HEC MONTRÉAL

L'équité dans les systèmes d'intelligence artificielle pour la prise de décision : atténuer les biais algorithmiques genrés et raciaux en apprentissage automatique

par

Balkissa Touré

Denis Larocque
HEC Montréal
Codirecteur de recherche

Joé T. Martineau
HEC Montréal
Codirectrice de recherche

Sciences de la gestion
(Spécialisation Intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Août 2025

Résumé

L'intelligence artificielle s'ancre de plus en plus dans nos vies et quotidiens. De l'IA générative aux agents intelligents, ces systèmes occupent désormais une place presque familière. Toutefois, leur rôle outrepasse désormais ces usages et on les retrouve de plus en plus dans des domaines sensibles, où leurs décisions influencent directement les trajectoires de vie. Ce constat nous a conduits à nous poser une question centrale : *les systèmes d'intelligence artificielle sont-ils réellement équitables ?*

Derrière l'image d'efficacité et de neutralité se cache une réalité plus contrastée, où les algorithmes peuvent reproduire ou accentuer les inégalités existantes. Ainsi, ce travail s'inscrit dans le champ de l'intelligence artificielle responsable et poursuit deux objectifs principaux. D'une part, nous cherchons à démystifier la notion de biais algorithmique. Et ce, en interrogeant leurs origines et leurs causes profondes, tout en illustrant leurs manifestations concrètes à partir d'exemples observés dans la réalité. Car tout effort de remédiation exige d'abord une compréhension approfondie de la complexité du problème.

D'autre part, ce travail adopte une approche empirique en mettant à l'épreuve différentes méthodes de mitigation sur des modèles d'apprentissage automatique dans un scénario simulé de recrutement, afin d'évaluer leur efficacité. L'étude s'interroge aussi sur la solidité de ces méthodes lorsqu'elles sont confrontées à des situations plus complexes en présence de biais explicites et implicites.

L'évaluation indique qu’aucune méthode ne supprime entièrement les biais. Toutes contribuent néanmoins à atténuer les écarts entre groupes protégés et non protégés, avec une efficacité qui fluctue selon les contextes et les critères d’évaluation. Ces constats renforcent l’idée que la réduction des biais n’est pas seulement possible mais constitue une étape indispensable dans la construction de systèmes de décision automatisés plus équitables et responsables.

Mots clés : Équité algorithmique, Biais algorithmiques, Métriques d’équité, Mitigation des biais, Intelligence artificielle responsable, Apprentissage automatique, Biais genrés, Biais raciaux, Recrutement automatisé

Méthodes de recherche : Expérimentation, Exploitation de données, Intelligence artificielle, Empirique, Quantitative et Comparative

Abstract

Artificial intelligence is becoming increasingly embedded in our daily lives. From generative AI to intelligent agents, these systems now occupy an almost familiar place. Yet their role extends beyond these common uses and increasingly reaches sensitive domains, where decisions directly shape individual life trajectories. This raises a central question: *are artificial intelligence systems truly fair?*

Behind the image of efficiency and neutrality lies a more complex reality, in which algorithms can reproduce or even amplify existing inequalities. This work therefore falls within the field of responsible artificial intelligence and pursues two main objectives. First, it seeks to demystify the notion of algorithmic bias by examining its origins and underlying causes, while illustrating its concrete manifestations through real-world examples. Indeed, any remediation effort requires a prior and thorough understanding of the complexity of the problem.

Second, the study adopts an empirical approach by testing various bias mitigation methods on machine learning models in a simulated recruitment scenario, in order to assess their effectiveness. It also examines the robustness of these methods when confronted with more complex situations involving both explicit and implicit biases.

Our evaluation shows that no method completely eliminates bias. However, all contribute to reducing disparities between protected and non-protected groups, with effectiveness fluctuating depending on the context and the evaluation criteria. These

findings underscore that reducing bias is not only possible but essential to the construction of automated decision-making systems that are fairer and more responsible.

Keywords: Algorithmic fairness, Algorithmic bias, Fairness metrics, Bias mitigation, Responsible AI, Machine learning, Gender bias, Racial bias, Automated recruitment

Research Methods: Experimentation, Data Analysis, Artificial Intelligence, Empirical, Quantitative, and Comparative

Table des matières

Liste des Tableaux

Liste des figures

Liste des abréviations

IA Intelligence Artificielle

AIF360 AI Fairness 360 (bibliothèque IBM de mitigation des biais)

SPD Statistical Parity Difference (différence de parité statistique)

EOD Equal Opportunity Difference (différence d'égalité des chances)

AOD Average Odds Difference (différence moyenne des chances**)**

ERD Error Rate Difference (différence des taux d'erreur)

DI Disparate Impact (impact disparate)

RQ Question de recherche

Remerciements

Je voudrais prendre ces quelques lignes pour exprimer ma plus profonde gratitude à toutes les personnes qui m'ont soutenue durant ce projet d'envergure et rendu possible la réalisation de ce mémoire.

En premier lieu, je tiens à remercier très sincèrement mes deux directeurs de recherche, Joé T. Martineau et Denis Larocque. Merci pour votre soutien indéfectible, votre disponibilité et vos conseils tout au long de ce projet. Je suis extrêmement reconnaissante d'avoir été mentorée par vous et je n'aurais pu rêver de meilleurs directeurs et mentors. Votre expertise, votre confiance et votre appui ont fait toute la différence.

Il m'est impossible de poursuivre ces lignes sans remercier mes parents, sans qui rien de tout cela n'aurait été possible. Vous êtes la source de mon parcours et de mes réussites. Je vous suis infiniment reconnaissante pour votre soutien indéfectible, vos conseils, vos sacrifices et l'accompagnement que vous m'avez offert tout au long de mon cheminement scolaire et de ma vie. Ce mémoire vous est dédié, car sans vous, il n'aurait jamais vu le jour.

À mon frère et à ma sœur, merci du fond du cœur. Ces quelques lignes ne suffiront jamais à exprimer toute ma gratitude.

À toute ma famille, vous êtes et restez ma source de motivation et de persévérance.

Je souhaiterais également remercier mes amis pour leur soutien inconditionnel, mais aussi de partager ma passion pour ce projet et de m'écouter en parler pendant des heures (ce qui, je l'avoue, est arrivé bien plus souvent que prévu).

Je suis reconnaissante envers toutes les personnes chères à mon cœur qui, par leur présence, leur soutien moral et leur confiance, ont rendu cette aventure possible. Merci d'avoir cru en moi lorsque moi-même je doutais, et d'avoir partagé avec moi toutes les étapes de ce projet.

C'est avec beaucoup de fierté, de joie et une certaine nostalgie que je dépose ce mémoire, que j'ai eu le bonheur de réaliser durant ces derniers mois. J'espère que vous prendrez autant de plaisir à le lire que j'en ai eu à l'écrire.

Chapitre 1 : Introduction

***« L'intelligence artificielle est la nouvelle électricité. »* — Andrew Ng (2017).**

Comme l'électricité a transformé la société il y a un siècle, la révolution technologique actuelle bouleverse nos modes de vie et redéfinit les bases de notre monde contemporain. Chaque révolution industrielle successive a profondément transformé les dynamiques du travail et de la société. Aujourd'hui, la quatrième révolution industrielle, portée par des technologies disruptives, poursuit cette transformation à une échelle inédite (Zhang et Chen, 2023). Cette révolution repose sur une synergie entre la démocratisation des technologies avancées et l'exploitation massive de volumes considérables de données (*big data*) (Chen 2023). Jadis confinées à des cercles restreints, ces technologies s'imposent désormais comme des moteurs de transformation à l'échelle mondiale. Les chiffres témoignent de cette ascension. En 2023, le marché mondial des technologies de l'information a atteint 8 852 milliards de dollars, enregistrant une croissance annuelle composée de 8,2 %. Selon les prévisions, ce marché devrait dépasser 11 995 milliards de dollars d'ici 2027 (ReportLinker 2023).

Au cœur de cette transformation, un acteur technologique occupe une place centrale : l'intelligence artificielle. De simples promesses techniques, elle s'impose aujourd'hui comme un catalyseur majeur de changement. En quelques années, elle est devenue un levier influençant non seulement la manière dont nous travaillons, mais aussi celle dont nous prenons des décisions. En 2023, son marché mondial était estimé à 193,63 milliards de dollars, avec des projections indiquant une croissance annuelle moyenne de 36,6 %

entre 2024 et 2030 (Grand View Research, 2024). Cette adoption s'accélère également au Canada, où 6,1 % des entreprises déclarent utiliser l'IA pour produire des biens ou fournir des services. Ces taux atteignent jusqu'à 20,9 % dans les industries de l'information et de la culture (S. C. Gouvernement du Canada 2024).

Au-delà de sa croissance économique fulgurante, l'intelligence artificielle ne se limite plus aux laboratoires de recherche et s'immisce dans notre quotidien. On la retrouve dans les voix familières de Siri ou d'Alexa, dans les recommandations que nous propose Netflix ou Spotify, ou encore dans les véhicules qui se conduisent (presque) tout seuls. Mais son influence ne s'arrête pas là. Dans le monde professionnel, elle est également perçue comme un levier d'efficacité considérable. D'après le Gartner CIO Survey (2019), environ 37 % des organisations mondiales affirment utiliser l'IA marquant une augmentation spectaculaire de 270 % en seulement quatre ans. Les algorithmes permettent d'automatiser des tâches répétitives, générant ainsi des économies de temps et des gains financiers significatifs (Gonzalez et al., 2019). Dans le même esprit, au Canada, les entreprises identifient les technologies émergentes et l'innovation comme des facteurs clés de succès , une opinion partagée par 28,3 % d'entre elles (Gouvernement du Canada, 2024)

Toutefois, cette révolution technologique n'est pas sans risques. Si l'intelligence artificielle promet des avancées spectaculaires dans de nombreux domaines, elle soulève également des préoccupations éthiques majeures. Lorsqu'elle est utilisée pour concevoir des systèmes décisionnels, comme dans le recrutement, elle peut amplifier des inégalités existantes et perpétuer des biais profondément ancrés (Chen, 2023).

Pourtant, ces biais ne sont pas nés avec les technologies modernes. Ce sont des problématiques qui trouvent leurs racines dans des comportements et des décisions humaines, façonnés par des préjugés conscients et inconscients. Ces discriminations structurelles ont été largement documentées dans la littérature. Par exemple une étude pionnière a montré que les candidats ayant des noms à consonance européenne recevaient 50 % plus d'appels pour des entretiens que ceux ayant des noms à consonance afro-américaine, malgré des qualifications (Bertrand et Mullainathan, 2004). Une autre étude révèle que les noms clairement associés aux Afro-Américains réduisent de 2,1 points la probabilité de contact de la part des employeurs par rapport aux noms clairement associés aux blancs. Cela correspond à une diminution équivalente à 9 % du taux moyen de contact pour les Afro-Américains (Kline et al., 2021).

Cependant, l'essor des systèmes d'intelligence artificielle confère une dimension nouvelle à ces biais et enjeux éthiques. En s'appuyant sur des données historiques déjà marquées par des déséquilibres, ces systèmes tendent à répliquer, voire accentuer, les discriminations sociales. Ils peuvent alors devenir des vecteurs de stéréotypes et biais, notamment de biais genrés et raciaux (Ferrara 2024). Ces enjeux touchent des principes clés comme l'équité et l'éthique. Ils soulèvent des questions sur la manière dont les systèmes automatisés influencent profondément la vie des individus, tout en suscitant des inquiétudes quant à leur risque d'institutionnaliser des inégalités préexistantes. Un exemple largement reconnu des dérives de ces systèmes de décisions automatisées se trouve dans le domaine judiciaire avec le logiciel COMPAS. Ce logiciel a été utilisé dans le système judiciaire américain pour prédire la probabilité de récidive d'un prévenu. Des enquêtes ont mis en lumière un traitement défavorable à l'égard des personnes afro-

américaines, fréquemment jugées à haut risque malgré l'absence de condamnations antérieures. Ce cas, loin d'être isolé, reflète une tendance plus large observée dans plusieurs états (comme le Wisconsin) où les biais algorithmiques renforcent les inégalités qu'ils étaient censés atténuer (Ferrara 2024). Cet exemple parmi tant d'autres illustre comment ces outils peuvent être biaisés contre certains groupes, augmentant les disparités au lieu de les réduire (Mehrabi et al., 2022).

Ces biais ne se limitent pas au seul domaine de la justice. Ils sont non seulement omniprésents, mais infiltrent divers secteurs et impactent profondément les dynamiques professionnelles, sociales et institutionnelles. En effet, partout où l'intelligence artificielle est mobilisée pour prendre des décisions automatisées (que ce soit dans la santé, l'éducation, les ressources humaines ou les services financiers etc.), des risques similaires émergent. Dès lors qu'un algorithme prédit ou recommande, il peut reproduire des inégalités, voire générer de nouvelles formes de discrimination. Le déploiement de ces biais dans des systèmes publics peut entraîner des conséquences graves, telles que le refus d'accès à des services essentiels, l'exclusion d'opportunités professionnelles, ou encore des arrestations et condamnations injustifiées. Ces dérives soulignent l'urgence d'une réflexion éthique et d'une régulation adaptée pour prévenir l'aggravation des discriminations systémiques et garantir une utilisation équitable et responsable de l'IA (Ferrara, 2024).

Face à ces défis, une prise de conscience émerge autour du concept d'IA responsable dans différents domaines (Kleinberg et al., 2016, Chen, 2023, Mehrabi et al., 2022, Ferrara, 2024, Hamida et al., 2024 etc.). Chercheurs, organisations et régulateurs

travaillent activement à identifier et à corriger les biais algorithmiques. À titre d'illustration, des outils comme l'IBM AIF 360 Toolkit (Bellamy et al., 2018) permettent d'évaluer et d'atténuer ces biais. Par ailleurs, de nombreuses initiatives réglementaires et des guides pratiques cherchent à encadrer l'utilisation éthique de ces technologies, notamment au Canada. Ces cadres sont encore en cours d'élaboration et leur mise en œuvre concrète dans les organisations soulève encore de nombreux défis.

Néanmoins, les questions d'équité et de biais en intelligence artificielle restent toujours moins discutées que leurs performances techniques. Le manque de transparence des modèles d'intelligence artificielle développés par les entreprises complique leur évaluation et rend difficile le développement sécurisé d'applications basées sur ces outils (Bommasani et al., 2023). La majorité des entreprises technologiques publient peu, voire aucune information sur les données utilisées, la conception et l'évaluation éthiques de leurs systèmes ainsi que sur leurs effets différenciés (Bommasani et al., 2023). De plus, l'insuffisance de la réglementation due à la rapidité du développement de l'IA et à sa nature transfrontalière, pose des défis majeurs (Walter 2024). Dans ce contexte, il est donc essentiel d'étudier et comprendre ces biais pour les atténuer, pendant que ces outils s'intègrent de plus en plus à nos quotidiens et prennent des décisions.

Cette recherche s'inscrit dans une démarche visant à explorer le concept d'intelligence artificielle responsable. Elle cherche à approfondir la compréhension des biais et des problématiques éthiques associées aux systèmes d'intelligence artificielle. Une attention particulière est accordée à leur impact dans des domaines sensibles comme le recrutement et les décisions automatisées. L'objectif est de réaliser une revue systématique des biais genrés et raciaux présents dans ces systèmes, en identifiant leurs origines et leurs causes

sous-jacentes. Nous analyserons également les principaux cadres conceptuels et métriques mobilisés pour évaluer l'équité algorithmique. Ensuite, notre attention se portera sur les solutions proposées dans la littérature, notamment les méthodes de mitigation des biais, afin d'évaluer leur pertinence et leur efficacité. L'approche adoptée se veut pragmatique, en dépassant le cadre purement théorique pour inclure une dimension expérimentale. Pour ce faire, nous simulerons un cadre de recrutement à l'aide de données test, afin d'examiner et de critiquer les méthodes existantes visant à réduire ces biais. Cette simulation permet d'étudier leur efficacité dans des conditions contrôlées, mais proches de leur application réelle, en particulier lorsque des biais explicites ou implicites (via des variables proxy) sont présents. Enfin, cette recherche a pour ambition de formuler des recommandations concrètes pour contribuer au développement de systèmes d'IA plus équitables et responsables.

Nous cherchons à fournir des réponses aux questions principales suivantes :

- Dans un contexte simulé de recrutement automatisé, dans quelle mesure les différentes méthodes de mitigation des biais algorithmiques (pré-traitement, in-processing, post-traitement) permettent-elles de réduire les inégalités de traitement entre groupes protégés et non protégés et de restaurer l'équité algorithmique ?
- Lorsque des biais indirects sont introduits via des variables proxies corrélées à un attribut sensible, les méthodes de mitigation conservent-elles leur efficacité ou voient-elles leurs gains d'équité se dégrader ?

