention et contrôle humains, le processus de l'évaluation et, en particulier, les critères d'évaluation sur lesquels se fonde le résultat de l'application de ce processus ainsi que la pondération de ces critères.* »[76].

A. Art. 22 RGPD

La littérature abondante est en constante évolution au vu des récentes propositions adoptées en matière d'IA, mais la majorité des auteurs souligne qu'il existe un droit de ne pas être soumis à une décision automatisée[77].

---

72 CABITZA *et. al*..

73 Voir pour cela l'ouvrage de NOWOTNY, not. p. 111-126.

74 ELI, Guiding Principles for Automated Decision-making in the EU, not. les principes 9, p. 22 et (Guiding Principle 9: Human oversight/action. The operator shall ensure reasonable and proportionate human oversight over the operation of ADM taking into consideration the risks involved and the rights and legitimate interests potentially affected by the decision.) et principe 10, p. 24 (Guiding Principle 10: Human review of significant decisions Human review of selected significant decisions on the grounds of the relevance of the legal effects, the irreversibility of their consequences, or the seriousness of the impact on rights and legitimate interests shall be made available by the operator.).

75 CJUE, arrêt C-817/19 du 21 juin 2022, *Ligue des Droits Humains*, par. 194.

76 CJUE, arrêt C-817/19 du 21 juin 2022, *Ligue des Droits Humains*, par. 194.

77 ROIG ; TOSONI, p. 145-162 ; DJEFFAL, p. 255-284.

On peut dire que dans le RGPD, l'UE a fait le choix conscient d'inclure l'humain dans le processus de décision automatisé, comme l'indique l'art. 22 par. 1 RGPD :

> « *La personne concernée a le 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.* »[78].

On pourra aussi parler d'un droit à s'opposer à l'IA dans des contextes de prise de décision spécifiques[79]. Ce droit/cette possibilité prévu(e) par le RGPD trouve toute sa signification dans le cadre des décisions automatisées ayant un potentiel de discrimination.

B. Art. 21 LPD

L'art. 21 LPD[80] intitulé « *Devoir d'informer en cas de décision individuelle automatisée* » précise à son paragraphe 1 que :

> « *Le responsable du traitement informe la personne concernée de toute décision qui est prise exclusivement sur la base d'un traitement de données personnelles automatisé et qui a des effets juridiques pour elle ou l'affecte de manière significative (décision individuelle automatisée).* »[81].

Le droit suisse prévoit donc un devoir d'informer le citoyen de l'utilisation exclusive d'un algorithme pour la prise de décision. Ce qui poserait par contre un problème en pratique, ce sont les scénarios dans lesquels les algorithmes contribuent à la prise de décision à une certaine hauteur (p. ex. 50 %, 75 %, 90 %) sans pour autant qu'une décision soit prise sans intervention humaine. Ici, le risque qu'un humain suive aveuglément la suggestion de l'IA pourrait être comparé à une situation ou la décision humaine est substituée à 100 % par un algorithme, ce qui semble être pris en compte par exemple dans la proposition du CdE[82].

78 Voir dans ce sens, HOOFNAGLE/VAN DER SLOOT/BORGESIUS, spécialement p. 68.
79 KAMINSKI/URBAN, p. 1957-2048.
80 Loi fédérale sur la protection des données (LPD) du 25 septembre 2020, RS 235.1.
81 Voir aussi MÉTILLE, p. 13.
82 CoE, Zero Draft Convention on AI, Art. 20 para. 1, qui parle de *substantially informs* : « […] *where an artificial intelligence system substantially informs or takes decision(s)* ».

C. Art. 14 *AI Act* et Art. 20 CdE *Zero Draft Convention on AI*

Les propositions de l'UE et du CdE sont pour l'instant les seules propositions législatives qui incluraient un tel droit au *human-in-the-loop* et qui concerneraient les scénarios ayant un impact pour le genre ou d'autres caractéristiques protégées dans le cadre d'une discrimination causée par un système de recrutement automatisé[83].

Le contrôle humain prévu par l'art. 14 du *AI Act* devrait être mis en place avant toute mise sur le marché et rendre possible aux personnes chargées d'effectuer le contrôle humain d'atteindre cinq objectifs : appréhender entièrement les capacités et les limites du système d'IA et surveiller son fonctionnement, afin de pouvoir détecter et traiter dès que possible les signes d'anomalies, de dysfonctionnements et de performances inattendues (a), avoir conscience d'une éventuelle tendance à se fier automatiquement ou excessivement aux résultats produits par un système d'IA à haut risque (« biais d'automatisation ») (b), être en mesure d'interpréter correctement les résultats du système d'IA à haut risque (c), être en mesure de décider, dans une situation particulière, de ne pas utiliser le système d'IA (d) et être capable d'intervenir sur le fonctionnement du système d'IA à haut risque ou d'interrompre ce fonctionnement au moyen d'un bouton d'arrêt ou d'une procédure similaire (e)[84].

Le projet législatif du CdE prévoit par exemple le droit à un examen humain (*human review*)[85].

Une question qui s'impose est de savoir si dans la collecte et l'utilisation des données une approche *discrimination awareness* est une possibilité[86]. Ceci dépendra largement de l'implication des chercheurs.

D. Rôle des chercheurs

Il est envisageable que les chercheurs fassent partie intégrante des commissions d'éthique, comités de surveillance ou de régulation établis à l'échelle européenne ou nationale. Toutefois, vu que les instances régulatrices traitent des questions d'algorithmes existants et prêts à être commercialisés ou

---

83 Pour l'UE, ceci en raison de la qualification des systèmes de recrutement comme IA à haut risque conformément à l'art. 6 conjointement avec l'Annexe 2. Pour le CdE, l'art. 12 CoE, Zero Draft Convention on AI prévoit l'application de principe d'égalité et de non-discrimination.

84 Art. 14 (4) a) - e).

85 CoE, Zero Draft Convention on AI, Art. 20 para. 1 : « […] *where an artificial intelligence system substantially informs or takes decision(s) affecting human rights and fundamental freedoms there is a right to human review of the decisions.* ».

