Rapport

A Survey on Bias and Fairness in Machine Learning

Encadrant :
Prof. Hoel Le Capitaine

Présenté par :
Safae Hassouni

Équipe de Recherche :
DUKE, LS2N, Université de Nantes

Date : 6 mars 2025

Table des matières

1 Introduction **3**

2 Exemples d'injustice algorithmique **4**
2.1 COMPAS et justice prédictive 4
2.2 Publicités en ligne et recrutement 4
2.3 Reconnaissance faciale et discrimination 4

3 Catégorisation des Biais en Intelligence Artificielle **5**
3.1 Biais dans les données 5
3.1.1 Biais de mesure 5
3.1.2 Biais d'omission 5
3.1.3 Biais de représentation 5
3.1.4 Biais d'agrégation 6
3.1.5 Biais d'échantillonnage 6
3.2 Biais dans les algorithmes 6
3.2.1 Biais algorithmique 6
3.2.2 Amplification des biais 6
3.3 Biais liés aux interactions utilisateurs 6
3.3.1 Biais de popularité 6
3.3.2 Biais émergent 7

4 Définitions et Métriques de l'Équité Algorithmique **8**
4.1 Définitions de l'Équité Algorithmique 8
4.1.1 Parité Démographique 8
4.1.2 Égalité des Chances 8
4.1.3 Contre-factualité 9
4.2 Limites et Incompatibilités entre Certaines Définitions 9

5 Méthodes d'Atténuation des Biais en Intelligence Artificielle **10**
5.1 Pré-traitement : Correction des biais dans les données 10
5.2 Entraînement équitable : Modification des algorithmes 10
5.3 Post-traitement : Ajustement des résultats après l'apprentissage . . . 11
5.4 Outils et métriques d'évaluation de l'équité 11

6 Applications et Domaines d'IA Concernés – Fair NLP **12**
6.1 Word Embedding et Biais de Genre 12
6.1.1 Exemple de biais 12

6.1.2 Méthodes de correction 12
6.2 Biais en Coreference Resolution 12
6.2.1 Exemple de biais 13
6.2.2 Solutions proposées 13
6.3 Biais dans les Modèles de Langage 13
6.3.1 Exemple de biais 13
6.3.2 Méthodes de correction 13
6.4 Biais en Traduction Automatique 13
6.4.1 Exemple de biais 13
6.4.2 Solutions proposées 13
6.5 Biais en Reconnaissance d'Entités Nommées (NER) 14
6.5.1 Exemple de biais 14
6.5.2 Méthodes d'atténuation 14

Chapitre 1

Introduction

L'intelligence artificielle (IA) est de plus en plus intégrée dans notre quotidien, jouant un rôle crucial dans la prise de décisions importantes, telles que l'embauche, l'octroi de prêts, ou encore le système judiciaire. Son adoption massive soulève cependant des questions éthiques, notamment en ce qui concerne l'équité et la présence de biais algorithmiques.

Un algorithme est dit "injuste" lorsqu'il favorise ou défavorise un groupe d'individus en raison de caractéristiques inhérentes ou acquises. Ces biais peuvent provenir des données d'entraînement, des mécanismes de l'algorithme lui-même, ou encore des interactions avec les utilisateurs.

Chapitre 2

Exemples d'injustice algorithmique

Les biais algorithmiques ont conduit à des décisions discriminatoires dans plusieurs domaines :

2.1 COMPAS et justice prédictive

L'algorithme COMPAS, utilisé pour évaluer le risque de récidive aux États-Unis, attribuait des scores plus élevés aux Afro-Américains, entraînant des peines plus sévères par rapport aux détenus blancs à profil similaire.

2.2 Publicités en ligne et recrutement

Les algorithmes de publicité ont favorisé les hommes dans les annonces d'emplois STEM, car les femmes étaient jugées plus coûteuses à cibler, influençant ainsi les opportunités professionnelles.

2.3 Reconnaissance faciale et discrimination

Des systèmes de reconnaissance faciale ont mal identifié les personnes à la peau foncée et les Asiatiques, en raison d'ensembles de données biaisés, illustrant un manque de diversité dans l'entraînement des modèles.

Ces exemples montrent l'impact des biais dans l'IA et soulignent la nécessité d'algorithmes plus équitables et transparents.

Chapitre 3

Catégorisation des Biais en Intelligence Artificielle

Ce chapitre présente une classification des biais en IA, en les regroupant selon leur origine : biais dans les données, biais dans les algorithmes et biais liés aux interactions utilisateurs.