- En présence de biais explicites et implicites, comment évoluent les métriques d'équité et de performance avant et après application de ces méthodes de mitigation ?

Les contributions de cette recherche sont multiples. Elle se donne pour ambition de combler plusieurs lacunes identifiées dans la littérature sur les biais des systèmes automatisés d'IA. Cette étude se distingue par une approche hybride, combinant une analyse théorique approfondie et une validation empirique.

Alors que l'intérêt pour les systèmes d'IA s'est intensifié avec leur démocratisation récente, la recherche sur leurs biais demeure fragmentée. La majorité des travaux se concentre sur les biais présents dans les données, au détriment d'autres étapes clés du cycle de vie des modèles, telles que la conception algorithmique ou les pratiques de développement (Mavrogiorgos et al., 2024). Pour combler cette lacune, notre étude adopte une approche holistique, en examinant l'ensemble du pipeline de développement et en envisageant les biais comme un phénomène systémique. Nous explorerons l'ensemble du processus de modélisation, depuis la compréhension des causes des biais jusqu'à l'expérimentation concrète de différentes méthodes de mitigation dans l'algorithme de prédiction.

De surcroît, les travaux sur l'équité algorithmique sont encore rares en langues non anglaises et restent marginaux, ce qui freine la diversification des perspectives culturelles sur le sujet (Ramesh et al., 2023). Ce mémoire vise donc à contribuer à combler cette lacune en apportant une perspective francophone à ces enjeux de biais algorithmiques.

Un autre enjeu critique réside dans le manque d'évaluations systématiques des outils d'automatisation dans des domaines tels que le recrutement. Contrairement aux méthodes traditionnelles comme les entretiens ou les tests cognitifs, qui sont validés par des protocoles bien établis, les systèmes d'IA sont souvent déployés sans validation scientifique approfondie (Hunkenschroer et Luetge, 2022). Cette recherche expérimente des techniques et méthodologies proposées pour évaluer ces outils dans un cadre simulé, offrant ainsi une perspective renouvelée sur les questions éthiques et les risques de biais dans ces systèmes. Cette approche permet de mesurer leur pertinence et leur efficacité dans des conditions proches de leur application réelle.

En outre, nous proposons une évaluation conceptuelle et empirique des biais dans les systèmes d'IA, tout en favorisant une réflexion interdisciplinaire. Elle mobilise des perspectives sociologiques, psychologiques et éthiques pour explorer les causes profondes des biais et leurs implications sociétales. En parallèle, elle examine des approches de mitigation tant techniques que non techniques, en s'intéressant non seulement aux solutions algorithmiques mais aussi aux cadres conceptuels permettant de mieux comprendre et atténuer ces biais.

Enfin, cette recherche propose des recommandations pratiques destinées aux entreprises et décideurs. En réduisant les biais, ces systèmes contribueront à renforcer la diversité organisationnelle tout en promouvant des pratiques responsables dès leur conception.

Afin de répondre à ces questions, ce projet s'articule autour de six grandes sections. Premièrement, l'introduction constitue l'amorce de l'étude et présente notre projet de recherche. Le deuxième chapitre est consacré à la revue de littérature. Il recense les

études antérieures et explore les biais genrés et raciaux, en s'appuyant sur des exemples concrets de discriminations observées dans les systèmes d'intelligence artificielle. Il en analyse les causes et les sources, puis aborde la notion d'équité dans ces systèmes, ainsi que les différentes méthodes de mitigation, qu'elles soient techniques ou non techniques. La troisième section détaille la méthodologie adoptée et les données mobilisées pour mener à bien le projet. La discussion qui suit examine ces résultats empiriques et la manière dont ils éclairent nos questions de recherche. Elle permet également de dégager les principaux enseignements de l'étude et d'établir une comparaison avec la littérature existante. Cette section met également en évidence les limites de notre travail. Enfin, la dernière section vise à conclure le mémoire en mettant en perspective les apports et les limites de cette recherche.

Chapitre 2 : Revue de la littérature

L'irruption des technologies d'intelligence artificielle dans nos vies transforme la manière dont nous travaillons et interagissons. Si cette ascension fulgurante ouvre un champ infini de possibilités, son usage croissant soulève également des enjeux éthiques fondamentaux. Ce chapitre poursuit trois finalités principales : d'une part, il illustre à travers des exemples concrets comment les systèmes d'intelligence artificielle, souvent considérés comme neutres et performants, peuvent engendrer des injustices sociales et perpétuer des biais et stéréotypes dans la vie quotidienne. Ensuite, il vise à approfondir les notions de biais, d'équité et d'éthique dans ces systèmes, en mettant en lumière leurs sources, leurs causes et les différents types de biais pouvant survenir, notamment dans le domaine du recrutement. Enfin, il explore les diverses solutions et méthodes disponibles pour atténuer ces biais et promouvoir des systèmes d'IA plus responsables.
Dans cette perspective, ce chapitre analyse les études existantes afin de mieux comprendre les biais de genre et de race dans les systèmes d'IA en général, et d'extrapoler ces analyses au domaine du recrutement. Les éléments présentés serviront de référence historique à notre recherche pour clarifier comment ces biais émergent et d'examiner les solutions pour les atténuer.

2.1. L'intelligence artificielle est-elle raciste et sexiste ? Études de cas concrets et leurs impacts

Les systèmes d'intelligence artificielle, souvent vantés pour leur objectivité et leur capacité à traiter d'immenses volumes d'informations, devraient en théorie réduire les biais humains. Pourtant, ils peuvent facilement perpétuer et amplifier ces biais,

reproduisant ainsi des discriminations, notamment liés au genre et à la race. Ce paradoxe soulève une question clé : l'intelligence artificielle est-elle vraiment impartiale ?
Cette section explore des cas concrets où les systèmes d'IA ont échoué à maintenir l'équité ainsi aggravant des inégalités existantes. L'objectif est de montrer que loin de relever de simples théories, ces problématiques impactent directement la vie de nombreuses personnes et communautés, soulignant l'urgence de concevoir des systèmes d'IA plus équitable.

2.1.1 Discrimination dans les systèmes de reconnaissance faciale

Nous utilisons quotidiennement les algorithmes de reconnaissance faciale : pour déverrouiller nos téléphones, surveiller les espaces publics ou même prendre des décisions cruciales en matière de sécurité. Cependant, ces technologies présentent des défauts notables, notamment des biais de genre et de race. Le projet « *Gender Shades*» (Buolamwini et Gebru, 2018) a réalisé une analyse critique des systèmes de reconnaissance faciale commerciaux de géants comme Microsoft, IBM et Face++. Ils ont évalué les biais présents dans les algorithmes d'analyse faciale automatisée et les ensembles de données en fonction des sous-groupes phénotypiques. Ensuite, ils ont mesuré les taux de précision pour chaque sous-groupe (hommes à la peau claire, hommes à la peau foncée, femmes à la peau claire, femmes à la peau foncée). Les résultats ont révélé des disparités significatives dans les taux d'erreur selon le genre et le ton de peau. Comme illustré dans la Figure 1, le taux d'erreur maximum parmi les trois systèmes de reconnaissance faciale (Microsoft, Face++, IBM) pour les femmes à la peau foncée atteignaient 34,7 %, contre seulement 0,8 % pour les hommes à la peau claire.

Ces biais ne sont pas de simples défauts techniques, ils ont des répercussions réelles et graves. Être mal identifié par un système de surveillance à cause de son origine peut entraîner des arrestations injustifiées et des erreurs d'identification, affectant principalement les minorités. Par exemple, depuis la levée de l'interdiction de la reconnaissance faciale dans la justice en 2022 à la Nouvelle-Orléans, cette technologie a conduit à plusieurs fausses identifications, dont des affaires de meurtre et d'agression. Sur 15 demandes, 9 n'ont donné aucun résultat et 3 des 6 correspondances étaient erronées, ciblant principalement les Afro-Américains (Ng, 2023). Ces incidents ne sont que des exemples parmi tant d'autres, montrant que ces technologies d'intelligence artificielle, loin d'être infaillibles, peuvent perpétuer les inégalités. Les conséquences de ces biais techniques peuvent être dramatiques dans la vie réelle.

Être systématiquement mal identifié en raison de la couleur de sa peau ou de son genre affecte directement la dignité et les droits des individus (Commission ontarienne des droits de la personne, 2024). Il est donc essentiel de reconnaître et de corriger ces biais pour éviter que ces outils ne deviennent des outils de discrimination systémique.

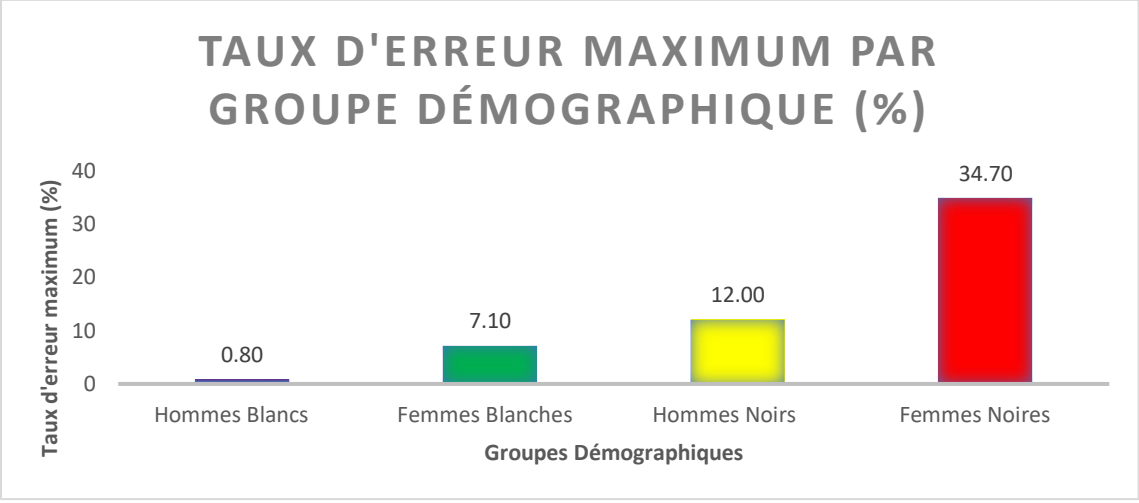

Figure 1: Taux d'erreur maximum de reconnaissance faciale par groupe démographique (figure basée sur les données de Buolamwini et Gebru, 2018, p. 9)

2.1.2 Les systèmes d'IA générative

Au cours des dernières années, l'intelligence artificielle générative (GenAI) a connu une croissance spectaculaire grâce à des modèles tels que GPT, DALL-E, Gemini, Claude et bien d'autres. Si ces avancées technologiques sont perçues comme des outils capables de produire du contenu innovant de manière objective, elles ne sont pas exemptes de défauts. En effet, elles perpétuent et amplifient souvent des stéréotypes sociaux, raciaux et genrés présent dans les données sur lesquelles elles s'appuient.

D'ici quelques années, 90 % du contenu généré sur internet pourrait provenir de l'intelligence artificielle (Nicoletti et Bass, 2023). Alors que des grandes entreprises de divers domaines adoptent ces outils pour générer du contenu nouveau à partir de grandes quantités de données, que ce soit pour l'art, la traduction, la rédaction automatique ou même le code, il est crucial de se demander si cette croissance ne légitime pas des préjugés. Cela nous emmène à réfléchir *: si les machines apprennent du passé, que se passe-t-il quand ce passé est truffé d'injustices ?*

Une analyse par Nicoletti et Bass (2023) incluant plus de 5000 images de personnes générées par l'IA a révélé des biais importants. Ils ont classé ces images selon l'échelle de Fitzpatrick[1] et leurs résultats montrent que les systèmes d'IA tendent à générer des

[1] L'échelle de phototypes de Fitzpatrick a été développée dans les années 1970 pour classer la manière dont les différentes couleurs de peau réagissent aux rayons UV, afin d'aider à prédire le risque de coup de soleil et de cancer de la peau. Elle constitue en soi une façon limitée de penser la carnation, mais elle reste aujourd'hui la norme utilisée par les dermatologues et par les chercheurs travaillant sur les biais en IA (Nicoletti et Bass, 2023).

images de personnes à la peau claire pour des professions valorisées comme « CEO » ou « scientifique », tandis que les représentations pour des emplois moins rémunérés, comme « travailleur de fast-food », incluent majoritairement des sujets à la peau foncée. Les stéréotypes de genre sont également reproduits. Des métiers comme « infirmier » (*nurse*) ou « enseignant » (*teacher*) génèrent principalement des images de femmes, alors que « ingénieur » et « développeur » sont associés aux hommes. Cela peut renforcer l'impression que certains groupes sont naturellement exclus de certaines sphères, impactant la perception de chacun quant à sa place dans la société. En effet, comme le souligne Heater Hiles, présidente de Black Girls Code, « on apprend, en ne se voyant pas représenté, que l'on n'a pas sa place » (Nicoletti et Bass, 2023). Cette absence de représentation illustre une forme subtile mais importante d'exclusion sociale et professionnelle chez les individus sous-représentés.

Ces technologies bien que s'appuyant sur des données massives reflètent simplement souvent les biais présents dans notre société, amplifiant parfois les disparités raciales et de genre. La technologie qui pourrait sembler impartiale premier abord, mais cette apparence d'objectivité est souvent trompeuse. Elle renforce l'idée que ces préjugés sont légitimes et inévitables, masquant ainsi leur caractère problématique. Or, ces biais sont loin d'être accidentels mais le reflet des données historiques biaisées utilisées pour entraîner ces modèles. En fin de compte, loin de corriger les disparités, les systèmes d'IA risquent de les institutionnaliser, exacerbant ainsi les inégalités sociales et professionnelles.

« Chaque étape d'un processus où un humain peut être biaisé, l'IA peut l'être aussi. »

Nicole Napolitano, Center for Policing Equity (Nicoletti et Bass, 2023).

2.1.3 Cas de biais dans le recrutement : Amazon et son algorithme de sélection biaisé

Les biais dans les processus de recrutement, qu'ils soient d'origine humaine ou algorithmique, entraînent des conséquences négatives à plusieurs niveaux (Chen, 2023). Ces biais peuvent affecter non seulement les individus, les organisations mais aussi la société. Prenons l'exemple d'Amazon. En 2014, le géant de la technologie a voulu améliorer son processus de recrutement en intégrant l'intelligence artificielle. L'objectif était de créer un algorithme capable de scanner les CV, d'identifier les candidats les plus qualifiés et de les noter d'une à cinq étoiles en fonction de leur qualification pour le poste.

Cependant, le système entraîné sur des CV majoritairement masculins des dix années précédentes, a pénalisé les candidats venant de collèges féminins ainsi que ceux dont les CV contenaient des termes tels que « féminin » (*women's*), comme dans « club d'échecs féminin » (*women's chess club*). Le système, présenté comme une avancée technologique, s'est révélé profondément biaisé. Amazon a conclu que la source du problème résidait dans les données d'entraînement biaisées, ce qui l'a contraint à retirer l'outil (Dastin, 2018).

En 2015, une autre tentative d'innovation a révélé un problème similaire. Cette fois, Amazon a essayé de développer un algorithme capable de parcourir le web rapidement pour repérer des candidats qualifiés. Cependant, cet algorithme a favorisé des verbes plus couramment utilisés par les hommes, comme « exécuté » (*executed)* et « capturé »

(*captured),* ce qui a conduit à des évaluations défavorables pour d'autres candidates, indépendamment de leurs compétences réelles (Dastin, 2018).

Ces exemples montrent que les biais dans l'IA dépassent les simples erreurs techniques. Bien que cela puisse sembler anodin pour certains, les conséquences peuvent être très réelles à différents niveaux. Au lieu de simplifier les processus, ces systèmes deviennent des obstacles pour certains groupes et mettent en lumière les dangers des outils de recrutement automatisés. Les systèmes d'IA ne peuvent pas être considérés comme intrinsèquement impartiaux. Lorsqu'ils sont formés sur des données biaisées, ils perpétuent les biais ou la discrimination, créant ainsi un risque d'inégalités généralisées (Chen, 2023). Ces biais contreviennent aux principes de non-discrimination et d'égalité de traitement consacrés par la Charte des droits et libertés de la personne du Québec (CDLPQ) (Gouvernement du Québec, s.d). Ils contreviennent également aux lois fédérales canadiennes telles que la Loi canadienne sur les droits de la personne et la loi sur l'équité en matière d'emploi (Gouvernement du Canada, 2023).