86 Berendt/Preibusch, p. 135-152.

qui sont déjà sur le marché, il est également important d'intégrer les questions éthiques et juridiques dès la conception des algorithmes[87]. Au plus les risques de discrimination sont connus dès la phase de conception des algorithmes, au plus les chances de réduire les biais et possibles discriminations sont élevées. Le rôle des chercheurs est d'abord pédagogique[88], afin de traduire en pratique la transparence tel que définie dans les propositions législatives. Ils pourront également alerter les citoyens, les entreprises utilisant l'IA et les administrations qui supervisent l'utilisation dans les cas où la substitution ou l'assistance de décisions humaines par l'IA pourrait porter atteinte aux principes de non-discrimination et donc recommander ou dissuader l'utilisation de l'IA compte tenu des risques. En ce qui concerne le droit d'information, le rôle des chercheurs est notamment d'expliquer le fonctionnement de l'IA dans des termes compréhensibles et de contribuer à la communication des informations essentielles aux utilisateurs finaux. Finalement, concernant la mise en place du *human-in-the-loop*, il est clair que la présence et/ou la disponibilité des chercheurs pour mettre en œuvre de tels droits sont importantes.

V. Sphère privée et non-discrimination dès la conception et par défaut

En ce qui concerne la discrimination algorithmique, le droit de la non-discrimination peut s'inspirer du droit de la protection des données en utilisant les concepts de vie privée dès la conception et par défaut (A.) et pour établir des concepts similaires de non-discrimination par conception et par défaut (B.). Comme déjà plus ou moins accepté pour la protection des données, le principe de non-discrimination de tout système algorithmique devrait être non seulement intégré dès la conception[89], mais également internalisé par défaut par toute entreprise créant ou utilisant des algorithmes.

A. Sphère privée dès la conception et par défaut

Peter SCHAAR, ancien préposé fédéral à la protection des données (*Bundesdatenschutzbeauftragter)* en Allemagne qui a largement contribué à renforcer la protection des données au niveau national et européen avait déjà

87 BAUMER, p. 2.

88 BERSINI, p. 84 qui suggère l'implication des informaticiens et des citoyens pour contrer l'analphabétisme informatique et faciliter une meilleure compréhension des algorithmes.

89 Pour quelques réflexions sommaires concernant le concept de *non-discrimination dès la conception* sans pour autant traiter les règles juridiques de la protection des données, voir LÜTZ, Discrimination by correlation, not. p. 278-279.

souligné l'importance de la sphère privée dès la conception en 2010[90]. Dix ans plus tard en 2020, il a réitéré la nécessité d'utiliser des solutions techniques afin d'assurer la protection des données[91]. Dans le droit de l'UE, on retrouve le modèle d'une régulation dès la conception notamment dans le droit de la protection des données, avec le concept de la sphère privée dès la conception et surtout à l'art. 25 RGPD[92]. Le but de ce concept est d'intégrer dans la conception de tout mécanisme qui utilise des données dans le sens du RGPD des mesures qui protègent les données et protègent la sphère privée des utilisateurs[93].

B. Non-discrimination dès la conception and par défaut

On pourra s'inspirer des méthodes et du cadre législatif instaurés par le RGPD afin d'élaborer un processus similaire pour le droit de la non-discrimination. Même si plusieurs problèmes persistent concernant cette approche en *privacy law*[94] et si certains outils du RGPD ne sont pas encore mis en place comme prévu, ce modèle semble la bonne voie à suivre.

D'abord, les chercheurs étaient largement impliqués lors du développement du concept de non-discrimination dès la conception[95]. La non-discrimination par défaut signifie que les principes juridiques de non-discrimination sont intégrés ou pris en compte le mieux possible lors de la conception des algorithmes[96]. Cela pourrait contribuer à éliminer des biais ou des discriminations dès la conception des algorithmes.

Ensuite, l'approche non-discrimination par défaut, requiert non seulement l'intégration technique dans le code des algorithmes, mais également et avant tout une connaissance, ouverture et compréhension des concepts juridiques de non-discrimination. Sans cela, il est difficilement envisageable que les concepteurs des algorithmes puissent intégrer de manière cohérente des spécifications et paramètres dans le modèle et le code des algorithmes qui soutiennent l'objectif

90 SCHAAR, Privacy by design, p. 267-274.
91 SCHAAR, Datenschutz, p. 179-185.
92 RUBINSTEIN/GOOD, p. 37-56 ; LAGIOIA/SARTOR, p. 75.
93 Voir CAVOUKIAN/DIXON ; SHEN/PERASON. Voir not. Comité européen de la protection des données, Lignes directrices 4/2019 relatives à l'article 25, Protection des données dès la conception et protection des données par défaut, Version 2.0 Adoptées le 20 octobre 2020, <https://edpb.europa.eu/system/files/2021-04/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_fr.pdf>.
94 GÜRSES/TRONCOSO/DIAZ, p. 25.
95 Voir TILBURG UNIVERSITY, Report Non-discrimination by design.
96 Pour une explication pratique sur comment les principes de non-discrimination devraient être intégrés lors de la conception des systèmes d'IA, voir REBSTADT *et al*., p. 495-511.

de non-discrimination[97]. Il est donc nécessaire de former les ingénieurs et développeurs en leur fournissant quelques bases en droit de la non-discrimination, ou au moins d'intégrer dans les toolkits ou les logiciels préfabriques et les bases de données qui servent comme exemple pour les développeurs des paramètres et spécifications qui permettent d'utiliser des blocs contenant déjà ces paramètres qui essaient de mettre en œuvre ce principe de non-discrimination.

C. Quel modèle à suivre pour la discrimination algorithmique ?

> « *The mechanics of modern algorithms offered promises of transparency and of equal, dispassionate treatment – behind the veil of ignorance – without making distinctions based on prohibited demographic characteristics such as race or gender.* »[98].

