Conformité Européenne des Systèmes d'IA: Outils Statistiques Élémentaires

Philippe BESSE[1]
Université de Toulouse – INSA, Institut de Mathématiques – UMR CNRS 5219, Université Laval – OBVIA

TITLE

European Compliance of AI Systems: Basic Statistical Tools

RÉSUMÉ

Suite à la publication du livre blanc pour une approche de l'IA basée sur l'excellence et la confiance, la Commission Européenne (CE) a publié de nombreuses propositions de textes réglementaires dont un (*AI Act)* (CE 2021) établissant des *règles harmonisées sur l'intelligence artificielle* (IA). Quels seront les conséquences et impacts de l'adoption à venir de ce texte du point de vue d'un mathématicien ou plutôt statisticien impliqué dans la conception de système d'intelligence artificielle (IA) à haut risque au sens de la CE? Quels outils et méthodes permettent de répondre aux obligations à venir de conformité: analyse rigoureuse et documentée des données traitées, des performances, robustesse, résilience de l'algorithme, de son explicabilité, des risques, pour les droits fondamentaux, de biais discriminatoires? Ces questions sont illustrées par un exemple numérique analogue à un score de crédit (cf. tutoriel) à la recherche d'un moins mauvais compromis entre toutes les contraintes. Nous concluons sur les avancées et limites du projet de règlement pour les systèmes d'IA à haut risque.

Mots-clés : *intelligence artificielle, apprentissage automatique, discrimination, effet disproportionné, AI Act, réglementation européenne.*

ABSTRACT

Following the publication of the white paper for an excellence and trust-based approach to AI, the European Commission (EC) has published numerous regulatory proposals including an AI Act (EC 2021) establishing harmonized rules on artificial intelligence (AI). What will be the consequences and impacts of the upcoming adoption of this text from the point of view of a mathematician or rather a statistician involved in the design of high-risk AI systems as defined by the EC? What tools and methods can be used to reach future compliance obligations? Rigorous and documented analysis of the data and performance, robustness, resilience of the algorithm, its explicability, and the risks of discriminatory bias for fundamental rights. These questions are illustrated by a numerical example analogous to a credit score (cf. tutorial) in search of the least bad compromise between all the constraints. We conclude on the advances and limitations of the proposed regulation for high risk AI systems.

Keywords: *artificial intelligence, machine learning, bias, discrimination, disparate impact, AI Act, european regulation.*

1. philippe.besse@insa-toulouse.fr

1. Introduction

L'adoption en 2018 du Règlement Général de la Protection des Données (RGPD) a profondément modifié les comportements et pratiques des entreprises dans leurs gestions des données, messageries et sites internet. Néanmoins, les condamnations récurrentes des principaux acteurs du numérique, notamment pour abus de position dominante, apportent les preuves de l'inutilité des chartes (*softlaw*) et résolutions éthiques (*ethical washing*). En conséquence et résistant aux accusations fallacieuses de freiner la recherche, l'Europe poursuit sa démarche visant à harmoniser réglementations et innovations technologiques pour le respect des droits humains fondamentaux mais aussi la défense des intérêts commerciaux de l'Union.

La publication par la Commission Européenne (CE) d'un livre blanc sur l'Intelligence Artificielle: une approche européenne axée sur l'excellence et la confiance (CE 2020) fait suite au guide pour une IA digne de confiance rédigé par un groupe d'experts (CE 2019). L'étape suivante est la publication de propositions de règlements dont certains en cours d'adoption :

- *Digital Market Act* (2020) : recherche d'équité dans les relations commerciales et risques d'entraves à la concurrence à l'encontre des entreprises européennes ;
- *Digital Services Act* (2020) : sites de service intermédiaire, d'hébergement, de plateforme en ligne et autres réseaux sociaux ; comment contrôler les contenus illicites et risques des outils automatiques de modération ;
- *Data Governance Act* (2020) contractualisation des utilisations, réutilisations, des bases de données tant publiques que privées (fiducie des données) ;
- *Artificial Intelligence Act* (CE 2021) : proposition de règlement établissant des règles harmonisées sur l'intelligence artificielle.

S'ajoutant au RGPD pour la protection des données à caractère personnel, l'adoption européenne à venir de ce dernier texte (*AI Act*) va profondément impacter les conditions de développements et d'exploitations des systèmes d'Intelligence Artificielle (systèmes d'IA). Cette démarche fait passer d'une IA souhaitée éthique (*ethical AI*), à une *obligation de conformité* (*lawfull AI*) qui confère le marquage "CE" ouvrant l'accès au marché européen. La CE veut ainsi manifester son *leadership* normatif à l'international afin que ce pouvoir de l'UE sur la réglementation et le marché lui confère un avantage concurrentiel dans le domaine de l'IA.

En conséquence, le présent document propose une réflexion sur la prise en compte méthodologique de ce projet de réglementation concernant plus spécifiquement les compétences usuelles en Statistique, Mathématiques, des équipes de développement d'un système d'IA, notamment ceux jugés à haut risque selon les critères européens. Il cible plus particulièrement certaines des sept exigences citées dans le guide des experts (CE 2019), reprises dans le livre blanc (CE 2020) et identifiées comme risques potentiels (Besse et al. 2019) : 1. confidentialité et analyse des données ; 2. précision, robustesse, résilience ; 3. explicabilité ; 4. non-discrimination.

La section 2 suivante extrait de l'*AI Act* les éléments clefs impactant les choix et développements méthodologiques puis la section 3 en commente les conséquences tout en proposant les outils statistiques bien connus de niveau Master et bagage d'un futur *scientifique des données*. Ceux ci semblent adaptés voire suffisants pour satisfaire aux futures obligations réglementaires de contrôle des risques afférents aux systèmes d'IA. Enfin la section 4 déroule un cas d'usage numérique analogue à la prévision d'un score de crédit sur un jeu de données concret. Cet exemple, extrait d'un tutoriel dont le code est librement accessible, permet d'illustrer la démarche de recherche d'un moins mauvais compromis à élaborer entre confidentialité, performance, explicabilité et sources de discrimination. Il souligne les difficultés soulevées par la rédaction de la documentation qui devra accompagner tout système d'IA à haut risque. En conclusion, nous proposons une synthèse des principales avancées

de ce projet d'*AI Act* et en relevons, dans la version d'avril 2021, les principales limites.

2. Impacts techniques de l'*AI Act*

2.1. Structure du projet de règlement

Castets-Renard et Besse (2022) détaillent une analyse du régime de responsabilité *ex ante*[2] proposé dans les 89 considérants[3] et 85 articles structurés en 12 titres de l'*AI Act* : entre auto-régulation, certification, normalisation, pour définir des règles de conformité notamment pour la défense des droits fondamentaux. L'objectif du présent article est plus spécifique, il est focalisé sur les éléments du projet de réglementation concernant directement le statisticien ou scientifique des données impliqué dans la conception d'un système d'IA jugé à haut risque car impactant des personnes physiques.

De façon générale, les considérants, introductifs au projet, listent donc les principes retenus par la CE et qui ont prévalu à la rédaction des articles. La CE insiste sur la nécessité de la construction de normes internationales en priorisant le respect des droits fondamentaux dont la non-discrimination. Consciente de la place occupée par les algorithmes d'apprentissage statistiques, elle souligne la nécessité de la représentativité statistique des données d'entraînement et l'importance d'une documentation exhaustive à propos de ces données et des performances d'un système d'IA. Consciente également de l'opacité de ces algorithmes, elle demande que les capacités d'interprétation de leurs sorties ou décisions en découlant soient à jour des recherches scientifiques en cours et qu'un suivi puisse être assuré grâce à une journalisation ou archivage des décisions et données afférentes.

2.2. Articles les plus concernés

La définition adoptée de l'IA (art. 3) est pragmatique et très flexible en se basant sur la liste exhaustive des algorithmes concernés (annexe I). Les algorithmes d'apprentissage automatique supervisés ou non, par renforcement, constituent actuellement l'essentiel des applications quotidiennes de l'IA. La représentation de connaissances, la programmation inductive et plus généralement les systèmes experts très développés dans les années 70s, restent présents dans certains domaines. Le troisième type d'algorithme cité cible les approches statistiques, inférences bayésiennes et méthodes d'optimisation. Les approches statistiques bayésiennes ou non conduisant très généralement à des prévisions pour l'aide à la décision peuvent être incluses dans la grande famille de l'apprentissage fondée sur des données. En revanche, les méthodes d'optimisation comme par exemple celles d'allocation optimale de ressources des sites d'intermédiation (*e.g.* Uber, ParcourSup,...) nécessitent une approche particulière. Cette liste peut être facilement adaptée en fonction des évolutions technologiques. Ces définitions reconnaissent la place prépondérante de l'apprentissage statistique et donc des données exploitées pour leur construction. Ils laissent de côté les algorithmes procéduraux basés sur les règles logiques d'une législation comme par exemple ceux présidant aux calculs des montants d'allocations.