Outre les conséquences juridiques, ces biais ont un impact économique significatif. La discrimination engendre des coûts importants pour les individus et la société. On estime que de mauvaises décisions d'embauche coûtent 1,6 million de dollars pour chaque tranche de 1 000 recrutements effectués (Beneduce, 2020). De plus, ces inégalités peuvent décourager certains groupes, comme les femmes, de poursuivre des carrières dans des domaines où ils se sentent marginalisés, exacerbant encore les disparités (Chang, 2023).

Ces biais dans les systèmes d'IA ne sont pas des abstractions théoriques mais ont des répercussions réelles et concrètes. S'ils ne sont pas adressés, ils continueront à perpétuer des injustices. Il est donc nécessaire d'entreprendre une réflexion éthique et mettre en place des encadrements prévenir ces dérives.

2.2 Comprendre les biais dans les systèmes d'IA

Les décisions qui sculptent nos parcours de vie, qu'il s'agisse d'accéder à l'éducation, de décrocher un emploi, d'obtenir un crédit ou de recevoir des soins reposent souvent sur des jugements critiques. Lorsque ces décisions sont biaisées, elles peuvent perpétuer les inégalités et restreindre les opportunités pour certains groupes.

L'Organisation internationale de normalisation (ISO) définit les biais dans les systèmes automatisés comme « Une différence systématique dans le traitement de certains objets, personnes ou groupes par rapport à d'autres » (ISO,2021).

Dans le cadre de processus de décision automatisée comme l'utilisation de l'IA dans le recrutement, où certaines étapes sont remplacées par des outils automatisés, les biais peuvent émerger de manière subtile mais persistante. Dans ce contexte, les biais font référence à des erreurs systématiques dans les processus de prise de décision qui entraînent des résultats injustes, souvent au détriment de groupes ou d'individus spécifiques (Ferrara, 2024,).

Les biais sont-ils toujours conscients ?

Les biais ne sont pas toujours intentionnels. Que ce soit dans les systèmes d'IA ou dans nos comportements humains, certains biais sont intentionnels, influencés par des stéréotypes connus, tandis que d'autres sont inconscients,profondément ancrés dans nos croyances implicites. Nos décisions sont souvent influencées par des préjugés implicites

que nous ignorons. Daniel Kahneman, dans « Thinking, Fast and Slow » (Kahneman, 2011), explique que la « pensée rapide » (système 1) est intuitive et automatique, alors que la « pensée lente » (système 2) est plus délibérée et consciente, mais que des biais peuvent se glisser dans les deux. La pensée rapide, peut être sujette aux erreurs et aux stéréotypes. La pensée lente quant à elle peut être influencée par le contexte ou des croyances inconscientes. La Figure 2 illustre ces deux modes de pensée et leurs liens avec les biais. Elle oppose la pensée rapide, intuitive et automatique, à la pensée lente, plus réfléchie mais influencée par le contexte.

Dans le cas des biais conscients, parfois ils sont délibérément intégrés dans les systèmes d'apprentissage automatique (*machine Learning*). Ils visent à simplifier les hypothèses et à améliorer la généralisation, permettant ainsi de produire des prédictions fiables sur de nouvelles données tout en évitant le surajustement (*overfitting*)[2].

En revanche, les biais inconscients, ou implicites, sont plus insidieux. Enracinés dans des normes sociales, ils influencent nos décisions sans que nous en ayons conscience. Ces biais reflètent des attitudes ou stéréotypes automatiques qui influencent nos perceptions et actions sans que nous en soyons pleinement conscients.

Dans le cadre du recrutement, ces biais peuvent subtilement orienter les décisions, même lorsque les recruteurs sont convaincus d'agir de manière totalement objective (Beatti et Johnson, 2012, p.11). Il est essentiel de prendre conscience de l'existence de ces biais et de leur potentiel impact sur nos décisions. Qu'ils soient conscients ou inconscients, leur

[2] Phénomène où un algorithme s'ajuste de manière excessive aux données d'entraînement, perdant ainsi sa capacité de généralisation et devenant incapable de faire des prédictions précises sur de nouvelles données (Ying,2019).

présence peut non seulement reproduire, mais aussi amplifier les inégalités existantes. Dans les systèmes décisionnels basés sur l'IA, comme ceux utilisés pour le recrutement, comprendre ces biais est crucial pour garantir des décisions justes et éthiques. La prise de conscience et la compréhension approfondie de ces mécanismes permettent de concevoir des solutions pour réduire leur impact et favoriser des pratiques plus équitables.

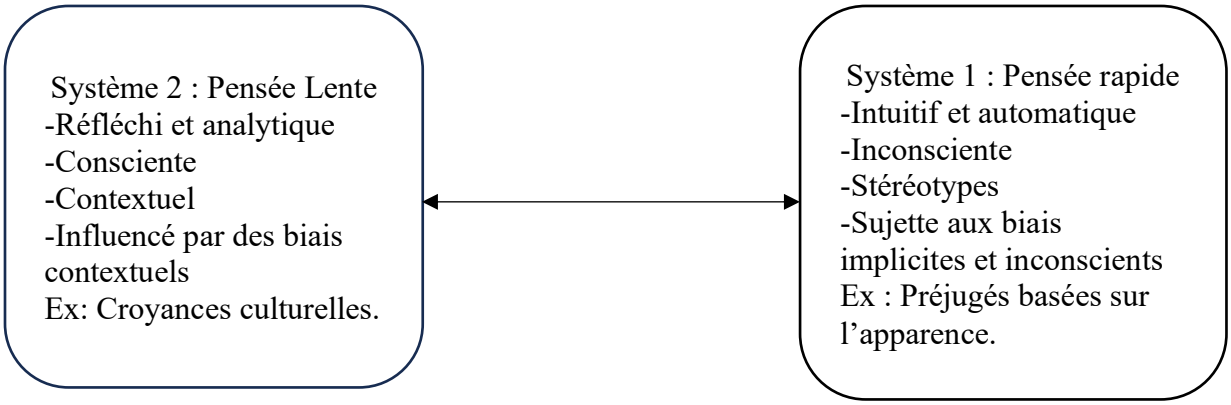

Figure 2: Interaction entre Système 1 et Système 2 : Modèle de Daniel Kahnemam et liens avec les biais (figure inspirée de Kahnemam, 2011)

Pourquoi les systèmes d'IA sont-ils souvent qualifiés de « boîtes noires », et en quoi cette opacité peut-elle amplifier les biais et compliquer leur détection ?

Les systèmes d'IA sont souvent qualifiés de « boîtes noires », une métaphore qui souligne leur manque de transparence et la complexité de leur fonctionnement interne. Une boîte noire fait référence à un système ou un modèle dont les processus internes sont opaques ou incompréhensibles pour les utilisateurs finaux et parfois même pour les scientifiques de données et les développeurs (Hamida et al., 2024). Ainsi, bien que les décisions ou prédictions produites par ces modèles soient visibles, les étapes intermédiaires et les logiques sous-jacentes restent cachées et il est souvent difficile de comprendre comment ces modèles parviennent à leurs prédictions (Gonzalez et al., 2019, p. 36).

Gonzalez et al. (2019) expliquent que lorsqu'ils sont appliqués à de grandes quantités de données, même si les algorithmes sont bien compris et génèrent des prédictions impressionnantes, les relations internes peuvent être trop compliquées pour être interprétées facilement. Par exemple, les forêts aléatoires[3] combinent les résultats de centaines, voire de milliers d'arbres de décision, chaque arbre capturant des interactions spécifiques dans les données, ce qui complique la justification de certains prédicteurs et l'explication de la façon dont certains résultats sont obtenus. D'autres modèles complexes également tels que les réseaux neuronaux profonds[4] (DNN), sont aussi qualifiés de « boîtes noires » car, bien que leur structure et leurs paramètres (poids) soient connus, leur comportement reste difficile à comprendre. Ces modèles peuvent contenir des milliers, voire des millions de paramètres (poids) (Hamida et al., 2024)

Ainsi, en raison de leur complexité et de leur opacité, les systèmes automatisés par l'IA présentent des défis importants en matière de transparence et d'équité, rendant difficile l'identification et la correction des biais. Si les données contiennent des biais implicites ou explicites, le modèle peut non seulement les apprendre mais aussi les amplifier (comme dans les exemples 2.1.2 et 2.1.2). En étant incapables de discerner les étapes

[3] **Les forêts aléatoires** sont une méthode d'apprentissage supervisé qui combine plusieurs arbres de décision structurés. Chaque arbre repose sur un vecteur aléatoire indépendant et identiquement distribué, appliqué uniformément à tous les arbres de la forêt. Lorsqu'une donnée est fournie en entrée, chaque arbre vote pour une classe, et la classe finale est déterminée par le vote majoritaire, permettant ainsi d'obtenir un résultat final unique et robuste (Breiman 2001).

[4] **Les réseaux neuronaux profonds (DNN)** sont des systèmes inspirés des neurones biologiques, capables de modéliser des relations complexes grâce à des couches multiples et des fonctions non linéaires, leur permettant d'approximer diverses fonctions et traiter des données séquentielles dans le cas des réseaux récurrents (Goodfellow et al., 2016, p. 351-354).

exactes de la prise de décision, les utilisateurs peuvent se retrouver à accepter des résultats biaisés sans possibilité de les remettre en question ou de les ajuster.
Cette opacité peut conférer aux biais une légitimité apparente, renforçant l'idée erronée que les décisions de l'IA sont neutres, alors qu'en réalité, elles peuvent perpétuer et même accentuer les préjugés existants.

Les causes des biais : préjugés et heuristiques liés à la complexité et à la société

En conclusion, les biais que nous rencontrons dans les systèmes d'intelligence artificielle ne sont ni accidentels ni le fruit d'un seul facteur isolé. Ils résultent d'une combinaison complexe de facteurs humains, sociaux et technologiques. Par conséquent, la solution ne réside pas uniquement dans l'application directe de méthodes techniques de mitigation, mais dans une démarche qui s'intéresse d'abord à comprendre les causes. En premier lieu, les biais humains, qu'ils soient conscients ou inconscients, influencent directement les décisions de conception des systèmes, intégrant des préjugés implicites ou des stéréotypes profondément ancrés dans nos esprits et nos structures sociales. Ces biais se transmettent dans les données utilisées et dans les modèles eux-mêmes.
Ensuite, il est impossible de comprendre les biais dans les systèmes d'IA sans reconnaître qu'ils ne sont pas uniquement causés par des technologies défaillantes ou des données biaisées, mais qu'ils reflètent les dynamiques sociales. Les inégalités historiques et les stéréotypes enracinés se répètent et se renforcent à travers nos structures sociales.
Comme l'explique (Zajko, 2020) il ne suffit pas d'améliorer les algorithmes, il faut interroger les fondements mêmes des données que nous utilisons, ces données qui incarnent des déséquilibres systémiques et des discriminations persistantes encrés dans la société.

Enfin, la complexité et l'opacité de certains systèmes d'IA amplifient ces biais en rendant leur détection et leur correction d'autant plus difficiles.

Cette dynamique est résumée dans la Figure 3. Elle met en relation trois dimensions : la psychologie humaine (biais cognitifs, heuristiques, préjugés implicites), la société (stéréotypes, inégalités sociales) et les systèmes d'IA (opacité et amplification des biais). La zone centrale souligne que les biais algorithmiques résultent de l'interaction entre ces trois niveaux, où les préjugés humains et sociaux se transmettent et se renforcent dans les modèles d'IA.

Ainsi, le premier pas pour s'attaquer aux biais dans les systèmes d'IA est de reconnaître l'importance de comprendre les causes et sources profondes afin d'appliquer les solutions adéquates pour les réduire.

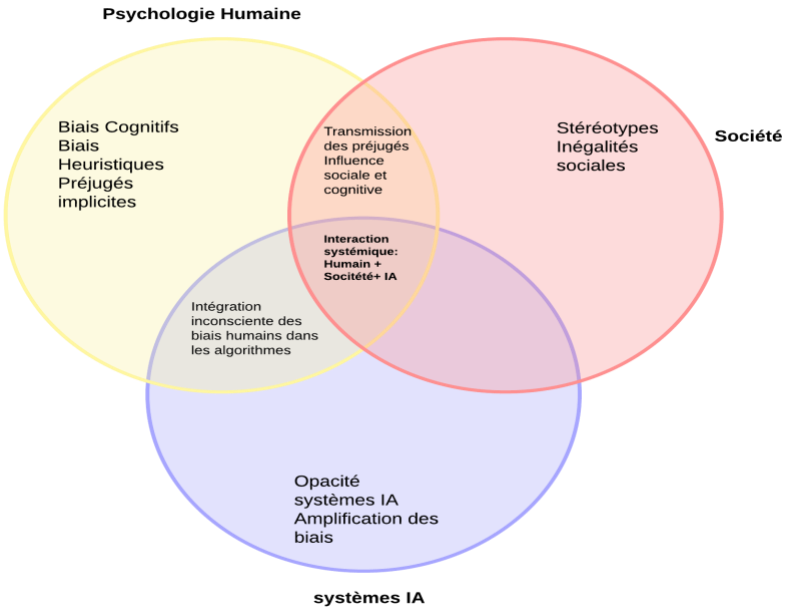

Figure 3: Résumé des causes des biais dans les systèmes d'IA

2.3 Les sources des biais

Les biais dans les systèmes d'IA peuvent provenir de multiples sources dont : les données d'entraînement biaisées, les choix de conception des algorithmes, et les biais provenant des utilisateurs.

Identifier et comprendre ces sources est crucial pour minimiser l'impact des biais et assurer des décisions équitables.

2.3.1 Biais des données

L'empreinte des biais dans les systèmes d'IA est souvent invisible mais profondément ancrée, façonnée par les imperfections et les préjugés présents dans les données qui nourrissent les modèles. Lorsque ces données sont non représentatives, incorrectes ou incomplètes, elles transmettent leurs imperfections aux modèles (Ferrara, 2024).

Les algorithmes dépendant de grandes quantités de données (*big data)* et absorbent les caractéristiques des données qu'ils traitent (Barocas et Selbst, 2016). Leur performance est étroitement liée à la qualité et à la représentativité de ces données. Si ces données sont biaisées, les modèles reproduisent et amplifient ces biais, affectant directement leurs prédictions et leurs décisions. (Mehrabi et al., 2022).

Les ensembles de données manquent souvent de diversité, limitant leur utilité et leur capacité à se généraliser. Par exemple, dans le domaine du recrutement, des données basées sur des décennies de pratiques de sélection peuvent sous-représenter certains groupes, tels que les femmes ou les minorités. Cela conduit les modèles d'IA à privilégier des profils historiquement valorisés. L'exemple d'Amazon, abordé en section 2.1.3, où le modèle était principalement entraîné sur des CV masculins, illustre de manière concrète

comment des biais dans les données peuvent engendrer des résultats discriminants dans des systèmes d'IA automatisés.

2.3.2 Biais des algorithmes

« Un modèle n'est rien d'autre qu'une opinion formulée en termes mathématiques. »

(O'Neil, 2016, p. 21)

Les algorithmes ne sont pas des entités neutres et objectives. Ils intègrent les choix et les hypothèses de leurs concepteurs. Contrairement aux biais des données, les biais algorithmiques ne sont pas introduits par les données elles-mêmes. Ils émergent directement des caractéristiques de l'algorithme utilisé. Les algorithmes conçus comme des séries d'instructions que les ordinateurs suivent pour accomplir des tâches, peuvent intégrer des biais intrinsèques qui se reflètent dans leurs sorties. Ces biais peuvent survenir lorsque les algorithmes sont basés sur des hypothèses biaisées ou utilisent des critères de décision qui favorisent involontairement certains résultats (Ferrara, 2024).

Des choix de conception algorithmique peuvent introduire des biais dans les résultats. Par exemple, l'utilisation de certaines fonctions d'optimisation, de régularisations, ou de modèles de régression peut influencer les décisions. L'utilisation d'estimateurs statistiquement biaisés peut également amplifier ces effets (Mehrabi et al., 2022).

De plus, la nature souvent opaque et complexe des systèmes d'IA, parfois qualifiés de « boîtes noires », peut rendre difficile l'identification et la correction de ces biais, car les processus décisionnels internes ne sont pas transparents (Hamida et al., 2024). Cette opacité peut empêcher les développeurs de détecter si et comment les biais se manifestent, rendant ainsi les systèmes moins justes et potentiellement discriminatoires.

2.3.3 Biais utilisateurs

Les biais utilisateurs se manifestent lorsque les personnes utilisant des systèmes d'IA introduisent, volontairement ou non, leurs propres préjugés dans le fonctionnement du système. Historiquement, les décisions humaines sont souvent influencées par des préjugés conscients et inconscients (Kahneman, 2011), un défi qui se transpose également dans l'utilisation des technologies d'intelligence artificielle. Contrairement aux biais des données ou aux biais algorithmiques, qui sont souvent implicites dans les données ou dans les choix techniques, les biais des utilisateurs sont directement influencés par les décisions humaines, consciemment ou non.