La réalité est toutefois très différente, comme cela a été montré dans la présente contribution, qui a développé la nécessité de transparence des algorithmes qui sont loin d'être neutres et qui requièrent une supervision afin d'empêcher des biais et discriminations. Détailler le modèle de régulation idéal pour contrer la discrimination algorithmique dépasserait largement l'espace et le but de la présente contribution[99]. Néanmoins, il est évident que les outils et les droits et éléments clés qui sont présentés ici devront faire partie intégrante de toute régulation. Idéalement, des outils comme l'analyse d'impact ou les audits devraient être intégrés dans un cadre juridique contraignant afin de pouvoir assurer efficacement la protection des droits humains.

D. Le rôle des chercheurs

Les chercheurs sont impliqués dès la conception des algorithmes, notamment en se basant sur la recherche effectuée par le monde académique[100], mais aussi par les chercheurs associés aux grandes entreprises technologiques[101].

97 THOMPSON, not. p. 1-3.

98 BURREL/FOURCADE, p. 222.

99 Pour des pistes de solution en Europe, voir LÜTZ, Gender Equality and AI in Europe. Pour l'exemple de la discrimination algorithmique dans le recrutement automatisé, voir LÜTZ, Discrimination algorithmique.

100 Il faut néanmoins garder en tête qu'une récente étude a déterminé que plus que la moitié des chercheurs a reçu un financement des grandes entreprises technologiques, voir GOFMAN/JIN ; voir aussi UNESCO/MILA, Missing Links in AI Governance, p. 35 ss (concernant l'industrie IA, l'éthique et l'équité).

101 Les chercheurs employés par ces entreprises ont publié des milliers des articles scientifiques sur l'IA : *Alphabet* (9000), *Microsoft* (8000) et *Meta* (4000) selon Big tech and

Ces recherches peuvent ensuite être utilisés dans les tests des systèmes d'IA à l'interne ou l'externe (p. ex. à travers des *hackathons*), et servir de base pour des conseils aux entreprises et pouvoirs publics. Finalement, la présence dans des structures publiques de surveillance peut contraindre une utilisation éthique et conforme aux règles juridiques en vigueur. Durant toute leur implication, il est essentiel que les chercheurs intègrent et respectent la conception et les valeurs des droits fondamentaux et le principe de la non-discrimination[102].

VI. Analyse d'impact pour les données et les algorithmes

Le RGPD prévoit à l'art. 35 par. 1 l'analyse d'impact pour la protection des données :

> « *Lorsqu'un type de traitement, en particulier par le recours à de nouvelles technologies, et compte tenu de la nature, de la portée, du contexte et des finalités du traitement, est susceptible d'engendrer un risque élevé pour les droits et libertés des personnes physiques, le responsable du traitement effectue, avant le traitement, une analyse de l'impact des opérations de traitement envisagées sur la protection des données à caractère personnel.* ».

Ces outils d'analyse d'impact sont incorporés dans le droit et envisagés dans les cadres juridiques et politiques proposés[103] ainsi que vivement discutés dans la doctrine[104].

Le rôle des chercheurs consiste notamment en la conception des outils permettant de détecter des biais et des discriminations[105] et à consulter les entreprises

artificial intelligence, Mastering the machine, *The Economist*, 1er Avril 2023, p. 54-55. Les centres de recherche en IA de grandes entreprises technologiques comme *Microsoft Research* emploient environ 8000 chercheurs en IA et sont donc souvent plus larges que les départements des grandes universités du monde, voir JUROWETZKI/HAIN/MATEOS-GARCIA/STATHOUPOULOS ; BISHOP, Microsoft AI Research.

102 EC, European Declaration on Digital Rights and Principles for the Digital Decade, par. 9 (f) : « *prendre des mesures pour faire en sorte que la recherche en matière d'intelligence artificielle respecte les normes éthiques les plus élevées et la législation pertinente de l'UE* ».

103 *AI Act* (art. 29 par. 6 qui réfère aux obligations découlant du RGPD) ; OECD, Recommendation on AI, not. V.) 2.3(b) (« *assessment mechanisms* ») ; CONSEIL DE L'EUROPE, Recommendation on the human rights impacts of algorithmic systems (qui fait référence aux *Human Rights Impact Assessment* 15 fois dans différents contextes).

104 MOSS/WATKINS/METCALF/ELISH.

105 Voir par exemple *Aequitas Bias & Fairness Audit* et leur *Bias and Fairness Audit Toolkit*, <http://aequitas.dssg.io> dont le code est disponible sur *GitHub.com* en open access: <https://github.com/dssg/aequitas> et où les entreprises peuvent auditer leurs propres données à travers leur site web.

et les administrations pour la mise en œuvre des outils. Au niveau européen, par exemple, le *European Centre for Algorithmic Transparency* a été créé auprès du *Joint Research Centre* (JRC) de la Commission européenne pour le suivi des questions de transparence algorithmique en lien avec le DSA :

> « *In addition to the Commission's supervisory role, such risk assessments and any accompanying mitigation measures will be subject to an external independent audit and oversight by researchers and civil society.* »[106].

Pour le droit de la non-discrimination, pour l'instant, sur la base du *AI Act*, seuls les systèmes d'IA à haut risque seront soumis à conditions. Cela concerne par exemple les systèmes de recrutement automatiques, mais pas d'autres applications, ce qui laisse un large vide par rapport au champ d'application classique du droit de la non-discrimination. Ce vide devra être comblé par d'autres outils législatifs, par exemple le projet du Conseil de l'Europe ou le droit national.

A. Analyse d'impact dans la protection des données

L'analyse d'impact est prévue dans le RGPD à l'art. 22 par. 3 qui spécifie que :

> « *le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision*. ».

L'intervention humaine est donc en théorie une possibilité qui pourrait être requise par l'individu afin de faire valoir son droit à la non-discrimination par exemple. Pour cela, l'individu doit être informé du fait qu'une décision est automatisée (de par l'utilisation d'un algorithme), par exemple aux fins du traitement de sa demande de crédit ou de sa candidature à une offre d'emploi.

De plus, en vertu de l'art. 22 par. 1,

> « *la personne concernée a le 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*. ».

[106] <https://algorithmic-transparency.ec.europa.eu/about_en>.

Donc, au-delà de demander l'intervention humaine dans le processus algorithmique, un individu peut demander à ne pas faire l'objet d'une décision automatisée, requérant en quelque sorte une intervention plus forte en comparaison avec le scénario d'acceptation d'une décision d'un algorithme sous une surveillance humaine accrue.