3.1 Biais dans les données

Les biais dans les données sont souvent introduits lors de la collecte, du traitement ou de l'échantillonnage des données. Ils peuvent fausser les performances des modèles d'apprentissage automatique.

3.1.1 Biais de mesure

Le biais de mesure survient lorsque les variables choisies pour représenter un phénomène sont imparfaites ou erronées. Par exemple, le système COMPAS d'évaluation des risques de récidive aux États-Unis utilise des données d'arrestation antérieure comme indicateur de criminalité, ce qui reflète davantage des pratiques policières biaisées que la dangerosité réelle des individus.

3.1.2 Biais d'omission

Ce biais est causé par l'absence de variables importantes dans un modèle. Par exemple, un modèle prédisant le taux de rétention des clients sans tenir compte de l'arrivée d'un concurrent proposant des prix plus bas risque d'aboutir à des conclusions erronées.

3.1.3 Biais de représentation

Ce biais se manifeste lorsque l'échantillon de données n'est pas représentatif de la population cible. Un exemple typique est le dataset ImageNet, largement biaisé vers les cultures occidentales en raison de son manque de diversité géographique.

3.1.4 Biais d'agrégation

Ce biais survient lorsqu'un modèle tire des conclusions sur des individus à partir d'observations faites sur une population globale. Dans le domaine médical, ignorer les différences biologiques entre groupes ethniques et de genres peut rendre les modèles inadaptés à certains patients.

3.1.5 Biais d'échantillonnage

Lorsque certaines sous-populations sont sous-représentées dans un jeu de données, le modèle formé sur ces données peut ne pas généraliser correctement à de nouveaux cas. Par exemple, une étude sur des patients basée sur un hôpital urbain pourrait ne pas être applicable aux populations rurales.

3.2 Biais dans les algorithmes

Les biais peuvent également être introduits au niveau des algorithmes, même si les données ne sont pas biaisées.

3.2.1 Biais algorithmique

Ce biais survient lorsque des décisions de conception introduisent une distorsion dans les prédictions. Par exemple, certains algorithmes de reconnaissance faciale ont montré une précision inférieure pour les personnes ayant une peau plus foncée, en raison d'un apprentissage biaisé.

3.2.2 Amplification des biais

Même si les données d'origine sont modérément biaisées, un modèle d'apprentissage automatique peut amplifier ces biais en surajustant certaines corrélations. Par exemple, un système de recommandation pourrait proposer principalement des contenus d'un certain groupe démographique, renforçant ainsi des stéréotypes existants.

3.3 Biais liés aux interactions utilisateurs

Les interactions entre utilisateurs et systèmes d'IA peuvent également créer ou renforcer des biais.

3.3.1 Biais de popularité

Ce biais se manifeste lorsque les contenus ou produits les plus populaires sont surrecommandés, au détriment des alternatives moins visibles. Par exemple, les plateformes de streaming favorisent souvent les artistes déjà bien établis, rendant plus difficile la découverte de nouveaux talents.

3.3.2 Biais émergent

Ce biais survient avec le temps à mesure que les utilisateurs interagissent avec le système. Par exemple, un moteur de recherche peut progressivement modifier son classement en fonction des clics des utilisateurs, renforçant des comportements et des choix spécifiques, parfois au détriment d'autres contenus plus pertinents.

Chapitre 4

Définitions et Métriques de l'Équité Algorithmique

L'équité algorithmique est un concept central dans le domaine du machine learning, particulièrement dans les applications impliquant des décisions automatisées ayant un impact social. Diverses définitions et métriques ont été proposées pour quantifier et garantir une prise de décision juste. Cependant, ces définitions ne sont pas toujours compatibles entre elles, et leur choix dépend souvent du contexte d'application.

4.1 Définitions de l'Équité Algorithmique

4.1.1 Parité Démographique

Aussi appelée "statistical parity" ou "demographic parity", cette métrique impose que la probabilité d'obtenir un résultat favorable soit identique entre les groupes protégés et non protégés. Formellement, un classificateur Y satisfait la parité démographique si :

$$P(Y = 1|A = 0) = P(Y = 1|A = 1)$$

où A est l'attribut sensible (par exemple, le genre ou l'origine ethnique). Cette définition garantit qu'aucun groupe ne soit systématiquement désavantagé, mais elle ne tient pas compte de différences de qualification entre les groupes.

4.1.2 Égalité des Chances

L'égalité des chances (Equal Opportunity) exige que le taux de vrais positifs soit identique pour les groupes protégés et non protégés. Un modèle respecte cette contrainte si :

$$P(Y = 1|A = 0, Y = 1) = P(Y = 1|A = 1, Y = 1)$$

Cela signifie que, parmi les individus qui devraient recevoir un résultat positif, la probabilité d'être correctement classé ne doit pas dépendre de l'appartenance

au groupe protégé. Cette métrique est particulièrement utile dans des contextes où l'accès à des opportunités est en jeu, comme les admissions universitaires ou le recrutement.