Les articles 5 et 6 adoptent également le principe de définitions pragmatiques en listant explicitement les applications prohibées (art. 5) et celles à haut risque de l'IA facilement adaptables en fonction des évolutions technologiques. L'article 6 fait la différence entre les systèmes faisant déjà l'objet d'une réglementation européenne (annexe II : systèmes de transports et de soins) qui nécessitent une certification *ex-ante* par un tiers, organisme de notification, contrairement aux autres (annexe III) impactant également des personnes physiques mais dont le processus de mise en conformité est

2. Par opposition à *ex-post*, *ex-ante* signifie ici que l'analyse ou audit de conformité d'un algorithme d'IA afin de valider sa certification (marquage "CE") est considérée ou effectivement réalisée *avant* sa diffusion ou commercialisation et donc avant sa mise en exploitation.
3. Les considérants sont une liste de principes qui motivent un décret, une loi ou un règlement et qui en précèdent le texte contenu dans la liste des articles.

seulement déclaratif. *Attention* : la consultation attentive de ces annexes, de leur évolution, est importante pour bien distinguer les systèmes à haut risque des autres. Les scores de crédit bancaire sont concernés (cf. exemple numérique section 4) ainsi que les évaluations *individuelles* de "police prédictive" ou les scores de récidive (justice) mais pas explicitement celles concernant des évaluations de risques de délits par bloc géographique telles *Predpol* ou Paved en France. Pour les applications dans le domaine de la justice, seuls sont concernés les systèmes d'IA à l'usage des autorités judiciaires (magistrats) tels le projet abandonné DataJust mais pas ceux à l'usage des cabinets d'avocats (*e.g. case law analytics*).

L'article 10 est fondamental, il insiste sur l'importance d'une exploration statistique préalable exhaustive des données avant de lancer les procédures largement automatiques d'apprentissage et optimisation. Il évite une forme d'hypocrisie en autorisant, sous réserve de précautions avancées pour la confidentialité, la constitution de bases de données personnelles sensibles permettant par exemple des statistiques ethniques. Cela autorise la mesure directe des biais statistiques, sources potentielles de discrimination.

L'article 11 impose la rédaction d'une documentation qui est essentielle pour ouvrir la possibilité d'audit *ex-ante* d'un système d'IA à haut risque relevant de l'annexe II ou celui d'un contrôle *ex-post* pour ceux relevant de l'annexe III. Avec un reversement de la charge de preuve, c'est au concepteur de montrer qu'il a mis en œuve ce qu'il était techniquement possible en matière de sécurité, qualité, explicabilité, non discrimination, pour atteindre les objectifs attendus de conformité.

L'article 12 impose un archivage ou journalisation du fonctionnement d'un système d'IA à haut risque. Cette obligation est nouvelle par rapport aux textes européens précédents. Elle est indispensable pour assurer le suivi des mesures de performances, de risques et donc pour être capable de détecter des failles nécessitant des mises à jour voire un ré-entraînement du système ou même son arrêt. Les conditions d'archivage sont précisées dans l'article 61 (*post-market monitoring*).

Selon l'article 13 un utilisateur devrait pouvoir interpréter les sorties, et doit être clairement informé des performances, éventuellement en fonction des groupes concernés, ainsi que des risques notamment de biais et donc de discrimination. Il s'agit ici d'un point sensible directement dépendant de la complexité des systèmes d'IA à base d'algorithmes sophistiqués donc opaques d'apprentissage statistique. Le choix des métriques de biais sont laissées à l'initiative du concepteur. De plus, le manque de recul sur les recherches en cours en matière d'explicabilité d'une décision algorithmique laissent beaucoup de latitude à l'interprétation de cet article qui devra être adaptée à l'évolution des recherches très actives sur ce thème. L'article 14 complète ces dispositions en imposant une surveillance humaine visant à prévenir ou minimiser les risques pour la santé, la sécurité ou les droits fondamentaux.

L'article 15 comble une lacune importante par l'obligation de déclaration des performances (précisions, robustesse, résilience) d'un système d'IA à haut risque. Il concerne également les algorithmes d'apprentissage par renforcement soumis à des risques spécifiques : dérives potentielles (biais) et attaques malveillantes (cybersécurité) comme ce fut le cas pour le *chatbot Tay* de *Microsoft*.

Les articles des chapitres suivants du Titre III notifient des obligations sans apporter de précisions techniques ou méthodologiques : obligations faites au fournisseur (art. 16), obligation de mise en place d'un système de gestion de la qualité (art. 17), notamment de toute la procédure de gestion des données de la collecte initiale à leurs mises à jour en exploitation, ainsi que de la maintenance post-commercialisation ; obligation de documentation technique (art. 18), d'évaluation de la conformité (art. 19), obligation des utilisateurs (art. 29)...

Les États membres sont, par ailleurs, invités à désigner une autorité notifiante comme responsable du suivi des procédures relatives aux systèmes à haut risque et un organisme notifié (art. 30 à 39) indépendant, tout à fait classique des mécanismes de certification déjà en œuvre. Un marquage "CE" sera délivré aux systèmes conformes (art. 49).

Ce processus de marquage "CE" est essentiel pour les systèmes d'IA à haut risque de l'annexe II, il repose sur un audit *ex-ante* requérant, dans le cas d'une évaluation externe, des compétences très élaborées de la part de l'organisme qui en porte la responsabilité afin d'être à même de pouvoir déceler des manquements intentionnels ou non. Sans évaluation externe, pour les systèmes d'IA de l'annexe II, c'est à l'utilisateur de prendre ses responsabilités vis-à-vis du respect, entre autres, des droits fondamentaux afin de pouvoir faire face à un contrôle si l'État membre désigne une autorité compétente à ce sujet et lui en fournit les moyens.

2.3. Conséquences

L'analyse de ces quelques articles amène des commentaires ou questions, notamment sous le prisme d'une approche mathématique ou statistique de conception d'un système d'IA.

Projet Le projet de règlement (`AI Act`) entre dans un long processus (3 ou 4 ans comme le RGPD ?) de maturation avant une adoption européenne et une application par les États membres. Les amendements à venir devront être successivement pris en considération pour en analyser les conséquences en espérant que des réponses, précisions, corrections, seront apportées aux points ci-dessous. Néanmoins et compte tenu des temps et coûts de conception d'un système d'IA, il est important d'anticiper dès maintenant l'adoption de ce cadre réglementaire.

Exigences essentielles À la suite du guide des experts, le livre blanc appelle à satisfaire sept *exigences essentielles* dont celles de non discrimination et équité, bien être sociétal et environnemental.

Environnement la prise en compte de l'impact environnemental reste anecdotique, simplement évoquées dans les considérants (28) et (81), puis l'article 69 (*codes de conduite*) 2. sans aucune obligation formelle de calculer une balance bénéfices / risques (environnementaux ou autres) d'un système d'IA. Ainsi, l'obligation de l'archivage des données de fonctionnement d'un système d'IA génère un coût environnemental qui mériterait d'être pris en compte dans les risques afférents à son déploiement au regard de son utilité.

Équité La demande exprimée qu'un système d'IA satisfasse au respect des droits fondamentaux en référence à la charte de l'UE, notamment celui de non-discrimination, est très présente dans le livre blanc (cité 16 fois), comme dans les considérants (15, 17, 28, 39) de la proposition de règlement. En revanche, ce principe n'apparaît plus explicitement dans les articles. Est-ce sa présence dans des textes de plus haut niveau comme la Charte des Droits Fondamentaux de l'UE qui n'a pas justifié ici une répétition ou encore un manque d'harmonisation entre les États membres à ce propos ? Il n'y a donc pas de précision sur les façon de "mesurer" une discrimination ou la nécessité de l'atténuer. En revanche, les recherches et documentations des biais potentiels sont clairement explicitées.

Normes Le considérant (13) appelle à la définition de normes internationales notamment à propos des droits fondamentaux. En l'absence d'une définition juridique de l'équité d'un algorithme, celle-ci est définie en creux par l'*absence de discrimination* interdite explicitement. Le souci est que la littérature regorge de dizaines de définitions de biais statistiques pouvant être à l'origine de sources de discrimination ; lesquels considérer en priorité ? Il est peu probable que les autorités compétentes se prononcent à ce sujet, elles se focalisent (LNE 2021) sur les mesures de performances des systèmes d'IA de l'annexe II, notamment les systèmes de transport et les dispositifs de santé en vue de leur certification (marquage "CE").

Néanmoins, la recherche d'un biais systémique ou de société est requise dans l'analyse préalable des données (art. 10, 2. (f)), ainsi que l'obligation de détailler les performances (précision) par groupe ou sous groupe d'un système d'IA (art. 13, 3., (b) iv). Ceci permet de prendre en compte certains type de biais, donc de discriminations spécifiques même en l'absence de dé-

finitions normatives. Des indicateurs statistiques de biais devenus relativement consensuels dans la communauté académique sont proposés dans la section suivante.

En revanche, il est regrettable qu'aucune indication, recommandation, contrainte, ne vienne ensuite préciser ce qui pourrait ou devrait être fait pour atténuer ou éliminer un biais discriminatoire. Ceci est laissé au libre arbitre du concepteur d'un système d'IA en espérant que les choix opérés soient explicitement détaillés en toute transparence pour le fournisseur qui en assume la responsabilité et pour l'utilisateur en relation avec les usagers. L'exemple numérique illustre une telle démarche.