Les biais utilisateurs peuvent survenir lorsque les utilisateurs fournissent des données de formation biaisées ou interagissent avec le système d'une manière qui reflète leurs propres préjugés (Ferrara, 2024). L'interaction humaine avec l'IA peut fortement influencer les résultats produits par ces systèmes.

Un autre exemple de biais des utilisateurs se manifeste dans l'interprétation des scores ou des recommandations produites par l'algorithme. Les utilisateurs peuvent accorder une confiance excessive aux résultats des systèmes d'IA, supposant que ces derniers sont neutres et objectifs. Ce phénomène, appelé « biais d'automatisation », conduit souvent les utilisateurs à accepter les décisions de l'algorithme sans les questionner, même si celles-ci sont biaisées ou inexactes (Goddard et al., 2012). Dans le contexte des décisions sensibles, comme le recrutement, cette confiance aveugle peut conduire à des discriminations injustes et à des erreurs difficiles à corriger.

Le Tableau 1 propose une synthèse des différentes sources de biais évoquées, en rappelant leurs définitions, leurs causes et leurs principales caractéristiques, tout en fournissant des exemples de leurs manifestations.

Source de biais	Définition et causes	Caractéristiques	Exemples
Biais des données	Survient lorsque les données sont non représentatives, incorrectes ou incomplètes. Provient de sources biaisées, incomplètes, erronées.	-Amplifie et perpétue les biais et stéréotypes existants. -Affecte les prédictions des algorithmes et leur performance -Souvent invisible	- Modèle formé sur des données non représentatives échouant à reconnaître d'autres groupes. -Outil de recrutement d'Amazon favorisant les CV masculins à cause de données historiques non représentatives.
Biais algorithmique	Survient lorsque les algorithmes sont basés sur des hypothèses biaisées ou utilisent des critères de décision qui favorisent involontairement certains résultats biaisés. Provient de choix de conception algorithmique tels que l'utilisation de fonctions d'optimisation favorisant certains résultats, l'application critères de décision discriminants, la complexité ou opacité des modèles (boîte noire), ou encore l'utilisation d'estimateurs statistiquement biaisé.	-Introduit dans le code -Difficile à détecter et corriger -Peut amplifier les biais de données	- Algorithme favorisant certains mots-clés ou des critères non neutres dans le recrutement. - Modèles excluant involontairement des sous-groupes.
Biais des utilisateurs	Introduit par les préjugés conscients ou inconscients des utilisateurs lors de l'interaction ou de l'interprétation des résultats d'un système IA. Survient pendant l'interaction ou l'interprétation des résultats.	- Directement influencé par les décisions humaines. - Peut fausser les résultats ou décisions critiques. -Souvent difficile à corriger.	-Utilisateur acceptant aveuglément les recommandations biaisées d'un système IA. - Interaction avec un système IA influencée par des préjugés conscients ou inconscients. Exemple : un recruteur qui configure le système pour privilégier des

		critères tels qu'une université spécifique, favorisant indirectement des candidats issus d'un certain milieu.

Tableau 1: Résumé des sources de biais, causes et exemples dans les systèmes d'IA

2.3.4 La boucle des biais dans les systèmes IA

Mehrabi et al. (2022) mettent en lumière un phénomène où les biais présents dans les données, les algorithmes et les interactions des utilisateurs forment une boucle de rétroaction. Ce cercle vicieux est particulièrement préoccupant dans le domaine du recrutement.

Tout commence avec les données d'entraînement. Lorsque les données d'entraînement contiennent des biais dès le départ, les algorithmes les absorbent, les reproduisent et bien souvent les amplifient. Ainsi, l'algorithme va fournir des résultats biaisés aux utilisateurs, qui contribue involontairement à les renforcer à leur tour. Cela aboutit souvent à des résultats injustes et discriminatoires persistants. Ainsi, les effets discriminatoires s'installent et durent.

Prenons un exemple concret. Imaginons un algorithme conçu pour trier des candidatures en se basant sur des mots-clés ou certaines compétences. Si cet algorithme a été alimenté par des données biaisées, il privilégiera des profils historiquement favorisés, Ces profils ayant souvent des parcours similaires. Ces candidats apparaîtront en tête du classement, attirant l'attention des recruteurs. Influencés par ces résultats, les recruteurs se

concentreront principalement sur ces profils, en consultant leurs CV ou en les invitant à des entretiens, tout en négligeant d'autres profils potentiellement qualifiés. Ces interactions des recruteurs, comme leurs consultations, leurs clics et décisions, sont ensuite collectées par l'algorithme et réintégrées dans le système comme nouvelles données et pour entrainer d'autres modèles. Sauf qu'en réalité, on ne fait que renforcer les mêmes travers.

Ainsi, les candidats déjà favorisés par les premières versions de l'algorithme continueront d'apparaître en tête des classements, non pas parce qu'ils sont objectivement plus qualifiés, mais parce qu'ils ont bénéficié de la visibilité et de l'attention supplémentaires dans le processus de sélection. À l'inverse, ceux qui sont écartés au départ le resteront. Le modèle, croyant apprendre, ne fait en réalité que tourner en rond.

Au fil du temps, cette boucle de rétroaction se poursuit. C'est ainsi qu'un biais au départ relativement discret peut à force de répétitions et d'interactions devenir une boucle. Le processus dans son ensemble finit par favoriser systématiquement les mêmes profils, excluant progressivement des candidats issus de parcours différents, même s'ils sont qualifiés.

Ce phénomène de rétroaction, illustré à la Figure 4, amplifie les préjugés initiaux, créant un cercle insidieux où biais des données, biais algorithmiques et biais des utilisateurs se renforcent mutuellement. Les décisions deviennent alors de plus en plus discriminatoires, compromettant l'équité du processus de recrutement.

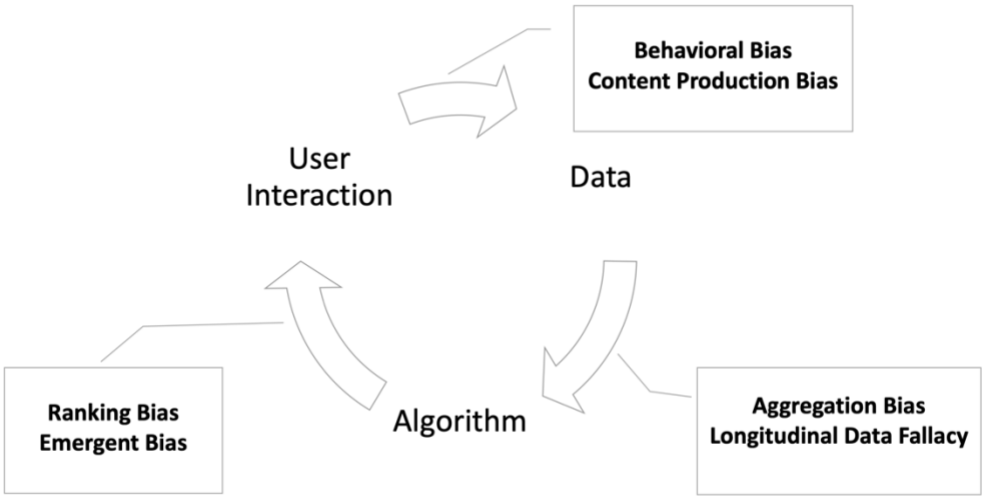

Figure 4: Boucle de rétroaction entre les données, les algorithmes et l'interaction utilisateur. Tiré de Mehrabi et al., 2022, p. 4

2.4 L'équité et l'éthique dans les systèmes de décisions automatisées par l'IA

Les notions d'équité et d'éthique sont des concepts aussi riches que complexes, traversant les disciplines et les époques avec une grande diversité de significations. Souvent évoqués dans des contextes sociaux, juridiques et technologiques, ces termes incarnent des enjeux fondamentaux de justice et de moralité, particulièrement pertinents à l'ère numérique. Ces notions varient significativement selon les perspectives culturelles et disciplinaires.

Tandis que la philosophie et la psychologie ont exploré ces concepts bien avant les sciences informatiques, l'introduction de l'intelligence artificielle a redéfini et complexifié leur compréhension. Ces divergences rendent difficile l'établissement de définitions universelles qui satisferait toutes les parties dans un contexte donné, soulignant ainsi leurs relativités intrinsèques (Barocas et al., 2023).

« L'éthique est ce que vous faites lorsque personne ne vous regarde. » (Villemure, 2023, p.26). Cela résonne dans l'univers de l'IA, où les décisions prises derrière les écrans, dans le code, les algorithmes et les ensembles de données doivent être régies par des principes moraux solides, souvent en l'absence de surveillance externe. L'éthique cherche ici à établir des principes moraux et guider le développement et l'utilisation des systèmes automatisés par l'IA pour assurer qu'ils respectent des normes de justice et d'équité.

Dans le contexte de la prise de décision, l'équité se définit par l'absence de tout préjugé ou favoritisme envers un individu ou un groupe, basé sur leurs caractéristiques inhérentes ou acquises (Mehrabi et al., 2022). Dans les systèmes d'IA, elle se réfère à l'absence de biais ou de discrimination, ce qui peut être difficile à atteindre en raison des différents types de biais susceptibles de survenir dans ces systèmes (Ferrara, 2024). C'est un concept à la fois difficile à définir qu'à atteindre.

L'équité peut être envisagée sous deux formes principales : l'équité individuelle et l'équité de groupe. L'équité individuelle exige qu'un modèle d'IA produise des résultats prédictifs similaires pour des individus similaires, tandis que l'équité de groupe requiert que les différents groupes soient traités de manière égale par le modèle (Chen et al., 2023). Pour ce mémoire, nous nous concentrerons sur la notion d'équité de groupe.

L'équité de groupe implique une division entre groupes privilégiés et non privilégiés, fondée sur des attributs protégés tels que le genre ou la race (Chen et al., 2023). Ces attributs, souvent associés à des injustices historiques, nécessitent une attention particulière pour garantir que les systèmes d'IA ne reproduisent pas ces inégalités.

Qu'est-ce qui est considéré comme équitable ou non dans le système de décisions automatisés par l'intelligence artificielle ?

Un système d'intelligence artificielle peut être considéré comme équitable lorsqu'il garantit que tous les individus, indépendamment de leur origine, genre ou statut socio-économique, disposent des mêmes opportunités d'obtenir un résultat favorable. Un système est équitable lorsqu'il garantit des décisions impartiales, basées sur des critères pertinents et accessibles à tous, tout en étant transparent et ouvert à la contestation. (Barocas et al., 2023)

Par exemple, un algorithme de recrutement qui sélectionne de manière équitable des candidats qualifiés, qu'ils appartiennent à des groupes sous-représentés ou majoritaires, pourrait être perçu comme équitable. Cependant, l'équité ne se limite pas aux résultats, elle englobe également le processus décisionnel. Des éléments tels que la transparence, l'interopérabilité, l'accessibilité, la responsabilité et la capacité d'expliquer les choix algorithmiques (Intelligence Artificielle explicable[5]) jouent un rôle crucial pour assurer la légitimité des décisions prises par ces systèmes. À l'inverse, un système est inéquitable lorsqu'il engendre des discriminations, amplifie les inégalités existantes ou manque de transparence et de responsabilité dans ses décisions (Barocas et al., 2023).

[5] L'Intelligence Artificielle Explicable (XAI) vise à améliorer la compréhension et la confiance des utilisateurs en fournissant des explications claires sur les décisions et les processus de l'IA. (Hamida et al., 2024)

Les biais évoqués dans les sections précédentes résultent de multiples facteurs comme des données de formation non représentatives, des choix de conception algorithmique ou encore d'interprétations humaines biaisées. Corriger ces biais constitue une étape essentielle pour progresser vers l'équité en IA. Les lignes directrices éthiques et les cadres réglementaires peuvent aider à mieux repérer les risques de biais et d'iniquité. Ces cadres offrent un repère essentiel pour garantir la responsabilité des systèmes d'intelligence artificielle (Oyeniran et al., 2022). Dans cette optique, plusieurs cadres éthiques ont émergé pour encadrer l'utilisation de l'intelligence artificielle. (Discuté à la section 2.6 de ce chapitre). Ces initiatives visent à promouvoir des normes fondamentales d'équité et à protéger les utilisateurs contre les impacts négatifs de ces technologies.

En somme, les concepts d'éthique et équité dans les systèmes d'IA ne se limitent pas à des concepts abstraits. Ce sont des objectifs dynamiques et évolutifs qui appellent une réflexion et vigilance constante sur les processus décisionnels et leurs impacts sociaux. Elles invitent concepteurs et utilisateurs à privilégier des solutions, dont nous discuterons dans les sections suivantes, respectant à la fois les données et les valeurs humaines fondamentales.

2.5 Le pipeline de recrutement : étapes et exemples de biais potentiels

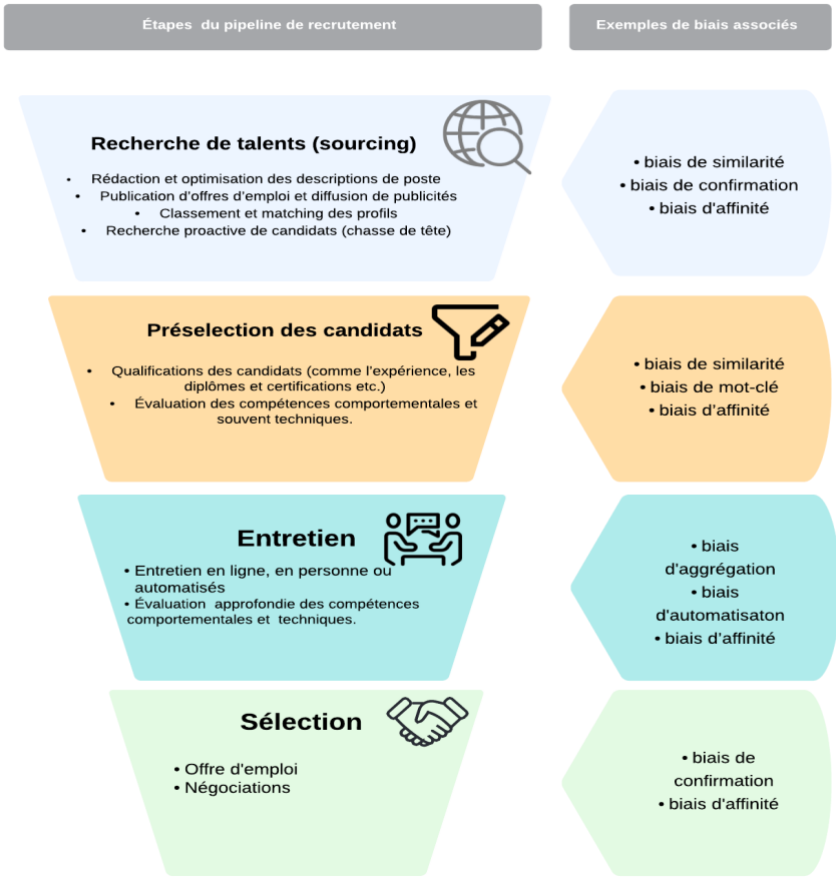

Figure 5: Pipeline de recrutement et exemples de biais à chaque étape (figure inspirée de Bogen et Rieke, 2018)

Pourquoi le recrutement est un pipeline?

Le processus de recrutement ne repose pas sur une unique décision, mais sur une série cumulative de petites décisions qui s'enchaînent (Chen, 2023). C'est pourquoi on le

considère souvent comme un pipeline[6]. Chaque étape agit comme un levier ou un filtre, visant à identifier les profils les plus adaptés. Cependant, chacune d'elles est susceptible d'introduire des biais, qu'ils proviennent des données, des algorithmes ou des décisions humaines. Ces différentes étapes successives, ainsi que quelques exemples de biais susceptibles d'apparaître à chacune d'entre elles, sont résumées à la Figure 5.

2.5.1 Les étapes du pipeline

Tout commence par la première étape, le sourcing, qui regroupe l'ensemble des actions entreprises par les employeurs pour repérer et attirer les candidats susceptibles de correspondre à un poste spécifique. Cela passe par la rédaction de l'offre d'emploi, son adaptation au public visé, le tri des premiers profils repérés, ainsi que la recherche proactive de candidats, notamment auprès de personnes qui ne sont pas nécessairement en recherche (ce qu'on appelle communément la chasse de têtes). Ces dernières années, les outils numériques et en particulier ceux propulsés par l'intelligence artificielle ont commencé à remodeler cette étape. Des outils comme Textio rendent les descriptions de poste plus inclusives et engageantes, tandis que des plateformes telles que LinkedIn ou ZipRecruiter exploitent des algorithmes pour repérer et classer les profils correspondant aux besoins de l'entreprise. Pour la recherche proactive, des solutions comme Searchlight ciblent des candidats en dehors des canaux traditionnels (Bogen et Rieke, 2018).