4.1.3 Contre-factualité

L'équité contre-factuelle (Counterfactual Fairness) stipule qu'une décision est équitable si elle resterait inchangée si l'individu appartenait à un autre groupe démographique tout en gardant les mêmes caractéristiques non sensibles. Formellement, cela signifie que :

$$P(Y_{A\leftarrow a}(U) = y|X = x, A = a) = P(Y_{A\leftarrow a'}(U) = y|X = x, A = a)$$

pour toutes les valeurs possibles de a et a'. Cette approche s'appuie sur des modèles causaux pour identifier les biais systémiques qui découlent de facteurs historiques ou institutionnels.

4.2 Limites et Incompatibilités entre Certaines Définitions

Les différentes définitions de l'équité ne sont pas toujours compatibles entre elles. Par exemple, il a été démontré qu'il est impossible de satisfaire simultanément la calibration (Test Fairness) et l'égalité des chances sauf dans des cas très spécifiques. En d'autres termes, tenter de corriger un type de biais peut en introduire un autre.

D'autres conflits existent entre la parité démographique et l'égalité des chances. Un modèle qui respecte strictement la parité démographique peut favoriser un groupe au détriment d'un autre en ignorant des différences pertinentes. À l'inverse, un modèle qui optimise uniquement l'égalité des chances peut ne pas garantir une représentation équitable de tous les groupes dans les résultats finaux.

Enfin, les approches causales, bien que conceptuellement séduisantes, nécessitent souvent des hypothèses fortes sur les relations causales entre les variables, ce qui peut être difficile à établir en pratique.

Chapitre 5

Méthodes d'Atténuation des Biais en Intelligence Artificielle

L'atténuation des biais en apprentissage automatique est un domaine clé pour garantir l'équité algorithmique. Les biais peuvent être réduits à différentes étapes du cycle de vie des modèles d'IA : avant l'entraînement (pré-traitement), pendant l'entraînement (in-processing), et après l'entraînement (post-traitement).

5.1 Pré-traitement : Correction des biais dans les données

Les méthodes de pré-traitement visent à modifier les données d'apprentissage avant l'entraînement du modèle pour corriger d'éventuels biais et assurer une meilleure représentativité des groupes. Ces méthodes incluent :

- **Re-échantillonnage des données** : Ajustement de la distribution des classes pour éviter une sous-représentation de certains groupes minoritaires.
- **Disparate Impact Remover** : Suppression ou transformation des attributs sensibles pour minimiser leur impact sur les prédictions.
- **Encodage équitable des caractéristiques** : Transformation des données en représentations latentes qui minimisent l'influence des attributs protégés.

Ces approches sont utilisées lorsque l'on a un accès complet aux données et que l'on souhaite s'assurer qu'aucune discrimination systémique ne soit intégrée dans le modèle d'apprentissage.

5.2 Entraînement équitable : Modification des algorithmes

L'approche in-processing vise à modifier les algorithmes d'apprentissage pour intégrer des contraintes d'équité directement dans le processus d'entraînement. Parmi les méthodes utilisées :

- **Pénalisation des biais dans la fonction de coût** : Ajout de termes dans la fonction de perte pour contraindre l'optimisation à réduire les disparités

entre groupes.

— **Apprentissage adversarial** : Utilisation de réseaux adversariaux pour empêcher le modèle de capturer des corrélations injustes entre les attributs sensibles et les décisions finales.

— **Méthodes de régularisation pour l'équité** : Contraindre le modèle à produire des prédictions similaires pour des individus de groupes différents mais présentant les mêmes caractéristiques non sensibles.

Ces méthodes nécessitent d'avoir un contrôle direct sur l'algorithme d'apprentissage et sont particulièrement adaptées lorsque l'accès aux données en amont est limité.

5.3 Post-traitement : Ajustement des résultats après l'apprentissage

Lorsque l'on ne peut pas modifier les données ni l'algorithme d'apprentissage, des techniques de post-traitement peuvent être appliquées sur les prédictions du modèle pour garantir l'équité. Ces approches incluent :

— **Recalibrage des scores de décision** : Ajustement des scores de sortie pour équilibrer les distributions entre groupes protégés et non protégés.

— **Changement de seuils de classification** : Modification des seuils de décision pour s'assurer qu'un groupe donné ne soit pas défavorisé par rapport à un autre.