Utilisateur & Usager Le règlement traite en priorité les considérations commerciales, donc des risques de défaillance inhérents de l'acquisition des données à la mise en exploitation d'un système d'IA. Tout système doit satisfaire aux exigences de performance annoncées selon un principe de sécurité des produits ou responsabilité du fait des produits défectueux. En revanche, l'usager final, les dommages auxquels il peut être confronté, ne sont pas du tout pris en compte. L'obligation d'information (art. 13) est ainsi au profit de l'utilisateur et pas à celui de l'usager, personne physique impactée, qui ne semble donc protégé à ce jour que par les seules obligations de l'article 22 du RGPD. Il est informé de l'usage d'un système d'IA le concernant, il peut en contester la décision auprès de l'utilisateur humain mais l'explication de la décision, des risques encourus, sont soumises aux compétences et à la déontologie professionnelle de cet utilisateur : conseiller financier pour un client, magistrat pour un justiciable, responsable des ressources humaines pour un candidat, à moins d'un cadre juridique spécifique (*e.g.* code de santé public).

Données le règlement reconnaît le rôle prépondérant des algorithmes d'apprentissage automatique et donc de la nécessité absolue (considérant 44) de qualité et pertinence des données conduisant à leur entraînement. L'article 10 impose en conséquence des compétences en Statistique pour conduire les études préalables à l'entraînement d'un algorithme. Nous assistons à un renversement de tendance, un retour de balancier, du tout automatique à une approche raisonnée sous responsabilité humaine de cette phase d'analyse des données longue et coûteuse mais classique du métier de statisticien.

Responsabilité De façon générale, l'objectif essentiel n'est plus la performance absolue comme dans les concours de type *Kaggle* et conduisant à des empilements inextricables d'algorithmes opaques mais de satisfaire à un ensemble de contraintes pour la mise en conformité, dont celle de transparence, sous la responsabilité du fournisseur du système d'IA. L'analyse des responsabilités en cas de défaillance ou de produit défectueux sera l'objet d'un autre texte.

Documentation Tous les choix opérés lors de la conception d'un système d'IA : ensembles de données, algorithmes, procédures d'apprentissage et de tests, optimisations des paramètres, compromis entre confidentialité, performances, interprétabilité, biais... doivent (art. 11 et annexe IV) être explicitement documentés en vue d'un audit *ex-ante* des systèmes de l'annexe II ou d'un contrôle *ex-post* d'un système de l'annexe III. C'est un renversement de la charge de preuve sous la responsabilité du fournisseur qui doit pouvoir montrer que le concepteur a mis en œuvre ce qui était techniquement possible pour satisfaire aux obligations (conformité) légales de sécurité, transparence, performances et non discrimination.

Autorité notifiante (Chapitre 4 Titre III) Chaque pays va se doter ou désigner (art. 30) un service chargé entre autres de superviser l'audit *ex-ante* d'un système d'IA à haut risque de l'annexe II avant son déploiement qu'il soit commercialisé ou non. L'autorité notifiante désigne l'*organisme de notification* qui exécutera l'audit. De façon assez étonnante, un système d'ascenseur élémentaire, n'embarquant qu'une "IA" logique rudimentaire mais dépendant de l'annexe II, est plus contraint par l'obligation de certification par un organisme tiers, au contraire d'applications des systèmes d'IA de l'annexe III (justice, emploi, crédit...) impactant directement des personnes physiques avec des risques réels envers les droits fondamentaux. Il faudra donc être

attentif à l'interprétation que fera un État membre de cette situation afin d'évaluer les possibilités de saisine et compétences de contrôle d'un système à haut risque de l'annexe III.

Archivage & confidentialité Le règlement cible donc, en première lecture, les obligations commerciales du fournisseur plutôt que celles étiques ou déontologiques envers l'usager. Néanmoins le règlement apporte la possibilité de prendre en compte des données sensibles (art. 10, 5.), les obligations d'archivage des décisions (art. 12), de suivi des performances selon les groupes (art. 13), une surveillance humaine (art. 14) pendant toute la période d'utilisation et de correction rétro-active des biais (art. 15). Cette obligation d'archivage et surveillance du fonctionnement notamment à destination des groupes sensibles oblige implicitement à l'acquisition, en toute sécurité (cryptage, anonymisation, pseudonymisation...), de données confidentielles (*e.g.* origine ethnique). Cela ne rend-il pas indispensable, selon le domaine d'application, la mise en place d'un protocole explicite de consentement libre et éclairé, d'un engagement éthique, entre l'utilisateur et l'usager, protégé par le RGPD ? Comment sont évalués les risques encourus d'un usager ou groupe d'usager par le recueil et l'exploitation de leurs données sensibles lors de l'exploitation d'un système d'IA face aux bénéfices attendus pour eux mêmes ou l'intérêt public ?

3. Prise en compte méthodologique de l'*AI Act*

3.1. Quels algorithmes

Dans l'attente d'une adoption effective du texte final qui risque d'être amendé, il est néanmoins prudent, compte tenu des investissements en jeu, d'anticiper des réponses techniques à certaines contraintes ou obligations faites aux système d'IA désignés à haut risque. Cet article laisse volontairement de côté certaines classes d'algorithmes mentionnées ou non dans l'annexe I dont la liste finale reste l'objet de débats entre les instances européennes.

Un système expert est l'association d'une base de règles logiques ou base de connaissances construites par des experts du domaine concerné, d'un moteur d'inférence et d'une base de faits observés pour une exécution en cours. Le moteur d'inférence recherche la séquence de règles logiquement applicables à partir des faits observés de la base qui s'incrémente comme conséquences du déclenchements des règles. Le processus itère jusqu'à l'obtention ou non d'une décision recherchée et expliquée par la séquence de règles y conduisant. Très développée dans les années 70, la recherche a marqué le pas face à un problème dit *NP-complet* c'est-à-dire de complexité algorithmique exponentielle en la taille de la base de connaissance (nombre de règles). Supplantée par la ré-émergence des réseaux de neurones (années 80) puis plus largement par l'apprentissage automatique, la recherche dans ce domaine dit d'IA symbolique est restée active. Elle connaît un renouveau motivé par les capacités d'explicabilité des systèmes experts.

Les approches statistiques bayésiennes ou non basées sur des données sont associées implicitement aux méthodes par apprentissage. En revanche, les algorithmes d'allocation optimale de ressources prennent une place à part. Si les principes d'allocation en tant que tels ne soulèvent pas de problème, ceux d'ordonnancement ou de tri des ressources peuvent amener des risques réels de discrimination indirecte. C'est notamment le cas de l'algorithme ParcourSup lorsque les établissements d'enseignement supérieur introduisent des pondérations selon le lycée d'origine des candidats : lycée de centre ville *vs.* lycée de banlieue. Cette situation rejoint alors le cas des algorithmes déterministes ou procéduraux. Il s'agit d'algorithmes décisionnels (*e.g.* calcul de taxes, impôts, allocations ou prestations sociales,...) basés sur un ensemble de règles de décision déterministes qui peuvent tout autant présenter des impacts, désavantages ou risques de discrimination indirecte, malgré une apparente neutralité. La Défenseure des Droits (2020) est très attentive en France à l'analyse

et détection de ces risques. Celle-ci relèvent de l'analyse experte des règles de décisions codées dans l'algorithme qui en l'état ne sont pas concernés par le projet de règlement. Néanmoins, la complexité de l'algorithme peut être telle qu'une analyse experte *ex-post* ne sera pas en mesure d'évaluer l'étendue des risques indirects. Aussi, un algorithme déterministe complexe peut être analysé avec les mêmes outils statistiques que ceux adaptés à un algorithme d'apprentissage automatique.

Nous insistons donc tout particulièrement sur les systèmes d'IA basés sur des algorithmes d'apprentissage supervisé ou statistique ou IA empirique par opposition à l'IA dite symbolique des systèmes experts. Ce sont très majoritairement les plus répandus au sein de ceux désignés à haut risque (art. 6) car susceptibles d'impacter directement des personnes physiques.

Même sans obligation de certification *ex-ante* par un organisme notifié, une documentation exhaustive (art. 11) d'un système d'IA à haut risque doit être produite et fournie à l'utilisateur. Cette section propose quelques indications méthodologiques pour répondre à cette attente.

3.2. Les données

Tout système d'IA basé sur un algorithme d'apprentissage statistique nécessite la mise en place d'une base de données d'entraînement fiable et représentative du domaine d'application visé qui doit en tout premier lieu satisfaire aux exigences de confidentialité du RGPD. Puis, le travail d'exploration statistique, généralement long et fastidieux d'acquisition, vérification, analyse, préparation, nettoyage, enrichissement, archivage sécurisé des données, est essentiel à l'élaboration d'un système d'IA performant, robuste, résilient et dont les biais potentiels sont sous contrôle. Construire de nouvelles caractéristiques (*features*) adaptées à l'objectif, traquer et gérer éventuellement par imputation des données manquantes, identifier les anomalies ou valeurs atypiques (*outliers*) sources de défaillances, les sources de biais : classes ou groupes sous représentés, biais systémiques, nécessitent compétences et expériences avancées en Statistique.

Ces compétences sont indispensables pour répondre aux attentes de l'article 10 ainsi qu'aux besoins de la documentation (annexe IV) imposée par l'article 11.

3.3. Qualité, précision et robustesse

Les articles 13 et 15 imposent clairement de devoir documenter les performances et risques d'erreur, éventuellement en fonction de groupes sensibles et protégés, ou de défaillance d'un système d'IA. Cela rend indispensable l'explicitation de choix notamment des métriques utilisées.