En principe, les exceptions prévues notamment à l'art. 22 par. 2 – nécessité à la conclusion ou exécution d'un contrat, autorisation par le droit national ou européen et le consentement – pourraient s'appliquer. Il n'est néanmoins pas très clair dans quel contexte par exemple un algorithme pourrait être requis à la conclusion d'un contrat de travail lors d'une procédure de recrutement (l'alternative étant de ne pas utiliser un algorithme), car le simple fait de dire qu'on utilise un algorithme pour le recrutement suffira à faire jouer l'exception établie par l'art. 22 RGPD. Le consentement, bien connu et débattu dans le cadre du droit de la protection des données, s'avère délicat s'il s'agit par exemple d'une procédure de recrutement, car refuser le consentement résulterait probablement dans l'impossibilité de postuler pour l'offre d'emploi. Dans ce cas spécifique, on pourrait parler d'un consentement qui n'est pas libre.

Avant la mise sur le marché, les analyses d'impact peuvent détecter les biais[107] ou stéréotypes, par exemple concernant le genre[108] ainsi que de possibles discriminations. Une telle vérification peut être mise en place soit par les entreprises elles-mêmes, soit par le pouvoir public, soit par des entreprises tierces ou chercheurs indépendants.

B. Analyse d'impact afin d'éviter les discriminations algorithmiques

Se distinguant de l'analyse d'impact dans le domaine de la protection des données, l'analyse d'impact algorithmique ou l'analyse des biais d'audit a pour but d'identifier des biais, des stéréotypes ou des risques de discriminations[109]. Il s'agit d'une certaine manière de tester les algorithmes avant leur

107 Voir ZERILLI, p. 43-45.

108 WAJCMAN/YOUNG/FITZMAURICE, not. p. 15 (concernant les biais et les *feedback loops*).

109 Voir pour des efforts d'opérationnaliser le concept de biais, JATON, p. 2-3 ; pour des propositions de régulation au niveau globale UNITED NATIONS, Commission on the Status of Women (CSW), 67th Conference, March 2023, E/CN.6/2023/3 – draft agreed conclusions (zero draft), para. 45 (u) « *Adopt regulations mandating evaluation and audit requirements for the development and use of artificial intelligence to provide a secure and high-quality data infrastructure and systems that are either continually improved or terminated if human rights violation or gendered bias are identified* », disponible : <https://www.unwomen.org/sites/default/files/2023-02/CSW67%20Agreed%20Conclusions_zero%20draft_1%20February%202023.pdf>.

mise sur le marché au moyen de *datasets* différents permettant de tester et vérifier si un algorithme montre des tendances biaisées ou produit des discriminations visibles. Tel qu'est le cas dans le monde non-algorithmique, l'exclusion *ex ante* de toute discrimination ou tout biais est illusoire, mais l'analyse d'impact permettra d'exclure *ex ante* certains dysfonctionnements graves pouvant causer des discriminations. La standardisation d'une telle procédure sur la base d'un standard technique ou une législation peut donc contribuer à réduire les effets néfastes des algorithmes. S'il est bien entendu toujours possible d'avoir recours à l'arsenal juridique du droit de la non-discrimination afin de faire valoir ses droits, au vu de l'opacité des algorithmes décrite dans la section *supra* II.), et afin de garantir un terrain de jeu plus au moins équivalent au monde d'avant les algorithmes, cette approche semble pouvoir porter ses fruits.

L'art. 29 par. 6 de l'*AI Act* fait référence à l'art. 35 RGPD pour les utilisateurs des systèmes d'IA à haut risque qui peuvent avoir recours aux données fournies sur la base de l'art. 13 *AI Act* afin de se conformer aux RGPD. Bien que préconisés par la majorité des chercheurs et faisant partie des outils préférés des législateurs afin d'adresser le problème de la discrimination algorithmique, les AIAs sont confrontés à un problème. En effet, leur efficacité est conditionnée à la coopération des opérateurs privés, car ce sont ces derniers qui ont accès aux informations, aux données ainsi qu'à l'expertise permettant d'analyser les algorithmes. Même en cas de cadre législatif contraignant, le résultat sera donc fort dépendant de la volonté des développeurs et des utilisateurs des algorithmes, ce qui peut menacer le bon fonctionnement de la régulation[110].

C. *Monitoring*

Après l'analyse d'impact en principe effectuée avant la mise sur le marché du produit IA, le monitoring est généralement mis en place une fois que l'algorithme est disponible pour le grand public. Le *monitoring* vise alors à identifier d'éventuels problèmes sur la base d'une collecte des données relatives à l'utilisation du système d'IA sur le marché. C'est notamment au cours de cette phase que les biais et les discriminations pourront remonter à la surface, les données utilisées par les individus étant forcément différentes de celles utilisées pour l'entraînement de l'algorithme. Il s'agit donc d'observer constamment le bon fonctionnement des algorithmes afin de documenter et de corriger d'éventuels impacts négatifs en matière de non-discrimination. Ceci nécessite une *data governance* bien établie comme requise par l'art. 10 par. 2 *AI Act* :

110 Voir SELBST, p. 117-118 et p. 152-153.

« *Les jeux de données d'entraînement, de validation et de test sont assujettis à des pratiques appropriées en matière de gouvernance et de gestion des données* », notamment en vue d'« *un examen permettant de repérer d'éventuels biais* ».

Dans le *AI Act*, cette étape est désignée comme le « *post-market monitoring* », tandis que le *US Accountability Act* parle de « *augmented decision-making proccesses* »[111].

Au vu de la récente popularité des applications d'IA se basant sur des *Large Language Models* (LLMs), comme *ChatGPT*, *Bard*, *Claude*, *etc*., le monitoring devient de plus en plus important[112].