— **Méthodes de réassignation** : Réattribution de certaines prédictions pour corriger d'éventuelles inégalités observées après entraînement.

Ces approches sont particulièrement adaptées aux contextes où l'accès aux données ou au modèle est restreint, mais où l'on peut contrôler la prise de décision finale.

5.4 Outils et métriques d'évaluation de l'équité

Plusieurs outils ont été développés pour évaluer et atténuer les biais dans les modèles d'intelligence artificielle. Parmi les plus connus :

— **Aequitas** : Outil open-source permettant d'analyser la disparité et les biais dans les prédictions des modèles d'apprentissage automatique.

— **AI Fairness 360 (AIF360)** : Bibliothèque développée par IBM contenant des implémentations de diverses métriques d'équité et de méthodes d'atténuation des biais.

— **Fairlearn** : Outil de Microsoft permettant d'évaluer et d'améliorer l'équité des modèles prédictifs en ajustant leurs décisions.

Ces outils fournissent des métriques d'équité telles que l'égalité des chances, la parité démographique et la calibration des scores, permettant d'identifier et de corriger les biais de manière systématique.

Chapitre 6

Applications et Domaines d'IA Concernés – Fair NLP

Ce chapitre explore différentes manifestations des biais dans le NLP et les solutions proposées pour atténuer ces problèmes.

6.1 Word Embedding et Biais de Genre

Les word embeddings sont des représentations vectorielles des mots utilisées par les modèles de NLP pour comprendre le langage. Cependant, ils peuvent capturer et amplifier des stéréotypes de genre.

6.1.1 Exemple de biais

Lors de tests d'analogie, des modèles d'embedding ont montré que "homme" était souvent associé à "programmeur informatique", tandis que "femme" était associé à "ménagère".

6.1.2 Méthodes de correction

Pour atténuer ces biais, plusieurs approches ont été développées :

— **Hard debiasing** : suppression des biais des mots neutres tout en conservant ceux des mots genrés.

— **Soft debiasing** : ajustement progressif des vecteurs de mots pour minimiser les biais tout en préservant la structure d'origine.

6.2 Biais en Coreference Resolution

La résolution de coréférence consiste à identifier les entités auxquelles se réfèrent les pronoms dans un texte. Des études ont montré que ces systèmes ont une tendance à assigner certains rôles en fonction du genre.

6.2.1 Exemple de biais

Un système pourrait associer “docteur” à un homme et “infirmière” à une femme, influençant ainsi les interprétations des textes.

6.2.2 Solutions proposées

— **WinoBias** : un benchmark conçu pour mesurer et corriger les biais dans les systèmes de coréférence.
— **Augmentation des données** : échange des rôles genrés dans les jeux de données pour équilibrer l'entraînement des modèles.

6.3 Biais dans les Modèles de Langage

Les modèles de génération de texte peuvent reproduire et amplifier les biais présents dans leurs corpus d'entraînement.

6.3.1 Exemple de biais

Un modèle entraîné sur des textes contenant des stéréotypes pourrait générer des phrases biaisées reflétant ces tendances.

6.3.2 Méthodes de correction

— Ajout d'une régularisation pour minimiser l'impact des biais dans les représentations vectorielles.
— Métriques spécifiques pour quantifier les biais dans les résultats des modèles.

6.4 Biais en Traduction Automatique

Les systèmes de traduction automatique peuvent introduire des biais en raison des associations genrées dans les corpus de formation.

6.4.1 Exemple de biais

Anglais → Français :
— “My friend is a doctor” → “Mon ami est médecin” (masculin)
— “My friend is a nurse” → “Mon amie est infirmière” (féminin)
Ce biais est dû aux associations statistiques entre certaines professions et les genres dans les jeux de données.

6.4.2 Solutions proposées

— Ajout de balises de genre dans les phrases sources pour guider la traduction de manière plus neutre.
— Utilisation d'embeddings débiaisés pour limiter l'influence des stéréotypes.

6.5 Biais en Reconnaissance d'Entités Nommées (NER)

La reconnaissance d'entités nommées (NER) consiste à identifier les noms propres (personnes, lieux, organisations) dans un texte. Les modèles de NER ont montré une tendance à mal classer ou à ignorer les noms féminins.

6.5.1 Exemple de biais

Certains modèles identifient moins fréquemment les noms féminins comme des entités correctes, les confondant avec d'autres catégories (ex. lieux, objets).

6.5.2 Méthodes d'atténuation

— Création de jeux de données équilibrés incluant une représentation égale des genres.
— Développement de nouvelles métriques d'évaluation pour mieux identifier les erreurs liées au genre.