Détecter,évaluer les risques des impacts discriminatoires des algorithmes d'IA

Philippe Besse

Détecter, évaluer les risques des impacts discriminatoires des algorithmes d'IA

Contribution au séminaire conjoint: Défenseur des Droits & CNIL

Philippe Besse*

28 Mai 2020

Résumé

Faisant suite au déploiement du RGPD, la Commission Européenne a publié, en février 2020 un livre blanc pour une approche de l'IA basée sur l'excellence et la confiance et dont les recommandations sont largement issues du guide pour une IA digne de confiance rédigé en 2019 par un groupe d'experts européens. Au delà des questions prioritaires de protection des données au cœur des missions de la CNIL, ce livre blanc soulève avec insistance d'autres questions relatives aux risques des impacts des algorithmes d'apprentissage automatique sur notre société : qualité, reproductibilité de décisions algorithmiques, opacité des algorithmes et explicabilité des décisions, biais et risques de discrimination. En nous basant sur des exemples concrets : score de crédit, pré-recrutement prédictif, nous décrivons quels outils, procédures, indicateurs (cf. tutoriel), pourraient participer à la construction d'un DIA ou *Discrimination Impact Assessment* souhaité par le rapport Villani (2018) et cohérent avec la liste d'évaluation du groupe des experts européens. Les exemples traités montrent les difficultés et même l'impossibilité d'un audit *ex post* d'un système d'intelligence artificielle (SIA) sur la base d'algorithmes d'apprentissage. Nous concluons sur quelques recommandations dont la nécessité de la mise en place d'audits sur la base de documentations précises et exhaustives *ex ante* d'un SIA.

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

*Université de Toulouse – INSA, Institut de Mathématiques – UMR CNRS 5219, Observatoire International des Impacts Sociétaux de l'IA et du Numérique – Université Laval

English title : Detecting, assessing the risks of discriminatory impacts of AI algorithms ; Contribution to the joint seminar : Défenseur des Droits & CNIL.

Abstract : Following the deployment of the RGPD, the European Commission published in February 2020 a White Paper for an approach to AI based on excellence and trust. Its recommendations are largely derived from the guide for trustworthy AI drawn up in 2019 by a group of European experts. Beyond the priority data protection issues at the very heart of the CNIL's missions, this white paper insistently raises other issues related to the risks of the impacts of machine learning algorithms on our society : quality, reproducibility of algorithmic decisions, opacity of algorithms and explanability of decisions, bias and risks of discrimination. Based on practical examples : credit scoring, predictive hiring, we describe which tools, procedures, indicators (cf. tutorial), could participate in the construction of a Discrimination Impact Assessment (DIA) desired by the Villani report (2018) and consistent with the evaluation list of the European experts group. These examples reveal the difficulties and even the impossibility of an ex-post audit of an artificial intelligence system (AIS) based on learning algorithms. We conclude with a few recommendations, including the need to set up audits on the basis of accurate and exhaustive ex-ante documentation of an AIS.

keywords : artificial intelligence, machine learning, bias, discrimination, disparate impact, testing, european regulation.

1 Introduction

1.1 Du RGPD à une IA éthique

La publication du RGPD (2018) et son intégration dans les lois nationales ont considérablement impacté la gestion des données dont celles impliquant des personnes physiques. La CNIL développe un outil d'évaluation du bon usage des données sous la forme d'un logiciel d'Analyse d'Impact relative à la Protection des données (*Privacy Impact Assessment*). Par ailleurs, l'Autorité de la Concurrence traque les risques d'*entrave à la concurrence* de la part des moteurs de recherche, sites de comparateurs de prix.

Ces points ne seront pas abordés mais rappelons le considérant (71) du RGPD (2018) qui recommande :

> [...] Afin d'assurer un traitement équitable et transparent à l'égard de la personne concernée, [...] le responsable du traitement *devrait* utiliser des *procédures mathématiques ou statistiques* adéquates aux fins du profilage, appliquer les mesures techniques et organisationnelles appropriées pour faire en sorte, en particulier, que les facteurs qui entraînent des erreurs dans les données à caractère personnel soient corrigés et que le *risque d'erreur soit réduit au minimum*, sécuriser les données à caractère personnel d'une manière qui *tienne compte des risques* susceptibles de peser sur les intérêts et les droits de la personne concernée, et prévenir, entre autres, les *effets discriminatoires* [...]

Les risques provoqués par les impacts dus à l'opacité, aux biais, aux erreurs des décisions algorithmiques n'ont finalement pas ou peu été pris en compte dans une réglementation visant en priorité la protection des données. Ils ont été en revanche largement commentés dans de très nombreuses déclarations, chartes pour une IA éthique au service de l'humanité. Citons par exemple :

- Enjeux Éthiques de l'Algorithme et de l'Intelligence artificielle (CNIL 2017) également très présents dans le
- 5ème partie du rapport Villani (2018), et la
- Déclaration de Montréal pour un développement responsable de l'IA (2018) qui cite par ailleurs pas moins de 28 rapports nationaux ou professionnels.

1.2 Vers un nouveau règlement pour l'IA

De son côté la Commission Européenne (CE) a réuni un groupe d'experts indépendants de haut niveau sur l'IA qui propose un ensemble de

- lignes directrices en matière d'éthique pour une IA digne de confiance (2019) qui ont été suivies par la publication d'un livre blanc :
- Intelligence Artificielle: une approche européenne axée sur l'excellence et la confiance (2020).