D. Rôle des chercheurs

Si en principe les développeurs des entreprises et les chercheurs indépendants effectuent des recherches pouvant mener à des produits d'IA qui seront ensuite mise en production, ils poursuivent des objectifs différents. En effet, les développeurs employés par une entreprise ont pour but de développer des produits spécifiques pour le compte de l'entreprise. Par contre, le but poursuivi par les chercheurs n'est pas principalement de développer une application spécifique pour une entreprise, mais de s'engager dans la recherche fondamentale. Le rôle des chercheurs pour l'avancement des technologies et pour empêcher des effets négatifs[113] sur les droits fondamentaux n'est de ce fait pas seulement essentiel, mais ancré dans les droits humains[114]. Quant aux juristes, un de leurs rôles serait sûrement d'informer quelles sont les lois applicables qui doivent être respectées et par conséquent intégrées dans les algorithmes, mais aussi prises en compte lors des analyses d'impact.

---

111 Voir pour une comparaison des approches EU et US, MÖKANDER/JUNEJA/WATSON/FLORIDI, not. p. 752.

112 MÖKANDER *et al*., Auditing LLMs.

113 Voir par exemple l'avertissement des chercheurs en IA sur le développement des LLMs qui sont plus performants que *GPT-4*, "Pause Giant AI Experiments: An Open Letter" of March 22, 2023: "We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.", <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> ; voir aussi CoE, Zero Draft Convention on AI, l'art. 24 par. 3 CoE qui prévoit un moratoire ou une interdiction sur certaines applications d'IA.

114 Voir l'art. 27 de la Déclaration universelle des droits de l'homme et l'art. 15 du International Covenant on Economic, Social and Cultural Rights (ICESCR) qui stipule une sorte de droit aux sciences, ce qui implique naturellement un rôle primordial des scientifiques d'éclairer le public par exemple sur les effets des algorithmes et peut être même un devoir d'anticiper les conséquences (négatives) des algorithmes sur les droits fondamentaux, voir DONDERS/PLOZZA.

Les chercheurs en sciences informatiques et les développeurs, sachant que les techniques informatiques sont en constante évolution, sont notamment chargés d'informer et d'éclairer sur les dernières techniques pouvant être utilisées par exemple afin d'éviter ou de détecter des biais ou des discriminations[115].

Il s'impose d'impliquer les chercheurs en droit et en sciences informatiques qui n'ont aucun lien contractuel avec l'entreprise en cause et qui seront donc considérés comme personnes tierces dans tout le processus d'analyse d'impact et des audits. Les audits pourront être demandés aux entreprises et administrations avant ou après de la mise sur le marché[116]. La preuve de l'implication d'un certain nombre de chercheurs experts dans le domaine pourrait être requise. Les autorités européennes ou nationales devraient pouvoir se reposer sur des informaticiens et des juristes spécialisés dans la discrimination. Afin de pouvoir contrôler et vérifier la mise en place des différentes mesures, il faut garder à l'esprit que l'instrument phare proposé par de multiples chercheurs et repris dans des propositions législatives diverses représente aussi des risques, notamment au vu de la dépendance de l'expertise du secteur privé qui détient le savoir-faire permettant de procéder aux *bias* audits. Afin d'exercer le contrôle par *AI Act* et le monitoring *ex-post*, ceux-ci doivent dès lors être encadrés par des règles juridiques strictes[117], comme c'est le cas par exemple dans le secteur pharmaceutique, où les entreprises mènent les études cliniques elles-mêmes et l'administration procède à des vérifications. De manière similaire, il est envisageable que les employés de l'entreprise elle-même procèdent à la vérification des impacts de l'IA en interne, à condition d'ensuite être soumis au contrôle d'une autorité de surveillance. Afin de garantir plus d'impartialité, une entreprise tierce pourrait également procéder à une telle analyse des impacts, également sous le contrôle (plus sommaire) de l'autorité responsable. Finalement, impliquer les chercheurs académiques comporte l'avantage de l'impartialité et de la transparence, car leur travail est souvent inspiré par l'*open access* en comparaison avec ceux travaillant pour les entreprises privées[118].

115 Voir p. ex. MÖKANDER *et al*., Auditing LLMs, p. 1 qui alertent au fait que les mesures d'*audit* utilisées à présent et prévues dans les propositions législatives peuvent potentiellement encourir des risques pour les LLMs qui sont des systèmes en constante évolution.

116 Voir l'exemple des Pays-bas qui procédent à de telles audits pour le secteur public, COURT OF AUDITORS NL, Algorithm Audit ; RAJI/CONSTANZA-CHOCK/BUOLAMWINI, p. 5-26.

117 Dans ce sens aussi SUSSKIND, p. 194 qui est en faveur des règles contraignantes favorisant l'approche par la loi plutôt que le libre choix des entreprises.

118 BERSINI, p. 147.

VII. Conclusion

L'implication des chercheurs dans les débats avant l'adoption des actes juridiques, en ce compris les groupes de travail[119], les comités élaborant les standards[120], la doctrine et les conférences en général, est évidente. Leurs rôles définis dans les propositions législatives, demeurent cependant moins clair.

En raison de l'importance des données et de l'objectif commun de protéger des droits fondamentaux, l'article a d'abord constaté l'interdépendance, la conflictualité, la complémentarité ainsi que le potentiel d'imitation du droit de la protection des données et du droit de la non-discrimination. Le droit de la protection des données ayant une demi-décennie d'avance en matière de décision automatisée et d'algorithmes, le droit de la non-discrimination peut s'inspirer de ses cadres juridiques, de sa pratique et de ses institutions.

Concrètement, la présente contribution a exposé que le droit de connaitre l'implication d'un algorithme dans la prise de décision est essentiel et constitue souvent la base de toute action juridique ultérieure. Le droit de savoir si et comment un algorithme prend ou contribue à une décision fait partie de ce qui est souvent appelé la transparence. Celle-ci est souvent qualifiée de véhicule (« *tool* »), mais elle constitue en réalité plutôt l'objectif à atteindre.

Ensuite, le droit à l'intervention humaine dans tout processus algorithmique, qui était déjà présent dans le RGPD, est envisagé pour les projets de régulation de l'IA à l'échelle européenne.

En ce qui concerne le concept de non-discrimination dès la conception et par défaut qui découle du *privacy law*, il est clair qu'il devrait constituer un principe de guidage pour la conception des algorithmes, idéalement ancré dans une règle juridique.