Vient ensuite la présélection, qui constitue une étape clé du recrutement. On y évalue les diplômes, les parcours professionnels, mais aussi de plus en plus les aptitudes

[6] **Pipeline de recrutement :** une série d'étapes successives et structurées, à travers lesquelles un candidat potentiel progresse pour parvenir à une décision finale d'embauche.

comportementales. Là encore, l'IA se fait une place. Certains outils scannent les CV à la recherche de mots-clés ou de correspondances implicites, attribuant des scores de compatibilité. Des jeux cognitifs comme ceux de Pymetrics pour évaluer les compétences comportementales et les tests de personnalité prédictifs comme Koru permettent d'estimer l'adéquation des candidats à un poste (Bogen et Rieke, 2018).

L'entretien est également une étape clé du processus. On y mesure bien plus que les compétences des candidats mais aussi la motivation et la compatibilité des candidats avec la culture de l'entreprise. Ce moment d'échange, jadis purement humain, s'automatise lui aussi de plus en plus. Des outils comme HireVue s'immiscent dans les interactions, analysant la voix, les expressions faciales, jusqu'aux tournures de phrases, dans le but de produire un score qui guidera la suite du tri (Bogen et Rieke, 2018).

Enfin, vient la sélection, qui correspond à l'étape finale dans le processus. Dans cette phase, des outils d'IA avancés sont souvent utilisés pour affiner le choix final. Par exemple, Fama analyse les traces numériques pour repérer d'éventuels signaux faibles de comportements à risque (harcèlement, violence, etc.). D'autres, comme Oracle Recruiting Cloud, vont jusqu'à estimer les chances qu'a un candidat d'accepter l'offre, et proposent des ajustements pour maximiser ces chances (Bogen et Rieke, 2018).

2.5.2 Biais possibles lors de l'introduction des systèmes IA dans le pipeline

Malgré les avancées technologiques, l'intégration des systèmes et outils IA dans le processus de recrutement peut perpétuer et amplifier des inégalités structurelles préexistantes ou introduire de nouveaux biais. Ces outils, bien que puissants, présentent des limites significatives. Leur efficacité repose largement sur la qualité des données

d’entraînement et les paramètres définis, ce qui peut introduire des biais systémiques compromettant l’équité du processus de recrutement (Raghavan et al., 2019).

Un biais pernicieux, le biais de similarité, survient lorsque les algorithmes ou les recruteurs favorisent inconsciemment des candidats dont les profils ressemblent à ceux des employés actuels ou aux recruteurs eux-mêmes. Ce biais se définit comme une préférence pour les individus qui partagent des caractéristiques communes avec l'évaluateur (Rivera, 2012). Alimenté par des données biaisées, ce phénomène reproduit des schémas historiques d’exclusion. Il perpétue des normes organisationnelles susceptibles de refléter des pratiques discriminatoires passées, limitant ainsi l’accès à des opportunités pour les groupes sous-représentés. Par exemple, un algorithme entraîné sur des données historiques biaisées peut systématiquement privilégier des profils issus des parcours professionnels similaires. Il freine les initiatives en faveur de la diversité et renforce une homogénéité qui peut entraver l’innovation organisationnelle.

Un autre biais, le biais de confirmation, se manifeste de manière subtile mais importante. Il se définit comme une tendance à rechercher et à accepter plus facilement les informations qui confirment nos croyances ou jugements préexistants, tout en ignorant ou en minimisant les informations qui les contredisent (Bashkirova et Krpan 2024).Par exemple, si un algorithme attribue un score élevé à un candidat en se basant sur des critères biaisés, un recruteur influencé par ce score pourrait se concentrer uniquement sur les éléments du CV qui confirment cette évaluation, ignorant des signaux contradictoires comme un manque d'expérience pertinente.

De plus, le biais de mot-clé agit comme un filtre linguistique. Il émerge lorsque les algorithmes privilégient certains termes spécifiques dans les CV ou les descriptions de poste. Cela peut conduire à l'exclusion de candidats compétents qui ne formulent pas leurs qualifications selon les standards linguistiques attendus par l'algorithme, même si leurs compétences correspondent parfaitement aux exigences du poste. Ce biais limite la diversité en favorisant des candidats issus de contextes où ces termes sont plus courants. (Comme illustré dans l'exemple de la section 2.1.3)

Le biais d'agrégation survient lorsque de fausses conclusions sont tirées sur des individus en se basant sur l'observation de l'ensemble de la population (Mehrabi et al., 2022). Par exemple, si les données historiques indiquent que les hommes ont réussi dans un poste technique, un algorithme de recrutement pourrait conclure que les candidats masculins sont globalement mieux adaptés à ces rôles. Ainsi, l'algorithme pourrait injustement accorder un avantage aux candidats masculins, sans égard aux compétences réelles des femmes postulantes.

Le biais d'automatisation, quant à lui, désigne la confiance excessive et aveugle accordée aux résultats produits par les systèmes d'IA. Les recruteurs peuvent supposer que les algorithmes sont neutres et impartiaux, acceptant leurs recommandations sans les remettre en question. Ce biais peut conduire à des décisions injustes, en excluant des candidats dont les qualités ne sont pas correctement traduites par les métriques numériques (Goddard et al., 2012).

Enfin, les biais des utilisateurs**,** tels que le biais d'affinité, aggravent également ces déséquilibres. Il est étroitement lié au biais de similarité. Le biais d'affinité reflète une

préférence pour les candidats partageant des similarités culturelles ou personnelles avec les recruteurs, souvent au détriment d'une évaluation objective des compétences (Rivera, 2012). Par exemple, un recruteur influencé par ses préjugés inconscients peut ajuster les critères de sélection ou modifier les recommandations algorithmiques pour favoriser des candidats lui ressemblant. Ces interventions involontaires peuvent amplifier les biais et compromettre l'équité du processus de sélection.

2.6 Méthodes de mitigation des biais

2.6.1 Les méthodes non techniques : Approches organisationnelles et éthiques pour la mitigation des biais

Bien que les méthodes techniques jouent un rôle crucial dans la réduction des biais, elles ne suffisent pas à elles seules. Les approches non techniques, qui s'appuient sur des dimensions organisationnelles, humaines et éthiques, offrent une perspective complémentaire et essentielle pour garantir une équité véritable et durable.

- **Sensibilisation : Construire une prise de conscience collective**

La sensibilisation joue un rôle central dans la promotion de l'équité dans les systèmes d'intelligence artificielle. Le manque de sensibilisation est particulièrement problématique dans des domaines essentiels où la compréhension de l'ensemble du processus de prise de décision est indispensable (Hamida et al., 2024). Elle a pour objectif d'éduquer les parties prenantes, notamment les développeurs, décideurs et utilisateurs, sur les biais implicites susceptibles d'influencer les décisions critiques à chaque étape du cycle de vie des systèmes. Comprendre les différents types de biais, leurs

sources, leurs origines ainsi que leurs impacts est fondamental pour élaborer des stratégies qui garantissent l'équité, la transparence et la responsabilité des systèmes d'IA (Oyekunle et al., 2022). En outre, parvenir à l'équité dans les systèmes d'IA exige une compréhension nuancée des différentes formes d'équité et des compromis nécessaires pour les appliquer dans des contextes variés (Ferrara., 2023).

Toutefois, cette prise de conscience doit s'accompagner d'outils concrets adaptés aux réalités des organisations et dépasser la simple identification des biais. Par exemple, des stratégies comme des ateliers interactifs et des formations continues donnent aux acteurs principaux les moyens de réfléchir de manière critique à leurs propres préjugés et à leurs impacts sur les systèmes. Ces initiatives permettent non seulement d'éviter l'apparition de nouveaux biais, mais aussi de corriger les distorsions déjà ancrées. Ainsi, les organisations renforcent la confiance des utilisateurs et favorisent la construction de systèmes véritablement équitables et responsables. Une organisation qui accorde la priorité à l'équité et aux considérations éthiques est mieux positionnée pour mettre en œuvre des pratiques capables de réduire efficacement les biais (Oyekunle et al., 2022).

- **Cadres éthiques : encadrer les efforts**

Les initiatives éducatives, aussi essentielles soient-elles, ne suffisent pas à elles seules. Elles doivent s'accompagner de cadres éthiques solides qui orientent le développement et l'utilisation des systèmes d'intelligence artificielle. Cela passe par l'adoption de directives claires et de régulations pensées pour garantir l'équité, la transparence et la responsabilité tout au long du processus (Chen, 2023).

Certaines organisations s'y attellent déjà. Des cadres et lignes directrices éthiques sont nés pour servir de fondation au développement éthique de l'intelligence artificielle. L'IEEE, par exemple, à travers son initiative The *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,* a développé plusieurs outils et lignes directrices pour une conception éthique de l'intelligence artificielle (IEEE Standards Association, 2016). Parmi eux figure notamment le standard IEEE P7003, qui vise à identifier, mesurer et atténuer les biais dans les systèmes d'IA (IEEE Standards Association, 2024). Des initiatives comme la déclaration de Montréal pour une IA responsable (Université de Montréal, 2018) ou les principes directeurs de l'OCDE (OCDE, 2019) offrent également des repères concrets pour promouvoir des pratiques éthiques et collaboratives. Ces cadres s'appuient sur des valeurs fondamentales : le respect des droits humains, la protection des libertés individuelles, la transparence, l'équité et l'adhésion aux principes démocratiques. Intégrer ces réflexions éthiques dès les premières étapes de conception n'est pas seulement une précaution, mais une nécessité pour prévenir des dérives potentielles et ancrer des pratiques responsables au cœur même des systèmes. C'est ce qui permet aux organisations de créer des technologies à la fois fiables, justes et dignes de confiance mais aussi de favoriser un climat de confiance entre les utilisateurs et les technologies.

- **Surveillance continue et audit : Prévenir les biais émergents**

Pour garantir une intelligence artificielle véritablement équitable, la surveillance continue et les audits réguliers des systèmes tout au long de leur cycle de vie sont indispensables. Ces pratiques permettent de détecter et de corriger les biais avant qu'ils ne deviennent structurels, assurant ainsi une justice et une transparence durables (Chen, 2023). Un audit

efficace évalue les performances des modèles sur différents groupes démographiques afin d'identifier et de corriger les disparités éventuelles (Oyekunle et al., 2022). Il scrute aussi les biais invisibles, glissés dans les résultats algorithmiques, tout en s'assurant que les données ne soient ni déséquilibrées, ni déformées par des héritages discriminatoires. L'objectif est de garantir des décisions qui sont non seulement compréhensibles, mais aussi justifiables, et acceptées par les utilisateurs. Quoi qu'il en soit, un audit ne s'arrête pas au lancement d'un système. Il doit se prolonger après le déploiement, en intégrant les retours d'une diversité d'acteurs. Cette vigilance permet de prévenir les dérives sociales et éthiques tout en favorisant la conception de systèmes équitables et adaptés aux besoins des utilisateurs.

- **Transparence et explicabilité: éclairer la boîte noire**

La transparence constitue un pilier essentiel pour garantir l'équité des systèmes d'intelligence artificielle. Trop souvent, les algorithmes fonctionnent comme des boîtes noires rendant leurs décisions difficiles à comprendre ou à contester. L'intelligence artificielle explicable (XAI) s'efforce de répondre à ce défi en développant des modèles capables de justifier leurs choix. Par exemple, un système explicable peut fournir des explications claires sur le rejet d'un candidat lors d'un processus de recrutement automatisé, ouvrant ainsi la voie à la correction des biais identifiés.

L'explicabilité constitue une condition essentielle pour identifier les biais et, ce faisant, instaurer la confiance des utilisateurs, tout en favorisant une utilisation responsable des technologies d'IA (Hamida et al., 2024).

- **Mécanismes de retour d'information : une boucle d'amélioration continue**

Supprimer tous les biais dans les systèmes d'IA reste une ambition louable, mais encore hors de portée. En revanche, il est possible de les surveiller et de les réduire grâce à des mécanismes de retour d'information. Ces dispositifs offrent aux utilisateurs la possibilité de signaler les anomalies ou les injustices qu'ils perçoivent dans les décisions algorithmiques. Ce dialogue, parfois absent dans les systèmes automatisés, est pourtant crucial. Il contribue non seulement à améliorer les modèles, mais peut aussi permettre d'identifier des biais jusqu'alors inconnus ou des conséquences involontaires qui ont pu survenir lors du fonctionnement du système (Samiksha et Ipsita, 2023).

Par ailleurs, il peut jouer un rôle clé en brisant les cercles vicieux où les biais des données, des algorithmes et des utilisateurs se renforcent mutuellement, comme abordé dans la section 2.3.d. En d'autres termes, c'est en impliquant les utilisateurs que l'on conçoit des systèmes plus équitables et véritablement adaptés à leurs attentes.

- **Limiter la confiance excessive aux systèmes d'intelligence artificielle : vers un équilibre entre humain et machine**

Faire aveuglément confiance aux systèmes d'intelligence artificielle comporte des risques non négligeables. Bien qu'efficaces et prometteuses, ces technologies peuvent insidieusement renforcer des biais préexistants ou engendrer de nouvelles formes d'injustice si elles sont utilisées sans discernement.

Les utilisateurs ont tendance à accepter les recommandations éthiques d'algorithmes, même lorsqu'ils ignorent tout des données d'entraînement ou qu'ils savent si celles-ci

sont biaisées (Krügel et al., 2021, p.3). Par ailleurs, cette confiance est d'autant plus problématique que la notion d'étalonnage de la confiance appliquée aux systèmes l'IA[7] demeure souvent mal comprise. Il n'est pas rare que l'on tienne pour fiable un système qui exprime un niveau de confiance trop élevé ou trop faible, sans chercher à en vérifier la légitimité (Li et al., 2024). Cette confiance excessive présente un risque d'acceptation non critique des résultats algorithmiques, pouvant mener à des erreurs coûteuses et injustes. Plusieurs facteurs expliquent cette adhésion, dont notamment la performance perçue, la transparence, l'expérience utilisateur et la propre confiance des individus. Mais cette confiance aveugle peut engendrer des erreurs graves, les utilisateurs acceptant des suggestions erronées sans discernement (Chong et al., 2022). Dans les domaines où les décisions influent directement sur des trajectoires humaines, cette confiance excessive est d'autant plus inquiétante.

L'objectif n'est pas seulement de limiter la confiance en soi, mais de l'étalonner. Un étalonnage adéquat implique que le niveau de confiance des utilisateurs dans un système d'IA corresponde à sa probabilité réelle de réussite, comme l'ont souligné Li et al. (2024) et Chong et al. (2022). Adopter une posture critique face aux résultats algorithmiques est essentiel, en particulier lorsqu'ils touchent à des enjeux sociétaux majeurs. Cela implique aussi de veiller à une calibration rigoureuse des modèles, de manière à garantir une cohérence dans leur fonctionnement, et ce, pour tous les groupes démographiques concernés.

[7] **L'étalonnage de la confiance en IA** désigne le degré auquel la confiance exprimée par une IA correspond à sa probabilité réelle de justesse (Li et al., 2024).

2.6.2 Les méthodes techniques de mitigation des biais

Après avoir exploré en profondeur les concepts de biais, d'éthiques et d'équité dans les systèmes automatisés d'IA et leurs enjeux, il est maintenant temps de se pencher sur les méthodes techniques de mitigation. Ces approches jouent un rôle crucial dans la réduction des biais. En effet, c'est sur le terrain technique que s'esquisse une première ligne de réponse.

Les biais dans les systèmes d'intelligence artificielle représentent une problématique complexe, mais la communauté du *machine Learning* a développé une large gamme de méthodes de mitigation pour y remédier. La littérature scientifique regroupe ces interventions en trois familles, selon le moment auquel elles s'intègrent dans le développement d'un modèle prédictif. Chacune de ces méthodes interviennent à différents niveaux du cycle de vie des modèles. Les méthodes de pré-traitement interviennent avant même l'entraînement des modèles, les méthodes d'in-traitement ajustent le processus d'apprentissage lui-même, et enfin, les méthodes de post-traitement corrigent les prédictions finales une fois le modèle entraîné.

Chacune de ces approches offre des leviers uniques pour lutter contre les biais avec leurs propres stratégies et objectifs, permettant de cibler précisément les sources d'injustice, qu'elles se trouvent dans les données, les algorithmes ou les résultats.