Ce livre souligne l'importance prise par l'IA, qui *combine données, algorithmes et puissance de calcul*, dans tous les aspects de la vie des citoyens, en liste les

bénéfices attendus, mais met également en exergue les *risques potentiels, tels que l'opacité de la prise de décisions, la discrimination*, qui accompagnent son développement et sa mise en œuvre. C'est un enjeu majeur car l'acceptabilité de l'IA et donc son adoption par les citoyens ne seront possibles que si celle-ci est *digne de confiance*. La CE, qui ambitionne de faire de l'Europe un *acteur mondial de premier plan en matière d'innovation dans l'économie fondée sur les données et dans ses applications*, insiste sur la nécessité de cette confiance fondée sur les *droits fondamentaux* de la *dignité humaine et la protection de la vie privée.*

Il s'agit donc pour la CE de proposer les *éléments clefs d'un futur cadre réglementaire* basé sur un *écosystème de confiance* en prenant en compte les lignes directrices en matière d'éthique élaborées par le groupe d'experts et dont la *liste d'évaluation* servirait de base pour un *programme indicatif destiné aux développeurs de l'IA* et une *ressource mise à la disposition des établissements de formation*. La CE insiste sur la liste des exigences énumérées par le groupe d'experts en remarquant que si certaines sont prises en compte par les régimes législatifs ou réglementaires existants, d'autres (*e.g.* transparence, contrôle humain) ne sont pas couvertes ou qu'il est de toute façon *difficile de déceler et de prouver d'éventuelles infractions à la législation, notamment aux dispositions juridiques qui protègent les droits fondamentaux*, à cause de l'opacité des algorithmes d'IA.

Par ailleurs, suivant en cela le groupe d'experts, la CE insiste tout particulièrement sur une classe de systèmes d'intelligence artificielle (SIA) basés sur des *algorithmes d'apprentissage automatique* et donc sur le rôle fondamental des *données* utilisées pour leur entraînement.

Remarque : algorithmes déterministes ou procéduraux ou encore d'IA symbolique. Cet article laisse apparemment de côté cette classe 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 de désavantage ou risques de discrimination indirecte malgré une apparente neutralité. La détection de ces risques relève de l'analyse experte des règles de décisions codées dans l'algorithme. Néanmoins, la complexité de l'algorithme peut être telle (cf. Parcoursup) qu'une l'analyse experte *ex post* ne sera pas en mesure d'évaluer l'étendue des risques. En conséquence, l'algorithme déterministe peut être traité avec le même niveau d'opacité et les mêmes outils qu'un algorithme d'apprentissage non linéaire.

1.3 Focus sur la non-discrimination ou équité d'une décision algorithmique

Parmi les exigences largement partagées pour une IA de confiance, trois questions sont essentielles (Besse et al. 2019-a) et doivent s'imposer à des SIA conduisant à des décisions impactant des personnes. Ces décisions peuvent être la conséquence d'une prévision d'un algorithme d'apprentissage opaques entraînés sur des données ou celle de l'exécution d'un ensemble complexe de règles de décision déterministes codées dans un programme.

1. Quel est le niveau de précision, de reproductibilité d'une décision issue d'un SIA par apprentissage ?
2. Quelle explication d'une décision peut-elle être apportée en fonction du niveau (concepteur, utilisateur, usager) et du domaine concernés ?
3. Quels sont les risques de biais discriminatoires à l'encontre d'un groupe vulnérable protégé par la loi ?

Nous allons nous intéresser plus particulièrement aux questions difficiles de *détection des risques de discrimination* tout en remarquant que celles-ci impliquent les deux précédentes de même que le type d'informations disponibles car autorisées par la loi (RGPD) dans les bases de données.

La liste d'évaluation du groupe d'experts, base de réflexion de la CE, réserve la section 5 *Diversité, non-discrimination et équité* aux questions de discrimination. Relevons seulement trois questions de cette longue liste adressées aux concepteurs d'un SIA :

— Avez-vous prévu une définition appropriée de l'équité que vous appliquez dans la conception des SIA ?
— Avez-vous mis en place des processus pour tester et contrôler les biais éventuels au cours de la phase de mise au point, de déploiement et d'utilisation du système ?
— Avez-vous prévu une analyse quantitative ou des indicateurs pour mesurer et tester la définition appliquée de l'équité ?

Ce document ainsi que le tutoriel disponible sur un dépôt public (Besse et al. 2020), qui permet de reproduire les résultats numériques et graphiques ci-après, sont une contribution pour aider un concepteur d'un SIA à apporter des éléments de réponse à ces questions. La section suivante s'intéresse aux modes de détection d'une discrimination issue d'une décision avant d'aborder, section 3, le cas où cette décision est issue d'un algorithme d'apprentissage automatique ou que la complexité rend opaque. Un exemple numérique analogue au calcul d'un score de crédit illustre section 4 ces outils quantitatifs avant d'étudier section 5 les risques spécifique de discrimination à l'embauche de la part d'algorithmes de pré-recrutement. Ces éléments concrets nous permettent de conclure sur les modes opératoires à nécessairement mettre en place pour l'audit d'un SIA et quelques recommandations regroupées dans le tableau 1.