Choix de métrique

L'évaluation de la qualité d'une aide algorithmique à la décision est essentielle à la justification du déploiement d'un système d'IA au regard de sa balance bénéfice / risques. Dans le cas d'un système IA empirique ou par apprentissage automatique, il s'agit d'estimer la *précision* des prévisions dont les mesures sont bien connues et maîtrisées, parties intégrante du processus d'apprentissage. Néanmoins parmi un large éventail des possibles, le choix, précisément justifié, doit être adapté au domaine, au type de problème traité, aux risques spécifiques encourus quelque soit le modèle ou le type d'algorithme d'apprentissage utilisé. Citons par exemple les situations de :

Régression ou modélisation et prévision d'une variable cible Y quantitative. Elle est généralement basée sur l'optimisation d'une mesure quadratique (norme L_2) pouvant intégrer, à l'étape d'entraînement, différents types de pénalisation dont celles de parcimonie (*ridge, Lasso*) afin de contrôler la complexité de l'algorithme et éviter les phénomènes de sur-apprentissage. D'autres

types de fonction objectif basée sur une perte en norme L_1 ou valeur absolue, moins sensible à la présence de valeurs atypiques (*outliers*) que la norme quadratique, permet des solutions plus robustes car tolérantes à des observations atypiques.

Classification ou modélisation, prévision d'une variable Y qualitative. Le choix d'une mesure d'erreur doit être opéré parmi de très nombreuses possibilités : taux d'erreur, AUC (*area under the ROC Curve* pour une variable Y binaire), score F_β, risque bayésien, entropie... avec la difficile prise en compte des situations de classes déséquilibrées qui oriente le choix du type de mesure et nécessite des précautions spécifiques dans l'équilibrage de la base d'apprentissage ou les pondérations de la fonction objectif en prenant en compte une matrice de coûts de mauvais classement éventuellement asymétrique.

Limites de la précision

Besse (2021) rappelle que les performances de l'IA sont largement surévaluées par le battage médiatique dont bénéficient ces technologies. Ces performances sont d'autant plus dégradées lorsque la décision concerne la prévision d'un comportement (achat, départ, embauche, acte violent, pathologie...) individuel humain dépendant potentiellement d'un très grand nombre de variables explicatives ou facteurs dont certains peuvent ne pas être observables. Il importe de distinguer les systèmes d'IA développés dans un domaine bien déterminé (*e.g.* process industriel sous-contrôle), où le nombre de facteurs ou dimensions est raisonnable et identifié, des systèmes d'IA où opère le fléau ou malédiction de la dimension (*curse of dimensionality*), lorsque celle-ci est très grande, voire indéterminée.

L'histoire de la littérature statistique puis d'apprentissage automatique peut être lue comme une succession de stratégies pour le contrôle du nombre de variables et ainsi de paramètres estimés dans un modèle statistique ou entraînés dans un algorithme. Il s'agit par exemple de contrôler le conditionnement d'une matrice en régression : PLS (*partial least square*), sélection de variables (critères AIC, BIC), pénalisations *ridge* ou *Lasso*, et ainsi l'explosion de la variance des prévisions. Plus généralement c'est aussi le contrôle du risque de sur-ajustement qui doit être documenté comme résultat de l'optimisation des hyperparamètres : nombre de plus proches voisins, pénalité en machines à vecteurs supports, nombre de feuilles d'un arbre, de variables tirées aléatoirement dans une forêt d'arbres, profondeur des arbres et nombre d'itérations en *boosting* ... structures des couches convolutionnelles et *drop out* des réseaux de neurones en reconnaissance d'images. Même si les stratégies d'optimisation de ces hyperparamètres par validation croisée ou échantillon de validation sont bien rodées, le fléau de la dimension peut s'avérer rédhibitoire (*e.g.* Verzelen 2012).

Échantillon test

En tout état de cause, il est indispensable de mettre en place une démarche très rigoureuse pour conduire à l'évaluation de la précision et donc des performances d'un système d'IA basé sur un algorithme d'apprentissage. Comme énoncé dans l'article 3, 31. ce sont des *données de test indépendantes* de celles d'apprentissage qui sont utilisées à cet effet. *Attention* néanmoins d'évaluer les performances sur des données telles qu'elles se présenteront *réellement* en exploitation, avec leurs défauts, et pas un simple sous-ensemble aléatoire de la base d'apprentissage comme c'est trop souvent le cas en recherche académique. En effet cet ensemble de données peut bénéficier d'une homogénéité d'acquisition (*e.g.* même technologie, même opérateur) et de prétraitements qui peuvent faire défaut à de réelles données d'entrée à venir en exploitation. Cela demande donc une extrême rigueur dans la constitution d'un échantillon test pour éviter ces pièges bien trop présents en recherche académique (*e.g.* Liu et al. 2019, Roberts et al. 2021) et conduisant, sous la pression de publication, à beaucoup trop de résultats non reproductibles et des algorithmes non certifiables. Enfin, une surveillance (art. 14) toute la durée de vie du système d'IA est indispensable afin d'en

détecter de possibles dérives ou dysfonctionnements (art. 12 et 15) affectant la robustesse ou la résilience des décisions.

Robustesse

L'évaluation de la *robustesse* est liée aux procédures de contrôle mises en place pour détecter des valeurs atypiques (*outliers*) ou anomalies dans la base d'apprentissage et au choix de la fonction perte de la procédure d'entraînement de l'algorithme. Impérativement, surtout dans les d'applications sensibles pouvant entraîner des risques élevés en cas d'erreur, la détection d'anomalie doit également être intégrée en exploitation afin de ne pas chercher à proposer des décisions correspondant à des situations atypiques, étrangères à la base d'apprentissage.

Résilience

La *résilience* d'un système d'IA est essentielle pour les dispositifs critiques (dispositifs de santé connecté, aide au pilotage). Cela concerne par exemple la prise en compte de données manquantes lors de l'apprentissage comme en exploitation. Il s'agit d'évaluer la capacité d'un système d'IA à assurer des fonctions pouvant s'avérer vitales en cas, par exemple, de panne ou de fonctionnement erratique d'un capteur : choix d'un algorithme tolérant aux données manquantes, imputation de celles-ci, fonctionnement en mode dégradé, alerte et arrêt du système.

3.4. Explicabilité

Une recherche active

Il est bien trop tôt pour tenter un résumé opérationnel de ce thème et fournir des indications claires sur la démarche à adopter pour satisfaire aux exigences réglementaires (art. 13, 15). Il faut pour cela attendre que la recherche ait progressé et qu'une sélection "naturelle" en extrait les procédures les plus pertinentes parmi une grande quantité de solutions proposées ; un article de revue sur ce sujet (Barredo Arrieta et al. 2020) listait plus de 400 références.

Arbre de choix

Tentons de décrire les premiers embranchements d'un arbre de décision en répondant à quelques questions rudimentaires qu'il faudrait en plus adapter au domaine d'application car le type de réponse à apporter n'est évidemment pas le même s'il s'agit d'expliquer le refus d'un prêt ou les conséquences d'une aide automatisée au diagnostic d'un cancer.

Il importe de bien distinguer les niveaux d'explication : concepteur, utilisateur ou usager, même si ce dernier n'est pas directement concerné par le projet de règlement. De plus, l'explication peut s'appliquer soit au fonctionnement général de l'algorithme soit à une décision spécifique.

Il y a schématiquement deux types d'algorithmes dontceux relativement transparents : modèles linéaires et arbres de décision. L'explication est dans ce cas possible à condition que le nombre de variables et d'interactions prises en compte ou le nombre de feuilles d'un arbre reste raisonnable. Toutes les autres classes d'algorithme d'apprentissage, systématiquement non linéaires et complexes, sont par construction opaques. Il s'agit alors de construire une explication par différentes stratégies comme une approximation explicable par un modèle linéaire, un arbre ou un ensemble de

règles de décision déterministes. Une autre stratégie consiste à fournir des indications sur l' *importances des variables* en mesurant l'effet d'une permutation aléatoire de leurs valeurs (*mean decrease accuracy* Breiman, 2001), en stressant l'algorithme (Bachoc et al. 2020) ou en réalisant une analyse de sensibilité par indices de Sobol (Bénesse et al. 2021).

Le concepteur d'un algorithme s'intéresse également à l'explication d'une décision spécifique afin d'identifier la cause d'une erreur, y remédier par exemple en complétant la base d'apprentissage d'un groupe sous-représenté avant de ré-entraîner l'algorithme. L'utilisateur d'un système d'IA doit être au mieux informé (art. 13, 15) des possibilités d'expliquer une décision qu'il pourra retranscrire à l'usager (client, patient, justiciable, citoyen...) selon sa propre déontologie, son intérêt commercial ou une contrainte légale par exemple pour des décisions administratives. Pour ce faire quelques stratégies sont proposées comme une *approximation locale* par un modèle explicable (linéaire, arbre de décision) ou par une liste d'exemples *contrefactuels* c'est-à-dire des situations les plus proches, en un certain sens, qui conduiraient à décision contraire, généralement plus favorable (attribution d'un prêt). Lorsque cela s'avère impossible, comme par exemple dans le cas d'un diagnostic médical impliquant un nombre important de facteurs opaques, il importe d'informer précisément l'utilisateur et donc le patient sur les risques d'erreur afin que consentement de ce dernier soit effectivement libre et éclairé.