Les analyses d'impact nécessaires afin d'éviter les biais et discriminations algorithmiques sont également prescrites par le RGPD. Par conséquent, il convient de créer des synergies entre le RGPD et le futur *AI Act* en la matière. Il en est de même pour le *monitoring* et les analyses *ex-post* tant pour la législation que pour le fonctionnement des algorithmes, qui devraient être surveillés constamment afin de signaler des dysfonctionnements ou des risques pour les droits fondamentaux.

Dans tous les domaines discutés, le rôle des chercheurs est crucial, non seulement avant la mise en place d'un cadre législatif, mais surtout au cours du fonctionnement des algorithmes et la mise en œuvre des lois ainsi que pour le

119 Des groupes d'experts sur l'IA, comme p. ex la *High level expert group on Artificial Intelligence*, <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>.

120 CEN-CENELEC ou ISO p. ex. dont la Suisse est membre.

futur développement et d'éventuelles révisions de celles-ci. Les chercheurs ne sont pas seulement (plus) indépendants que le pouvoir public et les entreprises, il est également dans l'intérêt commun et public de tester et aviser des risques concernant le fonctionnement des algorithmes et de l'application des règles. Afin de permettre aux chercheurs d'effectuer ce travail, il est important d'ancrer leurs rôles dans les textes législatifs ainsi que d'établir les mécanismes pertinents permettant leur participation[121]. Cela implique aussi de donner accès dans un cadre spécial aux données et aux modèles des algorithmes, sans mettre en danger les intérêts des entreprises protégés notamment par la propriété intellectuelle ou des secrets d'affaires.

VIII. Bibliographie

A. Littérature

Serge Abiteboul/Gilles Dowek, Le temps des algorithmes, Paris 2017 ; **Antonio Aloisi**, Regulating Algorithmic Management at Work in the European Union: Data Protection, Non-Discrimination and Collective Rights, International Journal of Comparative Labour Law and Industrial Relations, 2022 (*à paraître*) ; **Solon Barocas/Moritz Hardt/Arvind Narayanan**, Fairness in machine learning, 2022 ; **Eric Baumer**, Toward human-centered algorithm design, Big Data & Society 2017, Nº 4 ; **Tom Begley/Tobias Schwedes/Christopher Frye/Ilya Feige**, Explainability for fair machine learning, 2020 ; **Bilel Benbouzid/Yannick Meneceur/Nathalie Alisa Smuha**, Quatre nuances de régulation de l'intelligence artificielle : Une cartographie des conflits de définition, Réseaux (Centre national d'études des télécommunications (France)) 2022, Nº 232-233, p. 29-64 ; **Bettina Berendt/Sören Preibusch**, Toward accountable discrimination-aware data mining: the Importance of keeping the human in the loop – and under the looking glass. Big data 2017, Nº 5, p. 35-152 ; **Hugues Bersini**, Algocratie : Allons-nous donner le pouvoir aux algorithmes ?, Louvain-La-Neuve 2023 ; **Nadja Braun Binder/Thomas Burri/Melinda Florina Lohmann/Monika Simmler/Florent Thouvenin/Kerstin Noëlle Vokinger**, Künstliche Intelligenz: Handlungsbedarf im Schweizer Recht, Jusletter du 28 juin 2021 ; **Todd Bishop**, One year later, Microsoft AI and Research grows to 8k people in massive bet on artificial intelligence, GeekWire 22 septembre 2017, <https://www.geekwire.com/2017/one-year-later-microsoft-ai-research-grows-8k-people-massive-bet-artificial-intelligence/ (consulté le 5 mars 2023>) (cité : Bishop, Microsoft AI Research) ; **Frederik Borgesius**, Discrimination, artificial intelligence, and algorithmic decision-making, 2018, Study for the Council of Europe ; **Anu Bradford**, The brussels effect, Northwestern University Law Review 2012, 107, p. 1 ss (cité : Bradford, Brussels Effect) ; **Anu Bradford**, The Future of the Brussels Effect, The Brussels Effect, Oxford 2020 (cité : Bradford, The Future of the Brussels Effect) ; **Jenna Burrell**, How the

[121] Voir EC, European Declaration on Digital Rights and Principles for the Digital Decade, par. 9 (e) : « *prévoir des garanties et prendre des mesures appropriées, y compris en promouvant des normes fiables, pour que l'intelligence artificielle et les systèmes numériques soient, en permanence, sûrs et utilisés dans le plein respect des droits fondamentaux*. ».