2 Détecter une discrimination

Les lois européennes comme américaines interdisent explicitement toute forme de discrimination vis-à-vis d'une personne (discrimination directe) ou d'un groupe (discrimination indirecte). Néanmoins, qualifier, prouver une discrimination, restent des questions difficiles à traiter et qui dépendent des cadres juridiques, usages culturels et aussi des domaines concernés : accès à l'emploi, le crédit, l'assurance, la santé... dans l'exemple très sensible de l'accès à l'emploi, des approches radicalement différentes sont mises en œuvre : *testing vs.* effet disproportionné.

TABLE 1 – *Proposition de recommandations*

1. Anticiper un futur règlement européen pour une *IA digne de confiance* intégrant une liste d'évaluation *ex ante* des systèmes d'IA sur le même principe que l'analyse d'impact relatif à la protection des données.
2. Réglementer la production et la diffusion de la documentation produite *ex ante* sur les risques algorithmiques, support potentiel indispensable d'un audit indépendant.
3. Réglementer l'identification des responsabilités humaines à toutes les étapes — données, entraînement, test, certification, exploitation — d'un système d'IA pour un contrôle qualité et une amélioration tout le temps de sa durée de vie.
4. Promouvoir (normalisation ?) les indicateurs statistiques de détection et mesure de discriminations algorithmiques indirectes.
5. Faciliter la production de statistiques anonymes d'origine ethnique nécessaires à l'évaluation d'une discrimination potentielle.

2.1 Testing

La détection et même la preuve d'une discrimination directe envers une personne peut être obtenue par *testing*. Cette pratique consiste à adresser à des dates distinctes deux dossiers, par exemple de candidature à un emploi. A l'exception de la caractéristique discriminatoire à tester : genre, origine ethnique, tranche d'âge, quartier d'habitation... les dossiers sont strictement similaires tout en introduisant des différences mineures afin d'éviter d'éventer le procédé. Son usage à été élargi (cf. Riach et Rich 2002) avec le déploiement d'enquêtes systématiques afin de viser l'objectif d'une mesure statistique de la discrimination indirecte envers un groupe. Les communautés académiques en Sociologie et Économie ont produit une vaste bibliographie à ce sujet (Rich 2014). En France, c'est la doctrine officielle diffusée par le Comité National de l'Information Statistique et déployée périodiquement par la DARES (Direction de l'Animation, des Études, de la Recherche et des Statistiques) lorsqu'il s'agit d'étudier les risques de discrimination à l'embauche. D'autres enquêtes par *testing* se ont également ciblé l'accès à l'assurance, au crédit ou encore au logement (cf. les rapports de recherche du TEPP).

2.2 Effet disproportionné

Aux USA, une approche très différente est développée avec la notion d'*adverse* ou *disparate impact* (effet disproportionné). L'évaluation de l'effet disproportionné consiste à estimer le rapport de deux probabilités : probabilité d'une décision favorable pour une personne du groupe sensible au sens de la loi sur la

même probabilité pour une personne de l'autre groupe. Elle est appliquée depuis 1971 (Barocas et Selbst 2017) pour mesurer des discriminations indirectes dans l'accès à l'emploi, le logement, et a donné lieu à une réglementation officielle de son usage notamment pour l'accès à l'emploi :

> **Civil Rights act & Code of Federal Regulations**
> Title 29 - Labor: Part 1607—Uniform Guidelines on Employee Selection Procedures (1978)
> — D. *Adverse impact and the "four-fifths rule."* A *selection rate* for any race, sex, or ethnic group which is *less than four-fifths (4/5) (or eighty percent)* of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. *Smaller differences* in selection rate may nevertheless constitute adverse impact, where they are *significant in both statistical and practical* terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group.

L'estimation de ce rapport de probabilités (*odds ratio*) est donc comparée à $0,8$. Une valeur inférieure n'induit pas nécessairement des poursuites juridiques mais oblige une entreprise à justifier, pour des raisons économiques, les raisons de ce déséquilibre.

2.3 Remarques

Les éléments de ces approches statistiques sont également présents dans un guide publié par le Défenseur des Droits et la CNIL (2012). Il décrit une approche méthodologique à l'intention des acteurs de l'emploi pour mesurer et progresser dans l'égalité des chances sans volonté coercitive ni obligation juridique. En préalable, ce guide pose la question de l'opportunité de construire des statistiques ethniques alors que, contrairement aux USA, l'origine des personnes ne peut être enregistrée dans une base de données. Cette apparente protection des droits des personnes soulève un problème lorsqu'il est question d'évaluer une possible discrimination. La difficulté peut être contournée en adoptant une identification de l'origine par le patronyme au prix d'une perte sans doute mineure mais à évaluer de précision. Ce guide évoque la pratique du *testing* mais incite également les services de ressources humaines d'une entreprise à produire des tableaux statistiques (tables de contingence) desquels il serait facile d'extraire une évaluation quantitative de l'effet disproportionné.

Chacune des approches : *testing vs. disparate impact* présente des avantages

mais également des défauts, biais ou difficultés de mise en œuvre. Le *testing* met bien en évidence une discrimination directe, intentionnelle, et peut conduire à une action en justice lorsqu'une personne est concernée. En revanche, utilisée lors d'une enquête systématique, il déploie des dossiers fictifs, fournis des résultats indicatifs, qui ne sont pas représentatifs de la politique d'embauche effective d'une entreprise sur l'ensemble des ses postes. Les enquêtes menées par la DARES ne conduisent pas à des actions en justice et la récente stratégie *name and shame* du gouvernement stigmatisant certaines entreprises a suscité de vives polémiques en janvier 2020.