I. Les méthodes pré-traitement

Les méthodes de pré-traitement sont appliquées avant que le modèle ne soit entraîné. Elles modifient les données d'entraînement pour atténuer les biais potentiels avant que ces biais ne soient appris par le modèle. Parmi les méthodes de prétraitement, on trouve la pondération par rééquilibrage « *Reweighing* » (Kamiran et Calders, 2012) qui génère

des poids différents pour les échantillons dans chaque combinaison (groupe, étiquette) afin de garantir l'équité avant la classification. Une autre méthode est l'apprentissage de représentations équitables « *Learning Fair Representation* » qui masque les informations sensibles pour créer des représentations équitables (Zemel et al., 2013).

II. Les méthodes in-traitement

Les méthodes d'in-traitement interviennent directement au cours de l'entraînement des modèles. Leur objectif est d'adapter les algorithmes d'apprentissage pour réduire ou éliminer la discrimination en temps réel. Cela peut être réalisé en modifiant la fonction objective de l'algorithme ou en imposant des contraintes spécifiques qui favorisent l'équité dans les prédictions. (Mehrabi et al., 2022)

Parmi les méthodes d'in-traitement, on peut citer la méthode d'apprentissage adversarial «*Adversarial Debiasing*» (Zhang et al., 2018), qui utilise un modèle adversaire pour détecter et minimiser les biais pendant l'entraînement. Cette approche apprend un classificateur qui maximise la précision des prédictions tout en réduisant simultanément la capacité de l'adversaire à identifier l'attribut protégé. Cela permet de créer un classificateur équitable, car les prédictions ne contiennent plus d'informations discriminatoires. De plus, une autre approche est la réduction par gradient exponentiel «*Exponentiated Gradient Reduction*» (Agarwal et al., 2018). Elle transforme le problème de l'équité en une série de problèmes de classification sensibles au coût, en retournant un classificateur aléatoire minimisant l'erreur empirique tout en respectant les contraintes d'équité.

III. Les méthodes post-traitement

Les méthodes de post-traitement interviennent après l'entraînement du modèle, en utilisant un ensemble de validation distinct qui n'a pas été impliqué dans le processus d'apprentissage. Ces approches sont particulièrement pertinentes lorsque l'algorithme ou les données d'entraînement ne peuvent être modifiés, traitant ainsi le modèle comme une boîte noire. Dans ce contexte, les étiquettes prédites par le modèle sont réattribuées selon une fonction spécifique pour améliorer l'équité des résultats. Cette stratégie permet de corriger les éventuels biais identifiés sans toucher aux fondations initiales du modèle (Mehrabi et al., 2022). Parmi ces méthodes, on retrouve la méthode dite d'égalisation calibrée des chances « *Calibrated Equalized Odds* » (Pleiss et al., 2017), une technique de post-traitement qui optimise les scores d'un classificateur calibré pour trouver les probabilités avec lesquelles changer les étiquettes de sortie, en suivant un objectif d'égalisation des chances. Une autre stratégie, appelée classificateur à option de rejet « *Reject Option Classifier* » (Kamiran et al., 2012) qui attribue des résultats favorables aux groupes non privilégiés et des résultats défavorables aux groupes privilégiés dans une bande de confiance autour de la frontière de décision, là où l'incertitude est la plus élevée.

Chapitre 3 : Méthodologie

Ce chapitre présente la méthodologie suivie dans le cadre de cette recherche. L'étude repose sur un cadre expérimental appliqué à la prédiction algorithmique en contexte de recrutement. Ce cadre est détaillé dans la section B du présent chapitre. L'objectif est d'évaluer l'impact de biais, explicites ou implicites, sur les décisions générées par des systèmes d'apprentissage automatique. Nous cherchons à la fois à mesurer l'équité de ces décisions, à identifier les dérives possibles et à explorer les moyens de les atténuer.

Dans cette optique, un environnement contrôlé et reproductible visant à reproduire des conditions réalistes tout en permettant une évaluation de l'équité algorithmique a été mis en place. En alignement avec les principes FAIR (*Findable, Accessible, Interoperable, Reusable*) (Wilkinson et al., 2016), chaque composante du dispositif a été documentée : jeu de données, code, hyperparamètres et scénarios expérimentaux. Semmelrock H. et al. (2025) précisent que la reproductibilité est au cœur de toute démarche scientifique rigoureuse, notamment en apprentissage automatique, où les résultats peuvent varier fortement selon les paramètres choisis. Elle permet de confirmer les résultats, de les étendre et de faire face aux inquiétudes grandissantes liées à la crise de la reproductibilité en apprentissage automatique.[8]. Dans cette perspective, l'intégralité du code source développé dans le cadre de cette étude a été rendu accessible dans un dépôt GitHub[9].

[8] **La crise de la reproductibilité en apprentissage automatique** désigne l'ensemble des obstacles qui empêchent de reproduire les résultats scientifiques, en raison notamment du manque de code ou de données accessibles, de la variabilité des entraînements de modèles et de la complexité des pipelines. Ce phénomène soulève des doutes sur la fiabilité et la validité des découvertes produites dans ce domaine.(Semmelrock et al. 2025)

[9] https://github.com/balkissaa/Memoire_Biais_IA

Outre la reproductibilité, notre démarche repose aussi sur la transparence et la réutilisabilité. Deux outils à code source libre ont été introduits dans notre protocole : FairCVtest[10], conçu pour tester les biais dans des scénarios de recrutement simulés, et la bibliothèque IBM AIF360[11], qui propose un ensemble d'outils d'apprentissage automatique pour la détection et la réduction des biais.

La stratégie utilisée est à la fois empirique et quantitative. Elle repose sur un jeu de données structuré et sur l'entraînement de modèles prédictifs. Les données modélisées sont analysées à l'aide d'indicateurs statistiques axés sur l'équité. Ces stratégies permettent d'établir des relations de causalité à partir d'observations mesurables, c'est-à-dire de relier les variations observées à des causes précises, en s'appuyant sur des données mesurables (Lim 2024). Cela est particulièrement utile pour analyser les effets globaux des algorithmes. L'approche quantitative facilite la formulation d'hypothèses précises ainsi que leur vérification à l'aide de données chiffrées (Creswell et Creswell, 2014).

Consciente des limites associées à une évaluation menée dans un environnement unique, cette recherche adopte une perspective comparative tenant compte du contexte. Ganesh et al. (2024) mettent en garde contre les limites d'une telle évaluation. En effet, une comparaison fondée sur un seul paramètre ou un cadre uniforme peut favoriser injustement certains algorithmes. Dans cette perspective, cette étude adopte une approche comparative sensible au contexte. Les méthodes de mitigation ont ainsi été testées dans

[10] https://github.com/BiDAlab/FairCVtest
[11] https://github.com/Trusted-AI/AIF360

plusieurs configurations expérimentales (neutre, biais explicite, biais indirect) afin de mieux comprendre leurs effets relatifs selon les conditions.

3.1 Données

Dans cette étude, la base de données FairCVdb[12], introduite par Peña et al. (2020), est utilisée. Elle contient 24 000 profils synthétiques**,** répartis en 80 % pour l'entraînement (19 200 CVs) et 20 % pour la validation (4 800 CVs) avec des variantes dites « aveugles » et « biaisées » permettant une analyse fine des effets du genre et de l'ethnicité dans les processus de décision algorithmique. Les profils sont équilibrés en termes d'attributs démographiques (genre : homme/femme, ethnicité : 3 groupes dont noirs, asiatiques et caucasien), permettant d'étudier l'impact des biais algorithmiques dans un système fictif de recrutement automatisé.

La base est stockée au format NumPy (.npy), optimisé pour une intégration fluide dans les pipelines d'apprentissage automatique. Elle encapsule un dictionnaire Python sérialisé, dont chaque clé correspond à une structure de données (tableau, vecteur, texte ou image) simulant des composantes typiques d'un CV réel. Les profils sont décrits au travers de variables sociodémographiques, éducatives et professionnelles, enrichies par des représentations vectorielles extraites d'images faciales, ainsi que par des éléments textuels (biographies). Les embeddings faciaux sont extraits via ResNet-50, un réseau de neurones convolutif (CNN) pré-entraîné sur ImageNet (Peña et al., 2020).

[12] https://github.com/BiDAlab/FairCVtest/tree/master/data

Le Tableau 2 ci-dessous présente une description détaillée de ces attributs, en précisant pour chacun leur type de variable, leurs modalités de codage ainsi que leurs valeurs possibles. Les variables démographiques comprennent le genre (binaire : homme/femme) et l'origine ethnique (trois catégories codées de 0 à 2), qui constituent les principaux attributs protégés de notre analyse. Les variables professionnelles incluent l'occupation, représentée par dix catégories, ainsi qu'un indicateur ordinal du niveau d'adéquation au poste. Les variables de compétences reflètent le parcours académique et professionnel des candidats (niveau d'éducation, expérience, lettre de recommandation, disponibilité et compétences linguistiques), chacune étant représentée sous forme de variables ordinales discrètes ou binaires. Enfin, les représentations faciales sont encodées à travers des vecteurs continus de 20 dimensions extraits d'un modèle ResNet-50 pré-entraîné, avec une version standard et une version dépourvue d'informations démographiques sensibles. Les variables ordinales (niveau d'éducation, expérience, disponibilité, compétences linguistiques) sont codées sous forme de scores discrétisés entre 0 et 1 (par exemple, 0,2 ; 0,4 ; 0,6 ; 0,8 ; 1), où une valeur plus élevée indique un niveau plus élevé de la caractéristique considérée.

La Figure 6 illustre visuellement ces composantes en montrant les différents blocs d'un CV. Le nombre de croix représente le niveau d'information sensible (+++ = élevé, ++ = moyen, + = faible).

Catégorie	Attributs	Type de variable
Démographique	Origine ethnique	Catégorielle nominale :3 catégories,0 = G1, 1 = G2, 2 = G3
	Genre	Binaire : 0 = homme, 1 = femme
Professionnels	Occupation	Catégorielle nominale :10 catégories codées de 0 à 9,

		Infirmier (0), chirurgien (1), médecin (2), journaliste (3), photographe (4), cinéaste (5), enseignant (6), professeur (7), avocat (8), et comptable (9)
	Niveau d'adéquation au poste	Ordinale (Numérique discrète): 0,25; 0,5; 0,75; 1
Compétences	Niveau d'éducation	Ordinale (Numérique discrète) : 0,2; 0,4; 0,6; 0,8; 1
	Expérience professionnelle	Ordinale (Numérique discrète) :0; 0,2; 0,4; 0,6; 0,8; 1
	Lettre de recommandation	Binaire : 0 = non, 1 = oui
	Disponibilité	Ordinale (Numérique discrète); 0,2; 0,4; 0,6; 0,8; 1
	Compétences linguistiques	Ordinale (Numériques discrètes) :3 langues, codées en: 0, 0,2; 0,4; 0.6, 0,8; 1
Représentations faciales	Face embeddings extraits de ResNet-50	Numérique continue : Vecteurs de 20 dimensions (normalisés)
	Face embeddings sans information de genre/ethnicité	Numérique continue : Vecteurs de 20 dimensions sans information démographique

Tableau 2: Description des attributs FairCVdB des profils

Afin de mieux comprendre la structure du jeu de données avant l'apprentissage, nous avons réalisé une série d'analyses descriptives sur les ensembles train (19 200 profils) et test (4 800 profils). On remarque que la répartition est très équilibrée sur le genre (environ 50 % hommes / 50 % femmes) et sur l'ethnicité (trois groupes autour de 33 % chacun). Les métiers sont aussi relativement bien représentés, chacun se situant entre 8 et 13 % des profils. Pour les variables ordinales, l'adéquation au poste est parfaitement répartie sur ses quatre niveaux (≈ 25 % chacun). L'éducation et l'expérience se concentrent surtout sur les niveaux intermédiaires, tandis que la disponibilité est très majoritairement au maximum (≈ 65 %). La variable lettre de recommandation est en revanche déséquilibrée environ 90 % n'ont pas de lettre et seulement 10 % en ont une. Les compétences linguistiques sont globalement homogènes, avec des moyennes proches de 0,5 pour les trois langues. Les labels neutres (scores objectifs d'adéquation au poste, calculés à partir des embeddings neutres sans biais) sont bien équilibrés leur distribution continue est centrée autour de 0,41 et, après binarisation à la médiane, on obtient

exactement 50 % de positifs et 50 % de négatifs. Dans l'ensemble, le jeu de données présente une structure équilibrée et représentative, offrant de bonnes conditions pour l'apprentissage.

Quatre nouvelles variables ont été ajoutées à la base de données, chacune présentant un degré de corrélation différent (faible à fort) avec la variable protégée « genre », afin de simuler la présence de biais indirects via des variables proxy. Chaque variable a été construite selon la formule :

$$(1)\ \mathbf{X\rho} = \boldsymbol{\rho} \cdot \mathbf{A_{std}} + \sqrt{\mathbf{1} - \boldsymbol{\rho}\mathbf{2}} \cdot \boldsymbol{\varepsilon}$$

Où $\mathbf{A_{std}}$ est la version normalisée de la variable *genre,* $\boldsymbol{\rho} \in \{0.3,0.5,0.7,0.9\}$ contrôle le degré de corrélation visé, et $\boldsymbol{\varepsilon} \sim N(0,1)$ représente un bruit gaussien standard. Cette construction permet d'obtenir des variables artificielles dont la corrélation avec *genre* est précisément calibrée, reproduisant ainsi différents scénarios de biais indirects (faible à très fort).

Cette base de données constitue le cadre expérimental de notre étude, permettant d'analyser l'équité algorithmique des systèmes de recrutement automatisés. L'exploitation de données textuelles, visuelles et structurées permettra d'évaluer les biais algorithmiques et de tester diverses stratégies de mitigation, en utilisant des approches d'apprentissage supervisé.

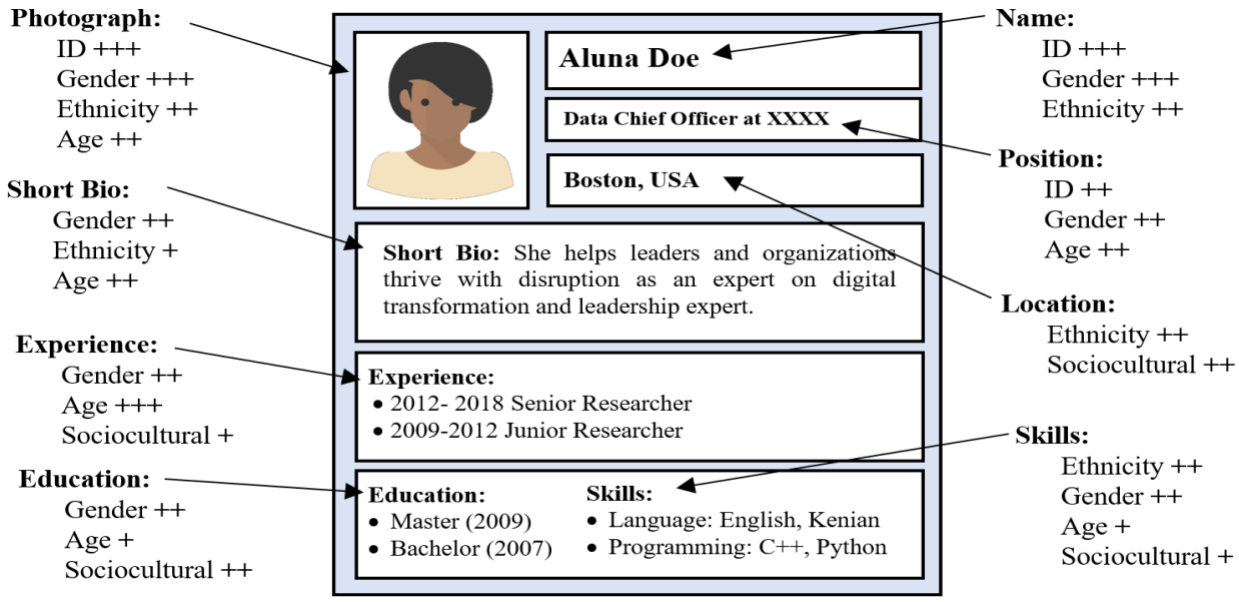

Figure 6: Blocs d'information dans un CV et attributs personnels (Peña et al., 2020)

3.2 Les mesures d'équité (métriques d'évaluation)

Nous nous penchons désormais sur les différentes mesures qui permettent d'évaluer l'équité. Ces mesures sont essentielles pour analyser l'impact des biais et mesurer l'efficacité des stratégies de mitigation dans les systèmes d'IA.