Quelques démonstrations de procédures explicatives sont proposées sur des sites en accès libre. Citons : gems-ai.com, aix360.mybluemix.net, github.com/MAIF/shapash

Réalité complexe

Ne pas perdre de vue que l'impossibilité ou simplement la difficulté à formuler une explication provient certes de l'utilisation d'algorithmes opaques mais dont la nécessité est inhérente à la complexité même du réel. Un réel complexe (*e.g.* les fonctions du vivant) impliquant de nombreuses variables, leurs interactions, des effets non linéaires voire des boucles de contre-réaction, est nécessairement modélisé par un algorithme complexe afin d'éviter des simplifications abusives pouvant gravement nuire aux performances. C'est tout d'abord le réel qui s'avère complexe à expliquer.

3.5. Biais & discrimination

Bien que très présente dans les textes préliminaires (livre blanc (CE 2021, considérants de l'AI Act) la référence au risque de discrimination ne l'est pas de façon explicite dans les projets d'articles. Apparaissent néanmoins l'obligation de détecter des biais dans les données (art. 10) ainsi que celle d'afficher des performances ou risques d'erreur par groupe (art. 13). Quelles en sont les conséquences au regard des difficultés de définir, détecter une discrimination qu'elle soit humaine ou algorithmique ?

Détecter une discrimination

Formellement, la stricte équité peut s'exprimer par des propriétés d'indépendance en probabilité entre la variable cible Y qui exprime une décision et la variable dite sensible S par rapport à laquelle une discrimination est en principe interdite. Cette variable peut être quantitative (*e.g.* âge) ou qualitative à deux ou plusieurs classes (*e.g.* genre ou origine ethnique) ou, de façon plus complexe, la prise en compte d'interactions entre plusieurs variables sensibles. Néanmoins cette définition théorique de l'équité n'est pas concrètement praticable pour détecter, mesurer, atténuer des risques de biais. De plus, les textes juridiques font essentiellement référence à un groupe de personnes sensibles par

rapport aux autres. En conséquences et pour simplifier cette première lecture pédagogique de la détection des risques de discrimination, nous ne considérons qu'une variable sensible à 2 modalités : jeune *vs.* vieux, femme *vs.* homme...

Une façon bien établie de détecter une décision humaine discriminatoire consiste à opérer par *testing*. Dans le cas d'une présomption de discrimination à l'embauche, la procédure consiste à adresser deux CV comparables, à l'exception (*counterfactual example*) de la modalité de la variable sensible (*e.g.* genre, origine ethnique associée au nom) afin de comparer les réponses : proposition ou non d'entretien. Cette démarche individuelle est rendue systématique (Rich, 2014) dans une enquête par l'envoi de milliers de paires de CV. C'est en France la doctrine officielle promue par le Comité National de l'Information Statistique et commanditée périodiquement par la DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques) du Ministère du travail.

Des indicateurs statistiques peuvent être estimés à l'issue de cette enquête mais, comme il n'existe pas de définition juridique de l'équité qui devient par défaut l'absence de discrimination, le monde académique a proposé quelques dizaines d'indicateurs (*e.g.* Zliobaitė 2017) afin d'évaluer des biais potentiels sources de discrimination. Il est nécessaire d'opérer des choix parmi tous les critères de biais en remarquant que beaucoup de ces indicateurs s'avèrent être très corrélés ou redondants (Friedler et al. 2019). Empiriquement et après avoir consulter une vaste littérature sur l'IA éthique ou plutôt sur les risques identifiés de discrimination algorithmique, un consensus émerge sur le choix en priorité de trois niveaux de biais statistique. Sont finalement considérés dans cet article élémentaire trois types de rapports de probabilités (égaux à 1 en cas d'indépendance stricte) dont Besse et al. (2021) proposent des estimations par intervalle de confiance afin d'en contrôler la précision.

Parité statistique et effet disproportionné

Le premier niveau de risque de discrimination algorithmique s'illustre simplement : si un algorithme est entraîné sur des données biaisées, il apprend et reproduit très fidèlement ces biais systémiques, de société ou de population, par lesquels un groupe est historiquement (*e.g.* revenu des femmes) désavantagé ; plus grave, l'algorithme risque même de renforcer le biais en conduisant à des décisions explicitement discriminatoires. Il importe donc de pouvoir détecter, mesurer, atténuer voire éliminer ce type de biais. L'équité ou parité statistique (ou *demographic equality*) serait l'indépendance entre la ou les variables sensibles S (*e.g.* genre, origine ethnique) et la variable de prévision $\widehat{Y}$ de la décision. Historiquement, l'écart à l'indépendance pour mesurer ce type de biais est évalué aux USA dans les procédures d'embauche depuis 1971 par la notion d'effet disproportionné ou *disparate impact* et maintenant reprises systématiquement (Barocas et Selbst, 2016) pour l'évaluation de ce type de discrimination dans un algorithme. L'effet disproportionné consiste à estimer le rapport de deux probabilités : probabilité d'une décision favorable ($\widehat{Y} = 1$) pour une personne du groupe sensible ($S = 0$) au sens de la loi sur la même probabilité pour une personne de l'autre groupe ($S = 1$) :

$$DI = \frac{\mathbb{P}(\widehat{Y} = 1|S = 0)}{\mathbb{P}(\widehat{Y} = 1|S = 1)}.$$

Cet indicateur est intégré au *Civil Rights act & Code of Federal Regulations (Title 29, Labor: Part 1607 Uniform guidelines on employee selection procedures)* depuis 1978 avec la règle dite des 4/5 ème ; si DI est inférieur à $0,8$, l'entreprise doit en apporter les justifications économiques. Les logiciels commercialisés aux USA et proposant des algorithmes de pré-recrutement automatique anticipent ce risque juridique (Raghavan et al. 2019) en intégrant une procédure automatique d'atténuation du biais (*fair learning*). Il n'y a aucune obligation ni mention en France de cet indicateur statistique, seulement une incitation de la part de la Défenseure des Droits et de la CNIL (2012) envers les services de ressources humaines des entreprises. Il leur est suggéré de tenir des statistiques ethniques,

autorisées dans ce cas sous réserve de confidentialité, sous la forme de tables de contingence dont il serait facile d'en déduire des estimations d'effet disproportionné.

La mise en évidence d'un biais systémique est implicitement citée lors de l'étape d'analyse préliminaire des données (art. 10, 2., (f)) mais sans plus de précision sur la façon dont il doit être pris en compte alors que renforcer algorithmiquement ce biais serait ouvertement discriminatoire. De plus serait-il politiquement opportun d'introduire une part de discrimination positive afin d'atténuer la discrimination sociale ? C'est évoqué dans le guide des experts (CE, 2019, ligne directrice 52) pour *améliorer le caractère équitable de la société* et techniquement l'objet d'une vaste littérature académique nommée apprentissage équitable (*fair learning*). Cette opportunité n'est pas reprise explicitement dans l'*AI Act* mais nous verrons dans l'exemple numérique ci-dessous qu'elle ne peut être exclue et peut même être pleinement justifiée en prenant en considération les autres types de biais ci-après.

Erreurs conditionnelles

Les taux d'erreur de prévision et donc les risques d'erreur de décisions sont-ils les mêmes pour chaque groupe (*overall error equality*) ? Autrement dit, l'erreur est-elle indépendante de la variable sensible ? Ceci peut se mesurer par l'estimation (intervalle de confiance) du rapport de probabilités (probabilité de se tromper pour le groupe sensible sur la probabilité de se tromper pour l'autre groupe) :

$$REC = \frac{\mathbb{P}(\widehat{Y} \neq Y | S = 0)}{\mathbb{P}(\widehat{Y} \neq Y | S = 1)}.$$

Ainsi, si un groupe est sous-représenté dans la base d'apprentissage, il est très probable que les décisions le concernant soient moins fiables. C'est une des première critiques formulées à l'encontre des algorithmes de reconnaissance faciale et ce risque est également présent dans les applications en santé (Besse et al. 2020) ou en ressources humaines (De-Arteaga et al. 2019). L'identification, la prise en compte et la surveillance de ce risque sont présents (art. 13, 3., (b), ii et art. 15, 1. & 2.) dans le projet de règlement et doivent donc être explicitement détaillés dans la documentation (art. 11).

Rapports de cote conditionnels

Même si les deux critères précédents sont trouvés équitables, les erreurs peuvent être dissymétriques (plus de faux positifs, moins de faux négatifs) au détriment d'un groupe avec un impact d'autant plus discriminatoire que le taux d'erreur est important. Cet indicateur (comparaison des rapports de cote ou *odds ratio* d'indépendance conditionnelle nommé aussi *equalli odds*) est au cœur de la controverse concernant l'évaluation COMPAS du risque de récidive aux USA (Larson et al. 2016). Il est également présent dans l'exemple numérique ci-après. Cet indicateur double est mesuré par l'estimation de deux rapports de probabilités : rapports des taux de faux positifs du groupe sensible sur le taux de faux positifs de l'autre groupe et rapport des taux de faux négatifs pour ces mêmes groupes.

$$RFP = \frac{\mathbb{P}(\hat{Y} = 1 | Y = 0, S = 0)}{\mathbb{P}(\hat{Y} = 1 | Y = 0, S = 1)} \quad \text{et} \quad RFN = \frac{\mathbb{P}(\widehat{Y} = 0 | Y = 1, S = 0)}{\mathbb{P}(\widehat{Y} = 0 | Y = 1, S = 1)}.$$

L'évaluation de ce type de biais n'est pas explicitement mentionné dans le projet de règlement. Néanmoins il fait partie de la procédure classique d'évaluation des erreurs en classification à l'aide d'une matrice de confusion ou de courbes ROC par groupes et ne peut être négligé.