Les enquêtes par *testing* ne nécessitent pas une participation des entreprises concernées mais sont d'un coût élevé et soulèvent de lourdes difficultés pour tenter d'approcher la réalité des embauches. En revanche, l'évaluation de l'effet disproportionné est de coût très faible mais implique une contribution loyale des services de ressources humaines ou une obligation réglementaire. Il est éventuellement biaisé puisque les dossiers ne sont pas identiques et nécessite donc une analyse ou la recherche d'autres explications possibles mais confondues des écarts observés.

3 Discrimination algorithmique

Une décision algorithmique ajoute une couche d'opacité sur une situation déjà complexe.

3.1 Algorithmes d'apprentissage automatique

Depuis une dizaine d'années, beaucoup plus pour l'accès au crédit, des procédures algorithmiques d'apprentissage supervisé sont déployées pour automatiser l'aide à la décision. Le principe consiste, sur la base de données massives historicisées, à estimer un modèle statistique ou entraîner un algorithme qui reproduit au mieux les décisions déjà connues afin d'appliquer ce modèle ou algorithme à de nouvelle situations. Les décisions automatisées qui en découlent, constituent l'application quotidienne et très largement la plus répandue des SIA dans nos quotidiens. Tous les domaines sont concernés : recommandations pour la vente en ligne, risque de rupture de contrat, de fraude, de défaillance d'un système, de récidive... aide au diagnostic médical.

3.2 Indicateurs statistiques de discrimination

Le problème émergeant de la discrimination algorithmique s'exprime simplement : si un algorithme est entraîné sur des données biaisées, il reproduit très fidèlement ces biais systémiques ou de société ; plus grave, il risque même de les renforcer. Très prolixe, le monde académique a proposé quelques dizaines d'indicateurs (Zliobaitė 2017) afin d'évaluer des biais potentiels. Néanmoins, beaucoup de ces indicateurs s'avèrent très corrélés ou redondants (Friedler et

al. 2019). Empiriquement, trois niveaux de biais discriminatoires doivent être pris en compte en priorité :

1. L'effet disproportionné reflet du biais social ou de population par lequel un groupe est historiquement (*e.g.* revenu des femmes) désavantagé. La mise en évidence de ce biais soulève des questions techniques, politiques évidentes. Renforcer algorithmiquement ce biais serait ouvertement discriminatoire, il importe de détecter, éliminer, un tel risque. Serait-il politiquement opportun d'introduire automatiquement une part de discrimination positive afin d'atténuer la discrimination sociale ? C'est techniquement l'objet d'une vaste littérature académique nommée apprentissage équitable (*fair learning*) et évoqué dans le travail des experts (ligne directrice 52) pour *améliorer le caractère équitable de la société.*
2. Les taux d'erreur de prévision et donc les risques d'erreur de décisions sont-ils les mêmes pour chaque groupe ? Ainsi, si un groupe est sous-représenté dans la base d'apprentissage, il est très probable que les décisions le concernant seront moins fiables. C'est typiquement le cas en reconnaissance faciale et ce risque est également présent dans les applications de l'IA en santé (Besse et al. 2019-b).
3. Même si les deux critères précédents sont trouvés équitables et surtout si les taux d'erreur identiques sont relativement importants, les erreurs peuvent être dissymétriques (plus de faux positifs, moins de faux négatifs) au détriment d'un groupe. Cet indicateur (comparaison des rapports de cote ou *odds ratio*) est ainsi au cœur de la controverse concernant l'évaluation COMPAS du risque de récidive aux USA (Larson et al. 2016).

3.3 Difficultés d'évaluation

Contrairement à des prises de position très naïves, des décisions algorithmiques ne sont pas plus objectives que des décisions humaines. Il est même facile de montrer sur des exemples (numériques ci-après, De Arteaga et al. 2019) que les biais humains sont fidèlement reproduits voire amplifiés même si la variable sensible (genre, origine, âge...) est absente de la base de données car cette information est présente, d'une façon ou d'une autre, dans les autres variables jouant le rôle de variables de substitution ou *proxy*. Autre conséquence importante de cette situation, le *testing* est complètement inopérant (cf. section suivante) pour détecter une discrimination face à un algorithme.

En conséquence, les biais, risques de discrimination, doivent être soigneusement évalués très en amont lors de la constitution des bases de données et lors de la procédure d'apprentissage afin de les corriger : *fairness by design*, au risque de ne plus être à même de pouvoir les détecter. C'est d'autant plus nécessaire que beaucoup des algorithmes d'apprentissage automatique (*e.g.* réseaux de neurones profond ou pas) parmi les plus performants sont opaques à toute interprétation élémentaire, à toute explication des décisions, contrairement à des modèles élémentaires de statistique ou d'arbre de décision.

4 Exemple numérique de discrimination algorithmique

Le problème s'énonce simplement : un algorithme entraîné à prendre des décisions à partir de données sociales biaisées reproduit fidèlement ces biais et peut même les amplifier et donc induire de fortes discriminations en lien avec le sexe, l'âge, l'origine des personnes. Nous proposons d'illustrer sur un exemple numérique élémentaire les difficultés rencontrées pour la détection et l'évaluation de ces risques fondamentaux.