machine 'thinks': Understanding opacity in machine learning algorithms, Big data & society 2016, N° 3 ; **Jenna Burrell/Marion Fourcade**, The Society of Algorithms. Annual Review of Sociology 2020, p. 47 ss ; **Federico Cabitza/Andrea Campagner/Gianclaudio Malgieri/Chiara Natali/David Schneeberger/Karl Stoeger/Andreas Holzinger**, Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable AI, Expert Systems with Applications 2023, Vol. 213, p. 118888 (cité : Cabitza *et al*.) ; **Ann Cavoukian/Mark Dixon**, Privacy and security by design: An enterprise architecture approach, Information and Privacy Commissioner, Ontario 2013 ; **Kathleen Creel/Deborah Hellman**, The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems, Canadian Journal of Philosophy 2022, p. 1-18 ; **Rebecca Crootof/Margot Kaminski/William Nicolson Price II**, Humans in the Loop, Vanderbilt Law Review, Forthcoming, 2023 ; **Livio Di Tria**, Comparaison entre la nLPD et le RGPD, 12 février 2021, <www.swissprivacy.law/55> ; **Yvonne Donders/Monika Plozza**, Look before you Leap: Anticipation Duties and Responsibilities under the Right to Science, in: International Journal of Human Rights (forthcoming) 2023 ; **Christian Djeffal**, AI, Democracy and the Law, *in* Andreas Sendmann (ed.), The Democratization of Artificial Intelligence, p. 255284, Bielefeld 2020 ; **Therese Enarsson/Lena Enqvist/Markus Naarttijärvi**, Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts, Information & Communications Technology Law 2022, N° 31, p. 123-153 ; **Florian Jaton**, Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application, Big Data & Society 2021, Vol. 8., Nr. 1 ; **Raquel Esther Jorge Ricart/Fiametti Rossetti/Luca Tangi/Vincent Van Roy**, AI watch, national strategies on artificial intelligence: a European perspective, 2022, Publications Office of the European Union (cité : Jorge Ricart et. al., AI watch, national strategies on AI) ; **Francesca Lagioia/Giovanni Sartor**, The impact of the general data protection regulation on artificial intelligence, European Parliament 2021, Publications Office ; **Michael Gofman/Zhao Jin**, Artificial Intelligence, Education, and Entrepreneurship, Journal of Finance 2022, Forthcoming ; **Seda Gürses/Carmela Troncoso/Claudia Diaz**, Engineering privacy by design, Computers, Privacy & Data Protection 2011, 14, p. 25 ; **Philip Hacker**, Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law, Common Market Law Review 2018, p. 55 (cité : Hacker, Teaching fairness to AI) ; **Philip Hacker/Ralf Krestel/Stefan Grundmann/Felix Naumann**, Explainable AI under contract and tort law: legal incentives and technical challenges, Artificial Intelligence and Law 2020, 28, p. 415-439 ; **Philip Hacker/Jan-Hendrik Passoth**, Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Revised and Extended Papers 2022, p. 343-373 ; **Ronan Hamon/Hendrik Junklewitz/Ignacio Sanchez/Gianclaudio Malgieri/Paul de Hert**, Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making, IEEE Computational Intelligence Magazine 2022, 17, p. 72-85 ; **Bert Heinrichs**, Discrimination in the age of artificial intelligence, AI & SOCIETY 2022, 37, p. 143-154 ; **Yoan Hermstrüwer/Pascal Langenbach**, Fair governance with humans and machines. MPI Collective Goods Discussion Paper 2022 ; **Chris Jay Hoofnagle/Bart Van Der Sloot/Frederik Zuderveen Borgesius**, The European Union general data protection regulation: what it is and what it means, Information & Communications Technology Law 2019, 28, p. 65-98 ; **Roman Jurowetzki/Daniel S. Hain/Juan Mateos-Garcia/Konstantinos Stathoulopoulos**, The Privatization of AI Research (-ers): Causes and Potential Consequences--From university-industry interaction to public research brain-drain? arXiv preprint

arXiv:2102.01648, 2021 ; **Florian JATON**, Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application, Big Data & Society 2021, 8(1), 20539517211013569 (cité : JATON, Assessing biases, relaxing moralism) ; **Margot E. KAMINSKI**, The right to explanation, explained. Berkeley Tech 2019, LJ, 34, p. 189 ss (cité : KAMINSKI, The right to explanation, explained) ; **Margot E. KAMINSKI**, The right to explanation, explained, Research Handbook on Information Law and Governance, Cheltenham 2021 (cité : KAMINSKI, The right to explanation, explained) ; **Margot E. KAMINSKI/Jennifer M. URBAN**, The right to contest AI, Columbia Law Review 2021, 121, p. 1957-2048 ; **Miriam KULLMANN**, Algorithmenbasiertes Personalrecruiting: antidiskriminierungs-und datenschutzrechtliche Aspekte, Zeitschrift für Arbeits-und Sozialrecht, 2021, p. 61 ; **Yann LE CUN**, Quand la machine apprend : la révolution des neurones artificiels et de l'apprentissage profond, Paris 2019 ; **Fabian LÜTZ**, How the 'Brussels effect' could shape the future regulation of algorithmic discrimination, Duodecim Astra 2021, 1, p. 142-163 (cité : LÜTZ, Brussels effect) ; **Fabian LÜTZ**, Discrimination by correlation. Towards eliminating algorithmic biases and achieving gender equality, (Dis) Obedience in Digital Societies, Bielefeld 2022, p. 250-293 (cité : LÜTZ, Discrimination by correlation) ; **Fabian LÜTZ**, Gender equality and artificial intelligence in Europe, Addressing direct and indirect impacts of algorithms on gender-based discrimination. ERA Forum 2022, p. 33-52 (cité : LÜTZ, Gender Equality and AI in Europe) ; **Fabian LÜTZ**, Le rôle du droit pour contrer la discrimination algorithmique dans le recrutement automatisé, *in* Florence GUILLAUME (éd.) La technologie, l'humain et le droit, Bern 2023 (cité : LÜTZ, Discrimination algorithmique) ; **Gianclaudio MALGIERI/Giovanni COMANDÉ**, Why a right to legibility of automated decision-making exists in the general data protection regulation, International Data Privacy Law 2017 ; **Sylvain MÉTILLE**, Le traitement de données personnelles sous l'angle de la (nouvelle) loi fédérale sur la protection des données du 25 septembre 2020, SJ 2021 II, p. 1-48 ; **Sylvain MÉTILLE/Annelise ACKERMANN**, RGPD : application territoriale et extraterritoriale. Datenschutzgrundverordnung (DSGVO): Tragweite und erste Erfahrungen = Le règlement général sur la protection des données (RPDG) : portée et premières expériences, 2020 ; **Sylvain MÉTILLE/Livio DI TRIA**, Protection des données: responsabilité croissante ? ; **Jakob MÖKANDER/Prathm JUNEJA/David WATSON/Luciano FLORIDI**, The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? Minds and Machines 2022, 32, p. 751-758 ; **Jakob MÖKANDER/Jonas SCHUETT/Hannah Rose KIRK/Luciano FLORIDI**, Auditing large language models: a three-layered approach, 2023 (cité : MÖKANDER *et al*., Auditing LLMs) ; **Emanuel MOSS/Elizabeth Anne WATKINS/Jacob METCALF/Madeleine Clare ELISH**, Governing with algorithmic impact assessments: six observations, 2020 ; **Helga NOWOTNY**, In AI We Trust: Power, Illusion and Control of Predictive Algorithms, 2021, Polity Press ; **Frank PASQUALE**, New Laws of Robotics: Defending Human Expertise in the Age of AI, Belknap Press 2020 ; **Inioluwa Deborah RAJI/Sasha COSTANZA-CHOCK/Joy BUOLAMWINI**, Change From the Outside: Towards Credible Third-Party Audits of AI Systems, In: Missing Links in AI Policy, UNESCO/MILA Paris 2023 ; **Jonas REBSTADT/Henrik KORTUM/Laura Sophie GRAVEMEIER/Birgit EBERHARDT/Oliver THOMAS**, Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services, HMD 2022 59, p. 495-511 ; **Antoni ROIG**, Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR). European Journal of Law and Technology 2017, Vol. 8, Nº 3 ; **Ira RUBINSTEIN/Nathaniel GOOD**, The trouble with Article 25 (and how to fix it): the future of data protection by design and default, International Data Privacy Law 2020, 10, p. 37-56 ; **Peter SCHAAR**, Privacy by design, Identity in the Information Society 2010, 3, p. 267-274 (cité : SCHAAR, Privacy by