Notons qu'il est d'autant plus difficile de faire abstraction du dernier type de biais que les trois sont interdépendants et même en interaction avec les autres risques : précision et explicabilité. Ceci est

clairement mis en évidence dans l'exemple numérique suivant. Il y a donc une forme d'obligation déontologique ou de cohérence statistique à devoir appréhender ces différents niveaux d'analyse.

4. Exemple numérique

L'exemple jouet ou bac à sable de cette section permet d'illustrer concrètement toute la complexité des principes précédemment évoqués en soulignant leur interdépendance. Ce jeu de données est ancien, largement utilisé pour illustrer tous les travaux visant une atténuation optimale du biais. Le monde académique espère avoir rapidement accès à bien d'autres "bac à sable" représentatifs dont la construction est l'objet de l'article 53 de l'*AI Act*.

4.1. Données

Les données publiques utilisées imitent le contexte du calcul d'un score de crédit. Elles sont extraites (échantillon de 45 000 personnes) d'un recensement de 1994 aux USA et décrivent l'âge, le type d'emploi, le niveau d'éducation, le statut marital, l'origine ethnique, le nombre d'heures travaillées par semaine, la présence ou non d'un enfant, les revenus ou pertes financières, le genre et le niveau de revenu bas ou élevé. Elles servent de référence ou *bac à sable* pour tous les développements d'algorithmes d'apprentissage automatique équitable. Il s'agit de prévoir si le revenu annuel d'une personne est supérieur ou inférieur à 50k$ et donc de prévoir, d'une certaine façon, sa solvabilité connaissant ses autres caractéristiques socio-économiques. Ces questions de discrimination dans l'accès au crédit sont toujours d'actualité (Campisi 2021, Hurlin et al. 2021, Kozodoi et al. 2021) même si le principe du *score de crédit* s'est généralisé dès les années 90 avec l'envol du *data mining* devenu depuis de l'IA.

L'étude complète et les codes de calcul sont disponibles dans un tutoriel (calepin *Jupyter*) mais l'illustration est limitée à un résumé succinct de l'analyse de la discrimination selon le genre.

4.2. Résultats

Une analyse exploratoire : nettoyage des données, description statistique, préalable doit être incluse dans la documentation. Elle est l'objet d'un autre tutoriel dont seuls quelques résultats sont retenus par souci de concision. Ils mettent en évidence un biais systémique ou de société important : seulement $11,6\%$ des femmes ont un revenu élevé contre $31,5\%$ des hommes. Le rapport $DI = 0,38$ est donc très disproportionné et peut s'expliquer par quelques considérations sociologiques bien identifiées sur le premier plan factoriel (fig. 1) d'une analyse factorielle multiple des correspondances calculée après avoir recodé qualitatives toutes les variables. Les femmes travaillent en moyenne moins d'heures (`HW1`) par semaine (occupations ménagères et enfants ?) ; même si le niveau de diplôme ne semble pas lié au genre, elles occupent un poste avec moins de responsabilité (`Admin`) (effet plafond de verre ?). Un autre type de biais semble présent dans ces données, les femmes sont associées (co-occurrences plus fréquentes que l'indépendance) à la présence d'enfants sans pour autant être en situation de couple contrairement aux hommes. Cette enquête s'adresse-t-elle de façon privilégiée au chef ou à la cheffe de famille éventuellement monoparentale ?

Les données ont été aléatoirement réparties en trois échantillons d'apprentissage (29 000), destinés à l'estimation des modèles ou entraînement des algorithmes, de validation (8000) afin d'optimiser certains hyper paramètres et de test (8000) pour évaluer les différents indicateurs de performance et biais. La taille relativement importante de l'échantillon initial permet de considérer un échantillon

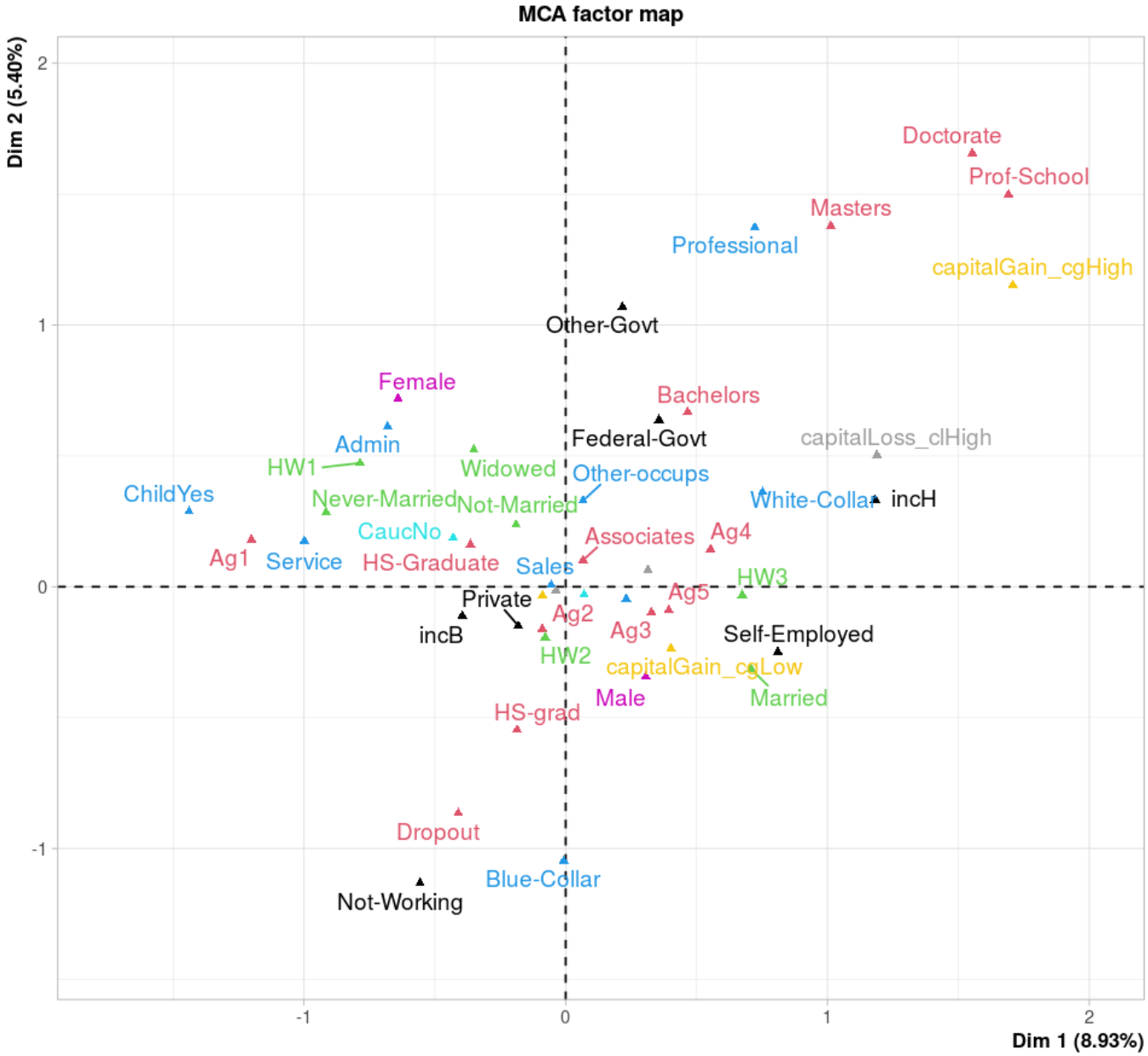

FIGURE 1 – Premier plan factoriel d'une analyse factorielle multiple des correspondances (librairie `FactoMineR`, Lê et al. 2008)

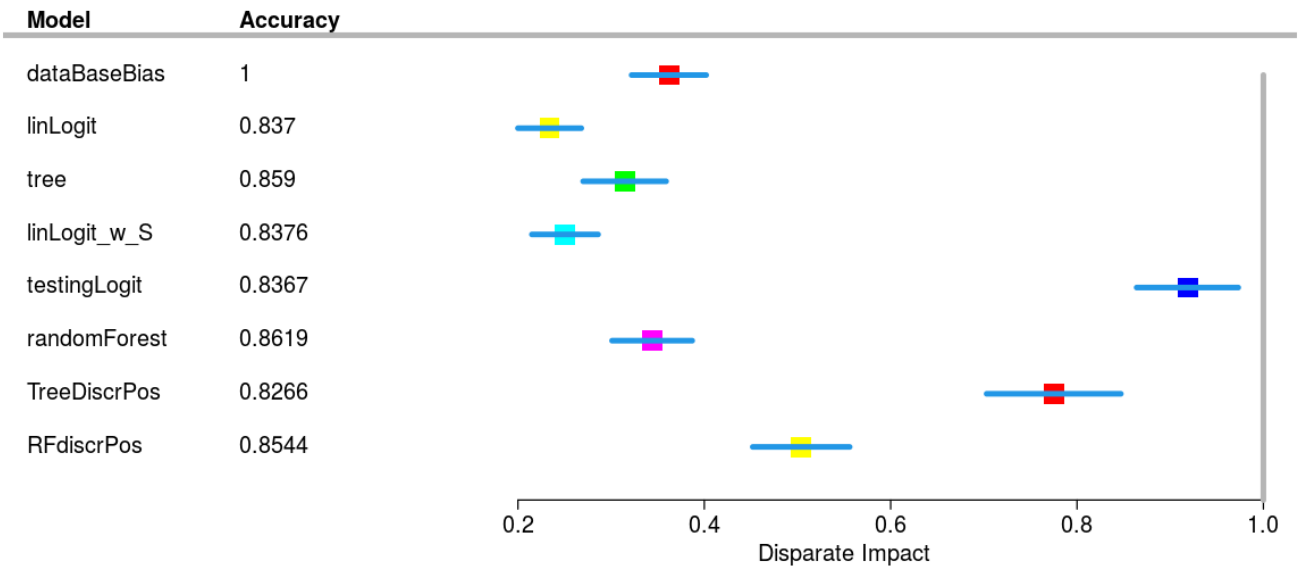

FIGURE 2 – Précision de la prévision (accuracy) et effet disproportionné (discrimination en fonction du genre) estimé par un intervalle de confiance sur un échantillon test (taille 9000) pour différents modèles ou algorithmes d'apprentissage.

de validation représentatif, comme demandé dans le règlement, afin d'éviter des procédures plus lourdes de validation croisée. Les résultats de prévision sont regroupés dans la figure 2.