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. L'étude complète et les codes de calcul sont disponibles (Besse et al. 2020) mais l'illustration est limitée à un résumé succinct de l'analyse de la discrimination selon le sexe.

4.2 Résultats

Les données ont été aléatoirement réparties en deux échantillons d'apprentissage (36 000), destinés à l'estimation des modèles ou entraînement des algorithmes, et de test (9000) pour évaluer les différents indicateurs. Les résultats sont regroupés dans la figure 1.

Ils mettent en évidence un biais 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. Il 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 est que le *testing* (changement de genre toutes choses égales par ailleurs) ne détecte plus ($DI = 0.90$) aucune discrimination !

Un algorithme non linéaire plus sophistiqué (*random forest*) est très fidèle au biais des données avec un indicateur ($DI = 0,36$) pas significativement différent de celui du biais de société et fournit une meilleure précision : $0,86$ au

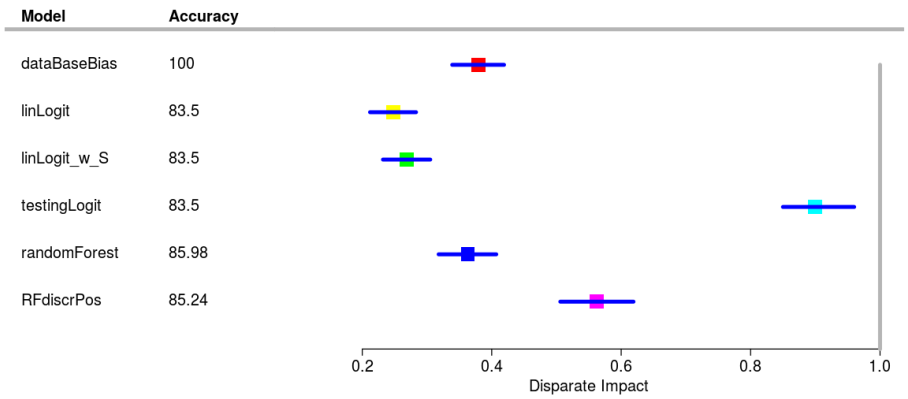

FIGURE 1 – *Précision de la prévision (accuracy) et effet disproportionné estimé par un intervalle de confiance sur un échantillon test (taille 9000) pour différents modèles ou algorithmes d'apprentissage.*

lieu de $0,84$ pour la régression logistique. Cet algorithme ne discrimine pas plus mais c'est au prix de l'interprétabilité 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. Enfin, la dernière ligne propose une façon simple, 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 : $0,3$ au lieu de celui par défaut de $0,5$ pour les hommes. C'est une façon, parmi beaucoup d'autres, d'introduire une part de discrimination positive et d'atténuer le biais pour une *société plus équitable*.

Les autres types de biais sont également à considérer. Par 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% pour les hommes. Il faut donc considérer le troisième type de biais pour se rendre compte que c'est finalement à leur désavantage. Le taux de faux positifs est plus important pour les hommes $(0,08)$ que pour les femmes $(0,02)$. 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 $(0,38)$. Noter que la procédure élémentaire d'atténuation du biais en entraînant deux algorithmes, un pour chaque genre, conduit à une légère augmentation du taux d'erreur pour les femmes, qui se rapproche un peu

de celui des hommes, et surtout produit un taux de faux positifs plus élevés pour les femmes. Aussi, sur cet exemple, l'introduction d'une dose de discrimination positive intervient sur les trois types de biais pour en réduire l'importance.

4.3 Discussion

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

- Sans précaution, si un biais est présent dans les données, il est reproduit et même renforcé par un modèle linéaire élémentaire.
- Un algorithme plus sophistiqué, non linéaire et impliquant les interactions entre les variables, ne fait que reproduire le biais mais, opaque, ne permet plus de justification des choix si l'effet disproportionné est juridiquement attaquable ($DI < 0,8$).
- La procédure de *testing*, déjà peut convaincante pour évaluer une discrimination indirecte *ex post*, est complètement inadaptée face à une procédure algorithmique.
- Actuellement en Europe, une ou un *data scientist* est libre de produire ce qu'il peut ou veut, en fonction de ses compétences et de sa déontologie personnelle : de l'algorithme élémentaire interprétable mais discriminatoire à celui incluant une part arbitraire de discrimination positive. Aucune procédure de contrôle que ce soit *ex ante* ou *ex post*, n'est en vigueur à ce jour pour le remettre en cause.
- La recherche d'une moins mauvaise solution sera l'affaire d'un compromis entre les trois exigences de base pour une IA de confiance : contrôle de la discrimination, qualité (robustesse réplicabilité) d'une décision et explicabilité de cette décision. En effet, le meilleur algorithme en termes de précision est opaque, ininterprétable, et donc inadapté pour éviter aux USA, une procédure judiciaire si l'effet disproportionné est trop important. De plus, la correction ou l'atténuation de l'effet disproportionné entraîne une dégradation de la qualité de la prévision. Les récents travaux de recherche en apprentissage équitable visent cette recherche de meilleur compromis.

En résumé, la détection d'un risque algorithmique de discrimination indirecte vis à vis d'un groupe est une question complexe basée sur l'estimation d'indicateurs statistiques impliquant également les autres exigences de qualité et explicabilité. Cette estimation est de plus soumise à l'accès à l'information sensible dont l'enregistrement (*e.g.* origine ethnique) peut être interdite par le RGPD ; interdiction contournable par des procédés (*e.g.* analyse du patronyme) pouvant nuire à la précision.