design) ; **Peter SCHAAR**, Datenschutz und Internet – Es ist kompliziert!, Informatik Spektrum 2020, 43, p. 179-185 (cité : SCHAAR, Datenschutz) ; **Andrew SELBST/Julia POWLES**, "Meaningful information" and the right to explanation. Conference on fairness, accountability and transparency, 2018. PMLR, p. 48-48 ; **Andrew SELBST**, An institutional view of algorithmic impact, Harvard Journal of Law & Technology 2021, p. 35 ss ; **Yun SHEN/Siani PEARSON**, Privacy enhancing technologies: A review, 2011, Hewlet Packard Development Company, disponible à : <https://bit. ly/3cfpAKz> ; **Nathalie SMUHA**, From a 'Race to AI' to a 'Race to AI Regulation': Regulatory Competition for Artificial Intelligence 2021 ; **Erica THOMPSON**, Escape from Model Land: How Mathematical Models Can Lead Us Astray and What We Can Do About It, London 2022 ; **Florent THOUVENIN/Nadia BRAUN BINDER/Melinda F. LOHMANN**, Regeln für den Einsatz von künstlicher Intelligenz (Gastkommentar), Neue Zürcher Zeitung, 27 février 2021 ; **Jamie SUSSKIND**, The Digital Republic, On Freedom and Democracy in the 21st Century, London 2022 ; **Florent THOUVENIN/ Markus CHRISTEN/Abraham BERNSTEIN *et. al.***, Positionspapier: Ein Rechtsrahmen für Künstliche Intelligenz ; **Luca TOSONI**, The right to object to automated individual decisions: resolving the ambiguity of Article 22 (1) of the General Data Protection Regulation. International Data Privacy Law 2021, 11, p. 145-162 ; **Marvin VAN BEKKUM/Frederic ZUIDERVEEN BORGESIUS**, Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? Computer Law & Security Review 2023, 48, p. 105770 ; **Sandra WACHTER/Bent MITTELSTADT/Luciano 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, p. 76-99 ; **Judy WAJCMAN/Erin YOUNG/Anna FITZMAURICE**, The digital revolution: Implications for gender equality and women's rights 25 years after Beijing 2020 ; **Betsy Anne WILLIAMS/Catherine F. BROOKS/Yotam SHMARGAD**, How algorithms discriminate based on data they lack: Challenges, solutions, and policy implycations, Journal of Information Policy 2018, 8, p. 78-115 ; **John ZERILLI**, A citizen's guide to artificial intelligence, Boston 2021.

B. Documents officiels

Conseil fédéral, « Intelligence artificielle » – lignes directrices pour la Confédération, Cadre d'orientation en matière d'IA dans l'administration fédérale 2020 ; **Organization for European Cooperation and Development (OECD)**, OECD/LEGAL/0449, Recommendation of the Council on Artificial Intelligence ; **College voor de Rechten van den Mens**, Inbreng College voor de Rechten van den Mens op verzoek van de commissie voor digitale Zaken 2022, 13 mai 2022 ; **Council of Europe**, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, 2020 ; **Council of Europe**, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Nº 108, https://rm.coe.int/1680078b37 ; **Council of Europe**, Preliminary draft Council of Europe study on the impact of artificial intelligence, its potential for promoting equality, including gender equality, and the risks to non-discrimination, GEC(2022)9 CDADI(2022)21 ; **Council of Europe,** Committee on Artificial Intelligence (CAI), Revised zero draft [framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, CAI(2023)01 (cité : CoE, Zero Draft Convention on AI) ; **European Commission (EC)**, European Declaration on Digital Rights and Principles for the Digital Decade, JO 2023/C 23/01 ; **European Law Institute (ELI)**, Guiding Principles for Automated Decision-Making in the EU, Vienne 2022, disponible :

<https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Innovation_Paper_on_Guiding_Principles_for_ADM_in_the_EU.pdf> ; **European Union Agency for Fundamental Rights (FRA)**, Bias in algorithms: artificial intelligence and discrimination, 2022, Publications Office of the European Union ; Netherlands Court of Auditors, An audit of 9 algorithms used by the Dutch government, The Hague 2022, disponible à : https://english.rekenkamer.nl/binaries/rekenkamer-english/documenten/reports/2022/05/18/an-audit-of-9-algorithms-used-by-the-dutch-government/An+Audit+of+Algorithms.pdf (cité : Court of Auditors NL, Algorithm Audit) ; **United Nations**, A/77/196: Principles underpinning privacy and the protection of personal data, 2022 ; **United Nations**, Commission on the Status of Women (CSW), 67th Conference, March 2023, E/CN.6/2023/3 – draft agreed conclusions (zero draft) ; **UNESCO**, Recommendation on the Ethics of Artificial Intelligence, 2021 (cite : UNESCO, Recommendation on the Ethics of AI) ; **UNESCO**, Guidelines for regulating digital platforms: A multistakeholder approach to safeguarding freedom of expression and access to information, 2023 ; **UNESCO/Mila**, Missing Links in AI Governance, Paris 2023 ; **Tilburg University**, Report Non-discrimination by design, 2019, disponible : <https://www.tilburguniversity.edu/sites/default/files/download/07%20Onderzoeksrapport%20non-discriminatie%20by%20design%20%28compr%29.pdf>.