Le biais systémique (`dataBaseBias`) des données est comparé avec celui de la prévision de niveau de revenu par un modèle classique linéaire de régression logistique `linLogit` : $DI = 0,25$. Significativement moins élevé (intervalles de confiance disjoints), il montre que ce modèle renforce le biais et donc discrimine nettement les femmes dans sa prévision. La procédure naïve (`linLogit-w-s`) qui consiste à éliminer la variable dite sensible (genre) du modèle ne supprime en rien ($DI = 0,27$) le biais discriminatoire car le genre est de toute façon présent à travers les valeurs prises par les autres variables (effet *proxy*). Une autre conséquence de cette dépendance aux proxys est que le *testing* ou *counterfactual test* (changement de genre toutes choses égales par ailleurs) ne détecte plus ($DI = 0,90$) aucune discrimination !

Un algorithme non-linéaire élémentaire (`tree`, arbre binaire de décision) augmente le biais mais pas de façon statistiquement significative car les intervalles de confiance ne sont pas disjoints. Sa précision est meilleure que celle du modèle de régression logistique mais, si l'objectif est une interprétation utile, il est nécessaire de réduire la complexité de l'arbre en pénalisant le nombre de feuilles, d'une centaine à une dizaine. Dans ce cas la précision se dégrade pour rejoindre celle de la régression logistique ; l'explicabilité a un coût.

Un algorithme non linéaire plus sophistiqué (`random forest`) est très fidèle au biais des données avec un indicateur ($DI = 0,36$) proche de celui du biais de société et fournit une meilleure précision : $0,86$ au lieu de $0,84$ pour la régression logistique. Cet algorithme ne discrimine pas plus, apporte une meilleure précision, mais c'est au prix de l'explicabilité du modèle. Opaque comme un réseau de neurones, il ne permet pas d'expliquer une décision à partir de ses paramètres comme cela est facile avec le modèle de régression ou un arbre binaire de décision de taille raisonnable.

Une question délicate concerne le choix politique de procéder ou non à une atténuation du biais systémique dans le cas d'un score de crédit. Contrairement à Hurlin et al. (2021), Goglin (2021)

l'aborde de façon très incomplète en ne considérant, de manière exclusive, que le biais des erreurs selon le genre. Cet auteur "justifie" de ne pas considérer le biais systémique car le corriger conduirait des femmes à des situations de surendettement tandis que le 3ème type de biais est purement oublié. Une analyse plus fine montre, à travers cet exemple, toute l'importance de prendre en compte simultanément les trois types de biais afin d'éviter un positionnement quelque peu "paternaliste".

En principe, la précision de la prévision pour un groupe dépend de sa représentativité. Si ce dernier est sous-représenté, l'erreur est plus importante ; c'est typiquement le cas en reconnaissance faciale mais pas dans l'exemple traité. Alors qu'elles sont deux fois moins nombreuses dans l'échantillon, le taux d'erreur de prévision est de l'ordre de $7,9\%$ pour les femmes et de 17% ($REC = 0,36$) pour les hommes (algorithme d'arbre binaire simplifié). Il est alors indispensable de considérer le troisième type de biais pour se rendre compte que c'est finalement au désavantage des femmes. Le taux de faux positifs est plus important pour les hommes $(0,081)$ que pour les femmes $(0,016)$ ($RFP = 0,20$). Ceci avantage les hommes qui bénéficient plus largement d'une décision favorable même à tort. En revanche, le taux de faux négatifs est plus important pour les femmes $(0,41)$, à leur désavantage, que pour les hommes $(0,38)$ ($Rfn = 1,08$) mais ces dernières différences ne sont pas significatives.

Dans une telle situation en choisissant le seuil de décision par défaut à $0,5$, une banque prendrait peu de risque : faible taux de faux positifs et taux élevés de faux négatifs mais, conclusion importante, il apparaît une *rupture d'équité* au sens où la banque prend *plus de risques au bénéfice des hommes* alors que les taux d'erreur les concernant sont plus élevés.

Une atténuation du biais des rapports de cotes se justifie donc afin de rendre comparables les chances d'obtention d'un crédit selon le genre et ce même à tort. Plutôt que d'équilibrer ces chances en pénalisant celles des hommes, une part de discrimination positive est introduite au bénéfice des femmes pour plus d'équité en cherchant à rendre égaux les taux de faux positifs selon le genre et évalués sur l'échantillon de validation.

Les deux dernières lignes de la figure 2 proposent une façon simple (*post-processing*), parmi une littérature très volumineuse, de corriger le biais pour plus de *justice sociale*. Deux algorithmes sont entraînés, un par genre et le seuil de décision (revenu élevé ou pas, accord ou non de crédit...) est abaissé pour les femmes : $0,3$ pour les forêts aléatoires, $0,2$ pour un arbre binaire, au lieu de celui par défaut de $0,5$ pour les hommes. Cette correction des faux positifs impacte également les taux d'erreur qui deviennent plus équilibrés selon le genre et provoque également une atténuation de l'effet disproportionné pour une *société plus équitable*. L'arbre binaire utilisé (`TreeDiscrPos`) est celui pénalisé (peu de feuilles) afin d'obtenir une interprétation facile au prix de la précision. Les seuils et le paramètre de pénalisation ont été déterminés sur l'échantillon de validation avant d'être appliquées indépendamment à l'échantillon test.

4.3. Discussion

Nous pouvons tirer quelques enseignements de cet exemple jouet imitant le calcul d'un score d'attribution de crédit bancaire.

- Sans précaution, si un biais est présent dans les données, il est appris et même renforcé par un modèle linéaire élémentaire.
- La suppression naïve de la variable sensible (genre) pour réduire le biais n'y change rien d'où l'importance (art. 10, 5.) d'autoriser la prise d'un risque contrôlé de confidentialité pour intégrer des données personnelles sensibles afin de pouvoir détecter des biais.
- Un algorithme sophistiqué, non linéaire et impliquant les interactions entre les variables, ne fait que reproduire le biais mais, opaque, ne permet plus de justification des décisions si l'effet disproportionné est juridiquement attaquable comme aux USA ($DI < 0,8$). Dans le cas pré-

sent, un simple arbre binaire pénalisé pour contrôler le nombre de feuilles permet de concilier accroissement peu important du biais et explicabilité sans trop pénaliser la précision.

- En présence de proxys du genre comme c'est le cas dans cet exemple, une procédure de *testing* (*counterfactual test*) est complètement inadaptée à la détection *ex-post* d'une discrimination algorithmique. Seule une analyse rigoureuse d'une documentation loyale (art. 11) décrivant les données, la procédure d'apprentissage, les performances, peut donc s'avérer convaincante sur les capacités non discriminatoires d'un algorithme.
- Sur cet exemple, le choix d'un *post-processing* permettant d'atténuer le biais des rapports de cotes conditionnels (taux de faux positifs similaires) selon le genre impacte les trois types de biais pour en réduire simultanément l'importance. C'est une façon de légitimer l'introduction d'une dose de discrimination positive qui réduit le désavantage fait aux femmes sans pour autant nuire aux hommes.
- Finalement dans cet exemple illustratif, un arbre pénalisé pour être suffisamment simple (nombre réduit de feuilles) et assorti d'une touche de discrimination positive fournit une aide à la décision explicable à un client et équitable en terme de risques de la banque vis-à-vis de son genre.
- Certes, dans le cas d'un score de crédit, cela aurait pour conséquence d'accroître le risque de la banque en réduisant la qualité de prévision et augmentant le taux de faux positifs pour les femmes mais lui fournirait des arguments tangibles de communication pour une image "éthique" : des décisions inclusives donc plus équitables et plus explicables sans trop nuire à la précision.

5. Conclusion

Comme le rappelle Meneceur (2021-b) dans une comparaison exhaustive des démarches institutionnelles, les très nombreuses approches éthiques visant à encadrer le développement et l'application des systèmes d'IA ne sont pas des réponses suffisantes et convaincantes pour développer la confiance des usagers. Ceci motive la démarche de la CE aboutissant à la publication de ce projet de règlement alors que le *Conseil de l'Europe envisage également un mélange d'instruments juridiques contraignants et non contraignants pour prévenir les violations des droits de l'homme et des atteintes à la démocratie et à l'État de droit*; la nécessité de conformité se substitue à l'éthique.