5 Discrimination algorithmique à l'embauche

5.1 Contexte

L'exemple précédent est certes très pédagogique car il permet d'expliciter tous les aspects techniques mais n'est pas complètement représentatif de la réalité des problèmes rencontrés lors du déploiement d'un SIA. Une approche plus réaliste se heurte immédiatement au manque de disponibilités de données pour des raisons légitimes et légales de confidentialité ou secret commercial. Les analyses de cas de discrimination algorithmique sont néanmoins très nombreuses dans la littérature essentiellement aux USA ; citons le cas des pratiques de police prédictive (Ferguson 2017), de justice prédictive (Larson et al. 2016) ou de l'accès aux soins (Lee et al. 2019). D'autres sont à rechercher dans les rapports annuels de l'institut de recherche AINow (Myers West et al. 2019) de l'Université de New-York. Pour des raisons tant juridiques que culturelles ou technologiques, la France est moins concernée par les exemples précédents alors que le risque de discrimination pour l'accès à l'emploi est un sujet d'actualité récurrent et identifié comme une application à risque élevé par la Commission Européenne.

Avec un retard classique sur la situation aux USA, nous assistons à une expansion très rapide des offres de plateformes, sites, logiciels de ressources humaines, proposant d'inclure des capacités d'IA dans les phases de gestion des candidatures à des offres d'emploi. Ceci peut concerner des analyses automatiques de CV, biographies ou vidéos. Une recherche sur internet avec les mots clefs "recrutement prédictif" est très informative à ce sujet. Nous ne ferons pas une étude exhaustive des offres proposées mais nous limiterons à quelques exemples à comparer à l'offre américaine.

5.2 Situation aux USA

Raghavan et al. (2019) proposent une étude systématique des annonces et pratiques de 18 entreprises qui vendent des solutions algorithmiques de pré-recrutement de candidats à l'embauche. L'article est focalisé sur les efforts mis en œuvre pour détecter et éventuellement atténuer les biais discriminatoires. Par souci d'homogénéité, l'analyse est volontairement limitée aux présentations disponibles sur internet. Toutes ces entreprises, sauf trois, se préoccupent des biais pouvant résulter des décisions automatiques, sept mentionnent la règle des 4/5èmes ou le besoin de contrôler l'effet disproportionné ; certaines proposent de l'atténuer, non pour des raisons éthiques, mais d'argumentation commerciale : économiser des procédures coûteuses de justification juridiques en cas d'effet disproportionné trop marqué. L'article se conclut par un ensemble de recommandations appelant à plus de transparence dans les choix technologiques, la prise en compte d'autres critères de biais que l'effet disproportionné qui peut être difficile voire impossible à évaluer sans l'accès à l'information sensible.

Un autre article (De-Arteaga et al. 2919) complète ce point de vue par l'analyse de près de 400.000 biographies professionnelles en anglais aspirées sur internet. Le contexte est différent car il ne s'agit pas d'une procédure de pré-

embauche mais de savoir si le contenu des biographies permet de prévoir le poste de travail occupé parmi les 38 identifiés comme étant les plus fréquents de professeur à rappeur. Différents types de représentation des données et algorithmes de prévision sont testés et l'étude des biais est abordée de façon spécifique. Les auteurs montrent que la précision de la prévision es sensiblement la même en prenant ou non en compte l'information relative au genre présente dans la biographie avec le prénom et le pronom personnel qui est à la troisième personne. Ils montrent également que pour les professions pour lesquelles le déséquilibre est le plus marqué, les erreurs commises aggravent le déséquilibre. Ceci est obtenu en comparant les taux de faux positifs selon le genre. Cet article ne met pas en évidence les biais de société de premier niveau (effet disproportionné ou de population) sur la sélection à l'emploi mais ceux de deuxième et troisième niveau qui impactent sur les déséquilibres de répartition des emplois. Ceci est observé même si l'information sur le genre n'est pas présente explicitement dans les données car elles est implicitement prévisible à partir des contenus des biographies.

Une illustration concrète de ces phénomènes a été observée chez *Amazon* (2018) dont la procédure automatique de présélection sur des postes techniques ne sélectionnait plus de femmes. Comme l'algorithme devait considérablement rendre opaque la prise de décision et sans doute pour éviter toute complication juridique sur la base du code de régulation fédéral, *Amazon* s'est empressé de stopper l'utilisation de ce système d'IA.

5.3 Quelques exemples en France

Les aides technologiques à l'embauche (*hiring tech*) sont bien moins développées en France mais de nombreuses initiatives émergent rapidement en surfant sur la vague des battages médiatiques *big data* et IA. Relevons quelques exemples représentatifs mais très peu renseignés issus d'une analyse rapide de l'offre présente sur internet.

CV Catcher de `JobiJoba` propose une stratégie spécifique d'analyse automatique de CV mais pas pour opérer une pré-sélection de ceux-ci. Il s'agit d'orienter le candidat, parmi les quelques centaines d'offres de postes permanents de l'entreprise, vers ceux les plus adaptés à son profil. La démarche est sans doute utile pour des types d'emplois spécifiques dont le marché est tendu mais aucune indication n'est fournie sur la manière de procéder.