Dans les systèmes d'intelligence artificielle, même un modèle performant peut générer des résultats inéquitables. Par exemple, des disparités peuvent survenir entre les groupes protégés (comme le genre ou la race) et les groupes non protégés. Ces écarts mettent en évidence l'importance de quantifier l'impact des biais et de disposer de mesures précises pour guider les efforts de mitigation.

Plusieurs métriques d'équité ont été développées pour évaluer ces écarts, chacune abordant un aspect spécifique de l'injustice algorithmique.

Dans ce travail, nous avons choisi de nous concentrer uniquement sur les mesures basées sur les groupes, qui permettent de comparer directement les résultats entre différentes catégories démographiques.

3.2.1 Notations et cadre de classification

Avant de présenter les différentes mesures d'équité, nous définissons ci-dessous le cadre de classification et les notations utilisées dans cette étude :

- **Y** désigne la véritable étiquette (ou label réel) d'un individu, où $\boldsymbol{Y = 1}$ indique qu'il est positivement qualifié (par exemple, jugé employable) et $\boldsymbol{Y = 0}$ le contraire.
- $\hat{Y}$ représente la prédiction du modèle pour cet individu, avec $\hat{Y} = \mathbf{1}$ une prédiction positive et $\hat{Y} = \mathbf{0}$ pour une prédiction négative.
- $\boldsymbol{A}$ correspond à l'attribut protégé (par exemple, le genre ou l'origine ethnique). Par convention, nous considérons $\boldsymbol{A = 0}$ pour le groupe de référence (non protégé) et $\boldsymbol{A = 1}$ pour le groupe protégé.
- **Probabilités conditionnelles** : les expressions de la forme $\boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{A = a, Y = y})$ désignent la probabilité qu'un individu appartenant au groupe $\boldsymbol{A = a}$ ayant un label réel $\boldsymbol{Y = y}$ reçoive une prédiction positive.
- **Erreur de classification** : l'événement $\hat{Y} \neq \boldsymbol{Y}$ indique que la prédiction du modèle ne correspond pas à la véritable étiquette.
- **Taux de vrais positifs (TPR)** et **taux de faux positifs (FPR)**, conditionnels à la valeur de l'attribut protégé A=a, sont définis comme suit :

$$\boldsymbol{TPR}_{A=a} = \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{Y} = \mathbf{1}, \boldsymbol{A} = \text{a}) \qquad \boldsymbol{FPR}_{A=a} = \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{Y} = \mathbf{0}, \boldsymbol{A} = \text{a})$$

Ces notations servent de base à la définition des principales mesures d'équité présentées ci-dessous.

3.2.2 Mesures d'équité

Premièrement, nous avons **la différence d'égalité des opportunités** (***equal opportunity difference*****)** ou évalue la différence dans les taux de vrais positifs entre les groupes protégés et non protégés, permettant de quantifier à quel point les opportunités sont inégalement réparties dans les prédictions du modèle(Chen et al., 2023; Mehrabi et al., 2022). Une valeur proche de zéro indique une meilleure équité, tandis qu'un écart important révèle une inégalité d'accès aux prédictions favorables.

$$(2)\ \boldsymbol{EOD} = \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{A} = \mathbf{0}, \boldsymbol{Y} = \mathbf{1}) - \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{A} = \mathbf{1}, \boldsymbol{Y} = \mathbf{1})$$

Ensuite, la **différence de parité statistique (statistical parity difference)** est une mesure utilisée pour évaluer si un modèle accorde les mêmes chances à différents groupes protégés et non protégés de recevoir une prédiction positive, indépendamment de la vraie classe des individus. En d'autres mots, cette mesure quantifie la différence de taux d'acceptation (ou de prédiction favorable) entre les groupes privilégiés (non protégés) et groupes non privilégiés (protégés) (Chen et al., 2023; Mehrabi et al., 2022). Une valeur proche de zéro indique l'équité, alors qu'un écart positif ou négatif signale une disparité dans les prédictions favorables.

$$(3)\ \boldsymbol{SPD} = \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{A} = \mathbf{0}) - \boldsymbol{P}(\hat{Y} = \mathbf{1} \mid \boldsymbol{A} = \mathbf{1})$$

La **différence des taux d'erreur (error rate difference)** est une mesure d'équité qui évalue la différence des taux d'erreur globaux entre les groupes protégés et non protégés. Elle prend en compte à la fois les erreurs de faux positifs et de faux négatifs pour fournir une vue d'ensemble du biais potentiel dans les décisions du modèle. Cette mesure permet de déterminer si le modèle commet proportionnellement plus d'erreurs pour un groupe spécifique par rapport à un autre (Chen et al., 2023). Une valeur proche de zéro indique que le modèle se trompe de façon comparable pour les deux groupes, tandis qu'une valeur élevée signale que l'un des groupes subit davantage d'erreurs de classification.

$$(4)\ \boldsymbol{ERD} = \boldsymbol{P}(\hat{Y} \neq \boldsymbol{Y} \mid \boldsymbol{A} = \boldsymbol{0}) - \boldsymbol{P}(\hat{Y} \neq \boldsymbol{Y} \mid \boldsymbol{A} = \boldsymbol{1})$$

La **différence moyenne des chances (average odds difference**) est une mesure d'équité qui combine les différences de taux de faux positifs et de vrais positifs entre les groupes protégés et non protégés. Elle permet d'évaluer si le modèle traite équitablement ces deux groupes en termes d'erreurs et de prédictions correctes (**Chen et al., 2023).** Une valeur proche de zéro reflète une meilleure équité globale, tandis qu'une valeur plus élevée indique que le modèle accorde des avantages ou commet davantage d'erreurs pour un groupe par rapport à l'autre.

$$(5)\ \boldsymbol{AOD} = \frac{\boldsymbol{1}}{\boldsymbol{2}} * [(\boldsymbol{P}(\hat{Y} = \boldsymbol{1} \mid \boldsymbol{A} = \boldsymbol{0}, \boldsymbol{Y} = \boldsymbol{0}) - \boldsymbol{P}(\hat{Y} = \boldsymbol{1} \mid \boldsymbol{A} = \boldsymbol{1}, \boldsymbol{Y} = \boldsymbol{0}))$$

$$+ (\boldsymbol{P}(\hat{Y} = \boldsymbol{1} \mid \boldsymbol{A} = \boldsymbol{0}, \boldsymbol{Y} = \boldsymbol{1}) - \boldsymbol{P}(\hat{Y} = \boldsymbol{1} \mid \boldsymbol{A} = \boldsymbol{1}, \boldsymbol{Y} = \boldsymbol{1}))]$$

Ou de manière équivalente

$$\boldsymbol{AOD} = \frac{\boldsymbol{1}}{\boldsymbol{2}} * [(\boldsymbol{FPR}_{\boldsymbol{A=0}} - \boldsymbol{FPR}_{\boldsymbol{A=1}}) + (\boldsymbol{TPR}_{\boldsymbol{A=0}} - \boldsymbol{TPR}_{\boldsymbol{A=1}})]$$

La mesure **d'impact disparate** (**disparate Impact**) évalue si un modèle accorde des résultats favorables à des taux proportionnels entre groupes protégés et non protégés. Contrairement à d'autres métriques qui se basent sur des différences absolues, celle-ci s'appuie sur un rapport de probabilités. Un score de 1 indique une parfaite parité, tandis qu'un écart significatif, notamment un score inférieur à 0,8, selon la règle dite des « 80 % » peut révéler une forme de discrimination indirecte. (Feldman et al., 2015)

$$(6)\ \boldsymbol{DI = P\big(\hat{Y} = 1 \mid A = 1\big)/P(\hat{Y} = 1 \mid A = 0)}$$

3.3 Mise en place expérimentale

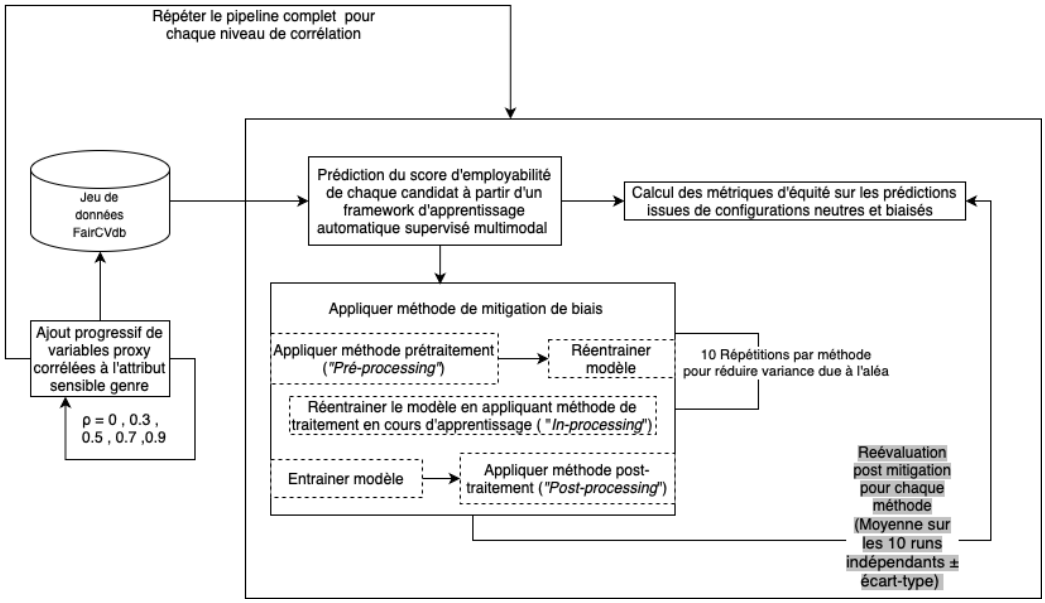

Figure 7: Architecture de notre protocole expérimental

Pour assurer la reproductibilité de notre étude, nous expliquons ici en détail le cadre expérimental mis en place. La Figure 7 présente l'architecture intégrale du protocole expérimental utilisé dans cette étude. Les différentes étapes de ce pipeline seront détaillées dans les sections suivantes.

3.3.1 Manipulation des données

Comme présenté dans la section A l'étude s'appuie sur la base de données *FairCVdb*, développée par Peña et al. (2020). Cette base a été conçue pour étudier les biais algorithmiques dans les systèmes de recrutement automatisé.

Afin d'explorer la sensibilité des modèles aux biais implicites, nous avons enrichi le jeu de données d'origine. Nous y avons ajouté quatre variables additionnelles créées artificiellement. Chacune joue le rôle de *proxy* du genre, avec des niveaux de corrélation différents : 0,3 (X_1), 0,5 (X_2), 0,7 (X_3) et 0,9 (X_4). Ces variables ont été insérées une à une, afin d'observer comment la force de la corrélation avec la variable protégée *genre* influence les prédictions du modèle. Chaque expérience est donc répétée quatre fois, avec une seule variable ajoutée à chaque itération.

La base *FairCVdb* est conçue de façon modulaire. Plusieurs versions, selon les configurations et les clés, permettent d'adapter les données à différents objectifs d'analyse :

- **Neutre *(neutral)*** : cette version emploie les *blind labels* (ou étiquettes aveugles) tant pour l'entraînement que pour le test. Elle représente un scénario idéalisé, où les labels sont supposés exempts d'influences liées aux attributs sensibles. Les

étiquettes sont établies indépendamment du genre ou de l'ethnicité. Elle permet d'évaluer les performances du modèle dans un contexte d'apprentissage dépourvu de biais explicite.

- **Version biaisé genre (*biased gender*)** les labels utilisés dans cette version sont influencés par le genre. Elle simule un biais de décision explicite. Elle permet de tester la réaction du modèle face à des biais directement présents dans les données de sortie.
- **Version biaisé ethnicité (*biased ethnicity*)** : cette configuration suit la même logique, mais avec un biais lié à l'ethnicité. Elle sert à analyser l'impact d'un biais systématique fondé sur l'origine ethnique des individus.

Ces différentes versions de la base permettent de comparer les résultats du modèle dans des situations variées : sans biais explicite, avec biais direct, ou avec biais indirect à travers les variables *proxy*. Dans notre étude, nous travaillons principalement avec les configurations neutres et biaisés genre.

3.3.2 Prédiction du score d'employabilité des candidats à partir du framework FairCvTest

Conformément au protocole décrit précédemment, les prédictions sont effectuées à l'aide du framework *FairCVtest*. La méthode suivie repose sur un processus d'apprentissage supervisé où les données d'entrée X sont associées à une cible y, représentant les scores d'employabilité. L'objectif est d'apprendre une fonction de prédiction $\mathbf{f: X \rightarrow \hat{y}}$, telle que les prédictions $\hat{y} \in [0,1]$ soient les plus proches possibles des scores réels **y.**

Comme expliqué dans la section précédente, les prédictions varient selon la configuration choisie (neutre, biaisée genre ou ethnicité).

En amont, un prétraitement est appliqué aux biographies textuelles des candidat(e)s. Celles-ci sont d'abord nettoyées, puis vectorisées à l'aide d'un *Tokenizer* (Keras). (Mielke et al., 2021; Watson et al., 2024). Chaque biographie est transformée en une séquence d'indices numériques correspondant aux mots, puis encodée sous forme de vecteurs de longueur fixe. Ce traitement garantit une entrée homogène pour le réseau de neurones.

Le modèle de prédiction repose sur une **architecture hybride**, combinant deux sous-réseaux neuronaux :

- **Un réseau textuel** : les biographies vectorisées sont injectées dans une couche LSTM bidirectionnelle (*Long Short-Term Memory*), qui permet de capturer les relations temporelles dans les deux sens de lecture. Cette structure est particulièrement adaptée au traitement de séquences, car elle préserve les dépendances contextuelles entre les mots tout au long du texte.(Graves 2012)
- **Un réseau tabulaire** : les attributs numériques du profil (âge, niveau d'études, domaine, etc.) sont traités via une couche dense (*fully connected*), qui apprend à extraire des représentations pertinentes pour la prédiction **(**Schwing, A. G. et Urtasun, R., 2015).

Les deux représentations (textuelle et structurée) sont ensuite concaténées pour former un vecteur combiné. Celui-ci est transmis à travers plusieurs couches densément connectées

avec activation ReLU, suivies d'un *dropout* pour limiter le surapprentissage, puis d'une couche de sortie sigmoïde (Peña et al., 2020). Cette dernière génère une la prédiction interprétée comme un score d'employabilité.

L'entraînement du modèle repose sur une fonction de perte de type **MAE (Mean Absolute Error)**, définie par la formule suivante :

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - \widehat{yi}|}$$

Où Y_i est le score réel du candidat i, et Ŷ la prédiction correspondante.
Cette mesure pénalise de manière linéaire toute différence entre la valeur prédite et la valeur réelle, sans distinction entre surestimation et sous-estimation. Le choix de la MAE s'inscrit dans une logique de robustesse et d'interprétabilité, car elle est moins sensible aux valeurs extrêmes que l'erreur quadratique moyenne MSE (*Mean Squared Error*) **(**Willmott, C. J et al., 2005).

Le modèle est entraîné sur les données d'apprentissage pendant plusieurs époques (*epochs*). Ainsi, à l'issue de cette phase, il est évalué sur les données de test issues d'un *split* initial, et les prédictions sont générées pour chaque profil.

3.2.3 Protocole expérimental

Afin d'évaluer la robustesse des méthodes de mitigation des biais en apprentissage automatique, notre démarche empirique repose sur une stratégie expérimentale en deux temps. L'objectif est double : valider l'efficacité de ces techniques de mitigation dans un

environnement neutre, puis analyser leur comportement en présence de biais indirects, introduits via des variables corrélées aux attributs protégés.

Dans un premier temps, nous établissons une expérience de référence, qui constitue notre point de comparaison. Celle-ci repose sur la base de données FairCVdb dans sa version originale, sans aucune variable artificiellement corrélée. Le modèle est entraîné et évalué sur ces données, puis les scores d'employabilité prédits sont utilisés pour calculer plusieurs métriques d'équité (telles **l'égalité des Opportunités**, ou **la différence de Parité Démographique**, **la différence Moyenne des Chances, la différence Moyenne des Chances** ou encore **Impact disparate).** Ensuite, les méthodes de mitigation sont appliquées. Ces métriques sont mesurées avant et après la mitigation, ce qui permet d'observer leur efficacité dans un contexte épuré de toute influence indirecte. Ce scénario « idéal » constitue ainsi une base solide pour interpréter les effets des biais ajoutés par la suite.

Dans un second temps, nous procédons à une expérience en contexte biaisé indirectement, en introduisant dans la base de données de nouvelles variables artificielles corrélées aux attributs protégés genre et ethnicité. Les quatre variables corrélées générées sont insérées une à une, et l'expérience est répétée pour chaque niveau de corrélation. Le même protocole est suivi : prédiction des scores, calcul des métriques d'équité, application des méthodes de mitigation, puis réévaluation post-mitigation.