L'analyse du projet de règlement européen montre des avancées significatives pour plus de transparence des systèmes d'IA :

- importance fondamentale des données et donc de leur analyse préalable fouillée et documentée,
- évaluation et documentation explicite des performances et donc des risques d'erreur ou de manquement : robustesse, résilience,
- documentation explicite sur les capacités d'interprétation d'un système, d'une décision, à la mesure des technologies et méthodes disponibles,
- prise en compte de certains types de biais : équité sociale dans les données, performances selon des groupes et suivi des risques possibles de discrimination associés,
- enregistrement de l'activité pour une traçabilité du fonctionnement,
- contrôle humain approprié pour réduire et anticiper les risques,
- obligation de fournir la documentation exhaustive à l'utilisateur (système d'IA de l'annexe III), qui est auditée *ex-ante* par un organisme notifié pour les systèmes d'IA de l'annexe II, pour l'obtention du marquage "CE".

Néanmoins ce projet de règlement principalement motivé par une harmonisation des relations commerciales au sein de l'Union selon le principe de sécurité des produits ou de la responsabilité du fait des produits défectueux ne prend pas en compte des dommages pouvant impacter les usagers. Les conséquences ou objectifs de la démarche adoptée par la CE rejoignent d'ailleurs les exigences de la FTC (*Federal Trade Commission*) (Jillson, 2021) de loyauté et transparence vis-à-vis des performances d'un système d'IA commercialisé. Aussi certains droits fondamentaux, bien que retenus comme *exigence essentielle* dans le livre blanc se trouvent pour le moins négligés et ce d'autant plus que les systèmes d'IA à haut risque de l'annexe III ne sont pas concernés par la certification d'un organisme notifié indépendant.

- Plus largement que les seules applications de l'IA, une prise en compte d'une forme de frugalité numérique afin de réduire les impacts environnementaux ne semblent pas, dans ce projet d'*AI Act*, une préoccupation majeure de la CE. Cela concerne la consommation énergétique pour le stockage massif et l'entraînement des algorithmes et la sur-exploitation des ressources minières nécessaires à la fabrication des équipements numériques.

- Il est certes conseillé de rechercher des biais potentiels dans les données (art. 10, 2., (f)) avec même la possibilité de prendre en compte des données personnelles sensibles (art.10, 5.) pour traquer des biais systémiques sources potentielles de discrimination. Néanmoins, l'absence de précisions sur la façon de mesurer ces biais, de les atténuer ou les supprimer dans les procédures d'entraînement laisse un vide potentiellement préjudiciable à l'usager. Alors qu'il est déjà fort complexe pour un usager d'apporter la preuve d'une présomption de discrimination, par exemple par *testing*, lors d'une décision humaine, l'exemple numérique ci-dessus montre que c'est mission impossible face à une décision algorithmique. Seule une procédure rigoureuse d'audit de la documentation décrivant les données, la procédure d'apprentissage et les dispositions mises en place pour gérer, atténuer les biais, peut garantir une protection *a minima* des usagers finaux contre ce type de discrimination. Cette mise en conformité agit comme un renversement de la charge de la preuve mais qui ne bénéficie, pour les systèmes d'IA de l'annexe III, qu'à l'information de l'utilisateur pas, dans l'état actuel, à la protection de l'usager.

- Consciente de ces problèmes la Défenseure des Droits a récemment publié un avis en collaboration avec le réseau européen EQUINET dont les principales conclusions sont résumées dans un communiqué de presse. Elle y *appelle à replacer le principe de non-discrimination (de l'usager) au cœur du projet d'AI Act*. Une des questions essentielles reste à savoir qui pourra, en dehors de l'utilisateur, avoir accès à la documentation d'un système d'IA à haut risque, et donc de pouvoir l'auditer dans de bonnes conditions. Ce sera sans doute à chaque État membre de légiférer sur ces questions.

- Notons que le Laboratoire Nationale de Métrologie et d'Essai (LNE) a pris les devants en proposant un référentiel de certification de processus pour l'IA (LNE 2021). Ce référentiel concerne le processus de conception d'un système d'IA et non la certification du produit final requérant la connaissance de normes encore à définir. Le LNE jouera le rôle d'organisme notifié pour les systèmes de transport de l'annexe II et sa filiale GMED pour les dispositifs de santé sous la responsabilité de l'Agence Nationale de Sécurité des Médicaments comme autorité notifiante.

L'exemple numérique jouet a également pour mérite de montrer clairement l'*interdépendance* de toutes les contraintes : confidentialité, qualité, explicabilité, équité (types de biais), que devrait satisfaire un système d'IA pour gagner la confiance des usagers. Il montre aussi que le problème ne se réduit pas à un simple objectif de minimisation d'un risque quantifiable pour l'obtention d'un meilleur compromis. C'est plutôt la recherche d'une moins mauvaise solution imbriquant des choix techniques, économiques, juridiques, politiques qu'il sera nécessaire de clairement expliciter dans la documentation rendue obligatoire par l'adoption à venir d'un *AI Act* qui serait, de toute façon et malgré les limites actuelles du projet de texte, une avancée notable pour plus de transparence.

Références

- Bachoc F., Gamboa F., Halford M., Loubes J.-M., Risser L. (2020). Entropic Variable Projection for Model Explainability and Intepretability, arXiv preprint : 1810.07924.
- Barocas S., Selbst A. (2016). Big Data's Disparate Impact, *California Law Review*, 104, 671.
- Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R., Herrera F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward Responsable AI, arXiv.
- Bénesse C., Gamboa F., Loubes J.-M., Boissin T. (2021). Fairness seen as Global Sensitivity Analysis, ArXiv, à paraître. responsible AI, *Information Fusion*, Vol. 58, pp 82-115.
- Besse P. (2021). Médecine, police, justice : l'intelligence artificielle a de réelles limites, The Conversation, 01/12/2021.
- Besse P., Besse Patin A., Castets Renard C. (2020). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, *Statistique & Société*, 3, pp 21-53.
- Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019). L'IA du Quotidien peut elle être Éthique ? Loyauté des Algorithmes d'Apprentissage Automatique, Statistique et Société, VOl6 (3), pp 9-31.
- Besse P., del Barrio E., Gordaliza P., Loubes J-M., Risser L. (2021) A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set, *The American Statistician*, DOI : 10.1080/00031305.2021.1952897, version en accès libre.
- Breiman L. (2001). Random forests, *Machine Learning* 45, 5–32.
- Campisi N. (2021). From Inherent Racial Bias to Incorrect Data—The Problems With Current Credit Scoring Models, Forbes Advisor.
- Castets Renard C., Besse P. (2022). Responsabilité ex ante de l'AI Act : entre certification et normalisation, à la recherche des droits fondamentaux au pays de la conformité, dans "Un droit de l'intelligence artificielle : entre règles sectorielles et régime général. Perspectives de droit comparé", dir. C. Castets-Renard et J. Eynard, Bruylant (à paraître).
- CE (2019) Lignes Directrices pour une IA digne de Confiance, rédigé par un groupe d'experts européens.
- CE (2020) Livre blanc sur l'intelligence artificielle: une approche européenne d'excellence et de confiance.
- CE (2021). Règlement du parlement et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union.
- De-Arteaga M., Romanov A. et al. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp 120–128.
- Défenseure des Droits (2020). Algorithmes: prévenir l'automatisation des discriminations, rapport.
- Défenseur des Droits, CNIL (2012). Mesurer pour progresser vers l'égalité des chances. Guide méthodologique à l'usage des acteurs de l'emploi.

– Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E., Roth D. (2019). Comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 329–38.

– Goglin C. (2021). Discrimination et IA : comment limiter les risques en matière de crédit bancaire, The Conversation, 23/09/2021.

– Hurlin C., Pérignon C., Saurin S. (2021) The fairness of credit score models, preprint SSRN.

– Jillson E. (2021). Aiming for truth, fairness, and equity in your company's use of AI, blog, consulté le 29/05/2021.

– Kozodoi N., Jacob, J. Lessman, S. (2021). Fairness in credit scoring: assessment, implementation and profit implications, preprint arXiv.

– Larson J., Mattu S., Kirchner L., Angwin J. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2020.

– Lê, S., Josse, J., Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18.

– Liu X., L. Faes, A. U. Kale et al. (2019), A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, *The Lancet Digital Health*, vol. 1, pp. e271–e297.

– LNE (2021). Référentiel de Certification du Processus IA, Laboratoire Nationale de Métrologie et d'Essais.

– Meneceur Y. (2021). Analyse des principaux cadres supranationaux de régulation de l'intelligence artificielle : de l'éthique à la conformité, projet d'étude, Institut des Hautes Études sur la Justice (IHEJ), version d'étude du 27/05/2021.

– Raghavan M., Barocas S., Kleinberg J., Levy K. (2019) Mitigating bias in Algorithmic Hiring : Evaluating Claims and Practices, *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

– Rich J. (2014). What Do Field Experiments of Discrimination in Markets Tell Us? A Meta Analysis of Studies Conducted since 2000, *IZA Discussion Paper*, No. 8584.

– M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, AIX-COVNET, J. H. F. Rudd, Evis Sala, C.-B. Schönlieb (2021), Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, *Nature Machine Intelligence*, 3, pages 199–217.

– Verzelen N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons, *Electron. J. Statist.*, 6, 38 - 90.

– Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min. Knowl. Disc.*, 31, p 1060–89.