`Assessfirst` est très présente dans les recherches sur les navigateurs ou les classements. Cette entreprise annonce un *algorithme de prédiction d'affinité*. L'objectif, pour les professionnels de ressources humaines, serait de *prévoir la capacité naturelle* de deux personnes à travailler ensemble : collaborateur et son futur manager, deux membres d'une équipe en mode projet. Cette affinité est calculée sur la base d'un questionnaire de personnalité, d'un test de raisonnement et d'un questionnaire de motivation. Ces données seraient *non-biaisées, exemptes de tout facteurs externes ou historiques pouvant affecter la relation.* Malheureusement aucune citation ou rapport de présentation ne viennent étayer ces affirmations.

Le cabinet de recrutement clémentine assure que les algorithmes de recrutement prédictif sont *loin de l'idée de déshumaniser les méthodes de recrutement traditionnelles* mais permettraient *avant tout de diminuer la discrimination à l'embauche pour que tout un chacun soit jugé selon des critères de comparaison identiques*. Les algorithmes ne se soucieraient pas de *savoir si un candidat a la peau noire, est une femme, n'est plus tout jeune ou est handicapé. Big Data* et algorithmes seraient donc utilisés pour prédire la probabilité de succès ou d'échec d'un candidat potentiel en comparant son profil à celui de milliers d'autres salariés occupant un poste similaire. Ces annonces n'en disent pas plus sur les manières de procéder et mesures prises pour en éviter les biais.

La société 365talents propose un autre type de service. Il s'agit d'offrir aux employés d'une entreprise un système automatique de recommandations de postes en mobilité. Le système serait capable de *capter les compétences implicites des employées afin d'estimer des niveaux d'expertise* à partir des sources de données connues (CV, entretien de carrière) ou déclarées. Le collaborateur peut alors recevoir des *suggestions personnalisées de postes en mobilité interne, de missions ou de formation*. L'entreprise est cette fois plus précise en indiquant que le système est basé sur un ou des algorithmes de traitement du langage naturel (*naturel language processing* ou NLP) utilisant une technique de représentation vectoriel des mots du dictionnaire (*word embedding*). Nous pouvons ensuite faire l'hypothèse qu'un système de recommandations du type de ceux utilisés dans la vente en ligne permet de rapprocher des représentations quantitatives des compétences avec celles des profils de poste. Il n'est fait aucune mention d'une évaluation de la qualité attendu de ce procédé ni des risques de biais discriminatoires alors que ceux-ci sont connus et documentés pour les méthodes de plongement de mots (Brunet et al. 2019).

easyrecrue propose de faciliter le pré-recrutement par une analyse automatique de vidéos. Pas de reconnaissance faciale, de la gestuelle ou des intonations mais une analyse du vocabulaire employé afin de sélectionner le groupe de candidats qui seront convoqués pour un entretien. La procédure est cette fois mieux documentée car accompagnée, par un conseil scientifique universitaire, elle est le résultat d'un travail de thèse (CIFRE) qui a donné lieu à une publication (Hemamou 2018). Cet article compare les performances des combinaisons de plusieurs descripteurs de diversité lexicale, de vectorisation de mots (`word2vec`...) et retient comme meilleur prédicteur un algorithme non linéaire dits de *forêts aléatoires* (agrégation opaque d'arbres binaires de décision). L'algorithme a été entraîné sur une sélection de 305 parmi 607 vidéos pour lesquels l'évaluation par un sélectionneur humain était tranchée : "avis favorable" ou "avis réservé". La qualité est évaluée par un critère classique d'aire sous la courbe ROC (0,69). En résumé et pour apporter des éléments succincts de réponse à la liste d'évaluation des experts de la CE, l'algorithme ne fait pas beaucoup mieux que le hasard (0,5) sur une version simplifiée du problème. Les auteurs ne s'inquiètent pas de la représentativité de ce faible échantillon ni des risques de reproduction des biais discriminatoires du recruteur humain.

5.4 Remarques

Il est remarquable qu'en France peu de regards critiques n'émergent sur le déploiement rapide de SIA pour l'aide automatique à l'embauche comparativement au volume des articles concernant les opérations de *testing* de la DARES. Seules des recensions des travaux de Cathy O'Neil (2016) ou des traductions de ses entretiens ne révèlent des remises en cause efficaces de ces procédés mais, comme précédemment, cela concerne principalement la situation aux USA. Il est pourtant facile de considérer que la situation est identique en France lorsque des entreprises comme *LinkedIn* sont concernées.

De toute évidence, rares sont les questions (qualité, transparence, non discrimination) présentes dans la liste d'évaluation du groupe d'experts européens auxquelles il serait possible d'apporter une réponse en considérant les présentations disponibles bien trop sommaires de ces algorithmes appliqués en France aux procédures d'embauche. Il est notable qu'aux USA, l'usage officiel de prise en compte de l'effet disproportionné a une influence réelle sur les pratiques. La leçon à en tirer est que l'adhésion d'une charte éthique, l'affichage de bonnes intentions ou des déclarations incantatoires sur l'objectivité des algorithmes, ne suffisent pas à changer les habitude ou protéger les usagers. Le déploiement d'un futur règlement européen encadrant la réalisation, la certification, l'exploitation des SIA est indispensable voire une urgence.

6 Conclusion

La détection d'une pratique discriminatoire est difficile. Cette difficulté est encore renforcée par la superposition d'une couche d'opacité lorsque la discrimination est la conséquence d'une décision ou aide à la décision algorithmique. Toutes les institutions et régulateurs s'accordent sur la nécessité de pouvoir auditer de tels systèmes que ce soit *ex ante* ou *ex post* (Castets-Renard 2020).

Une chose est à retenir de cette rapide présentation et des exemples qu'ils soient numériques ou d'analyse qualitative des produits sur internet. Sans documentation précise et exhaustive sur le processus qui a conduit à la mise en exploitation d'un SIA, du recueil des données à sa vérification, voire sa certification, un audit *ex post* est impossible. Les enquêtes classiques par *testing* sont hors-jeu et tester un système sur des données réalistes, nécessiterait une immersion complète dans la complexité du domaine d'application concerné. Le risque serait évidemment de ne tester que certains aspects du système, certaines situations ou types de données. La question principale reste donc la représentativité de ces tests par rapport à l'usage réel qui est fait du système. Cette question est *de facto* un préalable indispensable à la création d'un système d'IA : quelles données pour quel objectif ? Si elle n'a pas été posée explicitement et documentée *ex ante*, une analyse *ex post* ne peut conduire qu'à une remise en cause de la fiabilité du système, en termes de qualité de décision ou de biais, devant l'impossibilité d'en définir précisément le domaine d'usage. La mise en place de cette documentation *ex ante* sera la conséquence de l'exécution de la liste

d'évaluation du groupe des experts européens reprise par le livre blanc de la CE. Elle suit le même principe et la même logique que l'analyse d'impact relatif à la confidentialité des données (*privacy impact assessment*) et devrait être formalisée dans une réglementation à venir.

Des capacités et des compétences, à la fois techniques (statistique, apprentissage automatique) et juridiques de la part des régulateurs ou de sociétés ad hoc, sont indispensables pour auditer *a minima* une telle documentation. Il s'agira en tout premier lieu, de s'assurer que la vérification *ex ante* de conception d'un système d'IA a été mise en place très en amont dans un souci de contrôle exigeant de la qualité à toutes les étapes : représentativité statistique des données d'entraînement en fonction de l'objectif, procédure d'apprentissage et évaluation des erreurs, des biais, validation, éventuelle certification et mise en exploitation. Le cahier des charges doit également intégrer une surveillance du bon fonctionnement du SIA afin d'en contrôler tout risque de dérive et d'identifier les causes et responsabilités humaines en cas d'erreurs. Ce processus qualité peut imposer de devoir ré-entraîner périodiquement l'algorithme afin d'y intégrer des situations ou cas de figures initialement omis de la base de données. À titre d'exemple, dans le domaine de la santé aux USA, cette boucle de rétroaction est explicitement décrite dans les procédures de la FDA (*Food and Drug Administration*) pour la certification des systèmes de santé connectés (AI/ML-SaMD ou Artificial Intelligence and Machine Learning Software as a Medical Device).

Ces éléments conclusifs à un ensemble de recommandations listées en introduction (tableau 1).

Références

— Barocas S., Selbst A. (2016). Big Data's Disparate Impact, 104 *California Law Review*, 104 671.

— Besse P., Besse Patin A., Castets Renard C. (2019-a). Implications juridiques et éthiques des algorithmes d'intelligence artificielle dans le domaine de la santé, soumis, HAL preprint.

— Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2019-b). L'IA du Quotidien peut elle être Éthique? Loyauté des Algorithmes d'Apprentissage Automatique, *Statistique et Société*, **6**-3.

— Besse P., Castets Renard C., Loubes J.-M., Risser L. (2020). Évaluation des Risques des Algorithmes d'Apprentissage Statistique de l'IA: ressources pédagogiques, tutoriels R et python en ligne consultés le 8/05/2020.

— Brunet, M.-E., Alkalay-Houlihan C., Anderson A., Zemzl R. (2019). Understanding the Origins of Bias in Word Embeddings, arXiv preprint.

— Castets-Renard C. (2020). Comment construire une intelligence artificielle responsable et inclusive?, *Recueil Dalloz*, p 225.

— 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.*

— Ferguson A. (2017). *The Rise of Big Data Policing : Surveillance, Race,*

and the Future of Law Enforcement, NYU Press.
— Friedler S., Scheidegger C., Venkatasubramanian S., Choudhary S., Hamilton E., Roth D. (). Comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 329–38.
— Hemammou L., Wajntrob G., Martin J.-C., Clavel C. (2018). Entretien vidéo différé: modèle prédictif pour la pré-sélection de candidats sur la base du contenu verbal, *proceedings WACAI 2018*.
— Larson J., Mattu S., Kirchner L., Angwin J. (2016). How we analyzed the compas recidivism algorithm. ProPublica, en ligne consulté le 28/04/2020.
— Lee P, Le Saux M, Siegel R, et al (2019) Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review. *The American Journal of Emergency Medicine*, 37, p 1770–77.
— Myers West S., Whittaker M., Crawford K. (2019). Discriminating Systems : Gender, Race and Power, in AINowInstitute Report.
— O'Neil C. (2016). *Weapons of Math Destruction : How Big Data Increases Inequality*, Crown Random House.
— 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*.
— Riach P.A., Rich J. (2002). Field Experiments of Discrimination in the Market Place, *The Economic Journal*, Vol. 112 (483), p F480-F518.
— 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.
— Zliobaité I. (2017). Measuring discrimination in algorithmic decision making, *Data Min. Knowl. Disc.*, 31, p 1060–89.