Nous appliquons trois types de méthodes de mitigation : une méthode pré-traitement (Rééquilibrage des poids), une méthode in traitement (Débiasage adversarial) et une méthode post traitement (Classification avec option de rejet). Ainsi notre expérience a été

répétée quinze fois. Nous avons d'abord évalué les scores sur la base originale, avant et après mitigation, puis répété l'exercice avec l'introduction progressive des variables proxy (de X1 à X4), et ce pour chacune des trois méthodes, afin d'analyser l'évolution des effets observés.

Ce dispositif expérimental nous permet de tester deux hypothèses centrales. La première vise à vérifier si les méthodes de mitigation des biais algorithmiques sont efficaces dans un contexte simulé de recrutement. La seconde s'interroge sur la robustesse des méthodes afin de voir si elles conservent-elles leur efficacité lorsque des biais indirects sous forme de variables corrélées présents dans les données. Autrement dit, la présence de variables jouant le rôle de proxies (substituts statistiques aux attributs sensibles) nuit-elle à la capacité des algorithmes de mitigation à produire des résultats véritablement équitables ?

En comparant les performances des modèles dans ces deux contextes sans biais implicite (scénario neutre) puis avec biais implicite (via les proxies) nous cherchons à mieux comprendre la sensibilité des méthodes de mitigation aux corrélations latentes, un phénomène courant mais difficile à détecter dans les systèmes réels.

Afin de renforcer la validité des résultats, l'entraînement du modèle avec les méthodes de prétraitement et d'en traitement a été répété 10 fois avec des germes aléatoires (*random seeds)* différents (de 1 à 10). À chaque répétition, le modèle est réinitialisé, entraîné et évalué, puis les métriques d'équité et de performance sont enregistrées. Les résultats présentés sont la moyenne et l'écart-type calculés sur ces 10 exécutions. Cette répétition permet de mieux capter la variance introduite par l'aléa de l'apprentissage automatique.

La méthode de post-traitement quant à elle, n'étant pas liée à la phase d'apprentissage du modèle, n'a pas nécessité de répétitions multiples.

Le Tableau 3 résume les métriques d'équité utilisées dans cette étude et précise, pour chacune, la valeur cible recherchée. Les quatre premières métriques (SPD, EOD, ERD et AOD) mesurent des écarts de traitement entre groupes et sont donc idéales lorsqu'elles sont proches de zéro. L'Impact disparate, au contraire, repose sur un ratio de taux de sélection et est considéré comme acceptable lorsqu'il reste proche de 1, conformément à la règle des 80 %. Enfin, l'exactitude globale du modèle est rapportée pour apprécier les compromis éventuels entre performance prédictive et réduction des biais.

Métrique	Description	Valeur idéales
Différence de Parité Démographique (SPD) (3)	Évalue la différence de taux de sélection (prédictions positives) entre groupes protégés et non protégés.	Proche de 0
Différence d'égalité des opportunités (EOD) (2)	Mesure l'écart entre les taux de vrais positifs (TPR) des différents groupes. Indique l'équité dans la reconnaissance des candidats méritants.	Proche de 0
Différence de taux d'erreur (ERD) (4)	Reflète l'écart global entre les taux d'erreurs (faux positifs et faux négatifs combinés) selon les groupes.	Proche de 0
Différence moyenne des chances (AOD) (5)	Mesure la moyenne de la différence des taux de faux positifs (FPR) et de vrais positifs (TPR) entre groupes.	Proche de 0
Disparate Impact (DI) (6)	Ratio entre les taux de sélection des groupes. Une valeur éloignée de 1 peut indiquer une discrimination indirecte.	Proche de 1 (accepté entre 0.8 et 1.25) selon la règle de 80% (Feldman et al., 2015)
Exactitude (Accuracy)	Indique la performance globale du modèle, indépendamment de toute notion d'équité.	Plus élevé = meilleure performance

Tableau 3: Description et valeurs idéales des métriques

Chapitre 4 : Résultats empiriques

Nous répondons à deux questions principales dans notre expérience.

4.1 R.Q.1. Évaluation de l'efficacité des méthodes de mitigation dans un contexte de biais explicite

Cette section vise à analyser et valider l'efficacité des méthodes de mitigation des biais algorithmiques dans un contexte de biais explicite. Conformément au protocole expérimental décrit précédemment, un modèle d'apprentissage automatique a été entraîné pour prédire les scores d'employabilité des candidats à partir du framework FairCVtest, dans un scénario simulé de recrutement.

Trois familles de techniques de mitigation ont ensuite été appliquées sur un modèle biaisé : une méthode de prétraitement, une méthode en cours d'apprentissage, et une méthode de post-traitement. Pour chacune d'elles, les métriques d'équité ont été calculées avant et après application, afin de mesurer leur impact sur les disparités entre groupes protégés et non protégés.

L'objectif est de déterminer si ces approches permettent effectivement de réduire les écarts de traitement liés à l'attribut sensible (ici le genre), tout en préservant une qualité de prédiction acceptable.

RQ1.1 : Les méthodes de mitigation des biais algorithmiques permettent-elles de réduire les inégalités entre groupes protégés et non protégés, et de restaurer l'équité algorithmique ?

Avant d'apporter une réponse à cette question, nous procédons d'abord à une analyse de l'état initial de l'équité dans notre modèle. Cette première étape vise à établir une base de référence à partir des prédictions issues des configurations neutres et biaisées, sans application de techniques de mitigation. Les résultats (Tableau 4) permettent de mesurer les écarts d'équité présents avant toute intervention.

a. Calcul des métriques avant application des méthodes de mitigation

Configuration	SPD	EOD	ERD	AOD	DI	Exactitude
Neutre (Baseline)	0.00004	0.0193	-0.0150	0.0043	1.0008	0.9187
Biaisé genre	-0.3087	-0.2954	0.0063	-0.3017	0.5313	0.8258

Tableau 4: Évaluation initiale des biais - Comparaison des métriques d'équité avant mitigation pour les configurations neutre vs biaisée (genre)

Le Tableau 4 met en évidence les effets du biais dans les deux configurations testées, ainsi que la validité des hypothèses initiales. Dans la configuration neutre, les métriques sont proches des valeurs idéales. La distribution des prédictions y est globalement équitable entre groupes protégés et non protégés. En revanche, dans la configuration biaisée par le genre, les indicateurs révèlent des écarts significatifs, au détriment des femmes.

- La différence de parité statistique (SPD) passe de 0,00004 à -0,3087. Cela signifie que les femmes ont environ 30 % de chances en moins d'être sélectionnées que les hommes, à profil égal.
- La différence d'égalité des chances (EOD) chute de 0,0193 à -0,2954, indiquant qu'une bonne candidate a près de 30 % de chances en moins d'être reconnue comme telle par le modèle.
- L'impact disparate (DI) tombe de 1,00 à 0,53, bien en dessous du seuil de 0,80 couramment admis. Ce qui veut dire que les femmes ont été sélectionnées 47 % moins souvent que les hommes.
- La différence moyenne des chances (AOD) passe de 0,0043 à -0,3017, confirmant un écart fort entre les groupes. Cela confirme un déséquilibre important dans les décisions du modèle.
- Enfin, la précision du modèle diminue légèrement, de 91,87 % à 82,58 %.

Les écarts d'équité mettent en lumière un traitement inégal et potentiellement discriminatoire envers les candidates issues du groupe protégé (femmes).

b. Évaluation de l'impact des méthodes de mitigation sur les biais algorithmiques

Après avoir établi la présence de biais significatifs dans la configuration biaisée par le genre, nous examinons l'efficacité de trois techniques de mitigation appliquées à cette base : une méthode de pré-traitement *(Reweighing*), une méthode en cours d'apprentissage (*Adversarial Debiasing*), et une méthode de post-traitement (*Reject Option Classifier*).

On essaie non seulement d'évaluer la capacité de chaque méthode à corriger les disparités identifiées mais aussi, d'analyser leur impact sur les performances globales du modèle.

Métriques	Avant mitigation	Pré-processing	In processing	Post-processing
SPD	-0.3087	-0.2316 ± 0.0180	-0.2014 ± 0.0304	-0.1159
EOD	-0.2954	-0.1896 ± 0.0178	-0.2021 ± 0.0260	-0.0034
ERD	0.0063	0.0055 ± 0.0070	-0.0030 ± 0.0083	-0.1080
AOD	-0.3017	-0.2545 ± 0.0182	0.1989 ± 0.0304	-0.1153
DI	0.5313	0.6929 ± 0.0204	0.6573 ± 0.0389	0.8826
Exactitude	0.8258	0.8425 ± 0.0060	0.7761 ± 0.0053	0.5681

Tableau 5: Résultat des métriques après mitigation

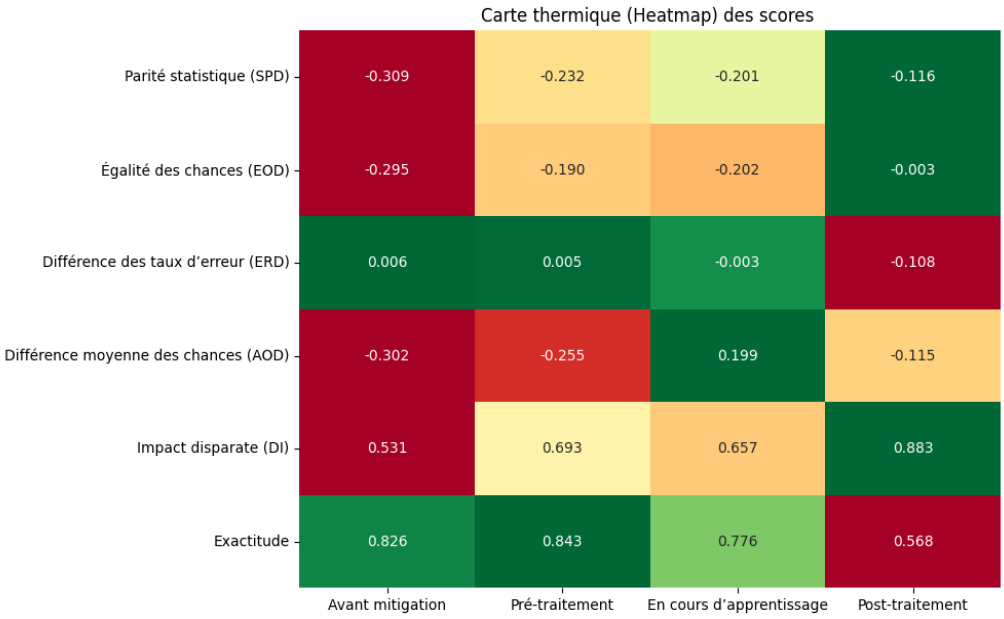

Figure 8: Comparaison des métriques d'équité par méthode de mitigation

Le Tableau 5 reprend les valeurs numériques des différentes métriques d'équité et de performance avant et après mitigation. Il permet de préciser les écarts mis en évidence visuellement par la Figure 8 qui présente une carte thermique (heatmap) des différentes

métriques d'équité et de performance avant et après application des méthodes de mitigation. Chaque ligne correspond à une métrique (parité statistique, égalité des chances, différence des taux d'erreur, différence moyenne des chances, impact disparate, exactitude) et chaque colonne à une étape (avant mitigation, pré-traitement, in-processing, post-traitement). Les couleurs indiquent l'intensité du biais ou de la performance : le rouge correspond aux valeurs les moins favorables, l'orange et le jaune à des valeurs intermédiaires et le vert à des résultats plus équilibrés.

Les résultats le montrent sans équivoque : appliquer des techniques de mitigation permet de réduire les biais entre les groupes. On observe que toutes les méthodes de mitigation appliquées sur la version biaisée du jeu de données contribuent à une amélioration des scores d'équité, quoique de manière inégale. Dans la version biaisée par le genre, les écarts étaient marqués. Mais après application des trois approches (pré-traitement, en cours d'apprentissage, post-traitement), on observe une nette amélioration des indicateurs d'équité.

- La différence de parité statistique (SPD) passe de –0,3087 à des valeurs comprises entre –0,23 et –0,11 selon la méthode. Cela reflète une meilleure répartition des opportunités entre groupes protégés et non protégés.
- La différence d'égalité des chances (EOD), qui mesure l'équité dans la reconnaissance des bons candidats, s'améliore nettement, surtout avec le post-traitement où elle atteint –0,0034 qui est une quasi-parité.
- L'impact disparate (DI), qui était tombé à 0,53, remonte jusqu'à 0,88 selon la méthode. Il franchit ainsi le seuil de 0,80.

- La différence moyenne des chances (AOD) confirme également cette tendance. Elle passe de –0,3017 à une valeur bien plus proches de zéro, selon les cas.

En somme, toutes les méthodes réduisent les écarts entre groupes protégés et non protégés. Les biais sont atténués et le système devient globalement plus juste dans ses décisions. Bien que l'ampleur de l'effet varie selon la méthode, toutes contribuent à restaurer, ne serait-ce que partiellement, une forme d'équité algorithmique.

4.2 RQ2. Évaluation de l'efficacité des méthodes de mitigation dans un contexte de biais implicite (biais indirect via des proxies corrélées à l'attribut protégé)

La notion de biais dans les systèmes d'apprentissage automatique n'est ni linéaire, ni triviale. Elle reflète des dynamiques sociétales complexes et renvoie à un enchevêtrement de facteurs techniques, sociaux et historiques (comme discuté dans la section 2.2, chapitre 2).

Dans les systèmes déployés en conditions réelles, les biais ne sont pas toujours visibles. Ils ne se réduisent pas à la simple présence d'attributs sensibles comme le genre ou l'ethnicité. Ils peuvent émerger de manière diffuse, à travers des variables indirectement corrélées aux attributs protégés. Ces proxies agissent comme des relais invisibles des inégalités. Par exemple, le code postal est souvent identifié comme un indicateur indirect de la race aux États-Unis, car il reflète des ségrégations résidentielles héritées (Barocas et Selbst, 2016 ; Suresh et Guttag, 2021). Ainsi, même lorsque l'attribut sensible est supprimé, un modèle peut continuer à produire des décisions discriminatoires. Plusieurs travaux ont souligné que des variables apparemment neutres, mais fortement corrélées à

un attribut protégé, peuvent induire des biais d'intensité comparable à ceux directement causés par cet attribut (Zafar et al., 2017).

Dans cette optique, nous avons simulé un scénario plus réaliste, dans lequel les biais ne sont pas explicitement présents, mais induits de manière implicite. Pour ce faire, une variable proxy artificielle, corrélée à l'attribut « genre », a été introduite dans notre base de données. Quatre niveaux de corrélation ont été testés : faible (0.3), modéré (0.5), élevé (0.7) et très élevé (0.9). Nous avons ensuite appliqué le même protocole expérimental que pour la RQ1, en mobilisant les trois grandes familles de méthodes de mitigation.

L'objectif est d'évaluer si ces méthodes demeurent efficaces lorsque les discriminations sont plus subtiles, dissimulées dans des structures corrélées, et donc plus difficiles à détecter.

R.Q.2.1. Quelle est la robustesse des méthodes de mitigation face à des biais indirects, induits par des proxies corrélés à un attribut sensible ?

Pour répondre à cette question, nous avons observé l'évolution des performances des trois méthodes de mitigation lorsque le biais devient plus diffus, dissimulé dans des variables corrélées à l'attribut protégé. Cette configuration vise à reproduire les formes de biais implicites fréquemment rencontrées dans les systèmes d'IA en conditions réelles.

Afin de rendre ces dynamiques visibles, nous présentons une série de six figures (Figure 9, Figure 10, Figure 11, Figure 12, Figure 13, Figure 14). Chacune illustre l'évolution d'une métrique d'équité : **écart de parité statistique, écart d'égalité des chances, écart moyen des chances, impact disparate, écart de taux d'erreur** et **précision** (Tableau 3 pour une description détaillée des métriques).

Chaque figure distingue trois contextes expérimentaux :

- Avant toute mitigation, dans un environnement neutre, sans variable corrélée,
- Après mitigation, toujours dans un environnement neutre,
- Après mitigation, dans des contextes biaisés, où la corrélation entre la variable proxy et le genre varie progressivement (0.3, 0.5, 0.7 et 0.9), selon les trois approches testées.

Ces représentations graphiques permettent de suivre :

- La dégradation graduelle de l'équité provoquée par l'introduction de biais latents,
- La capacité des méthodes à compenser ces distorsions,
- Les différences de comportement observées entre les métriques elles-mêmes.

Elles offrent une lecture synthétique, mais nuancée, de la robustesse des approches face à des biais invisibles mais persistants.

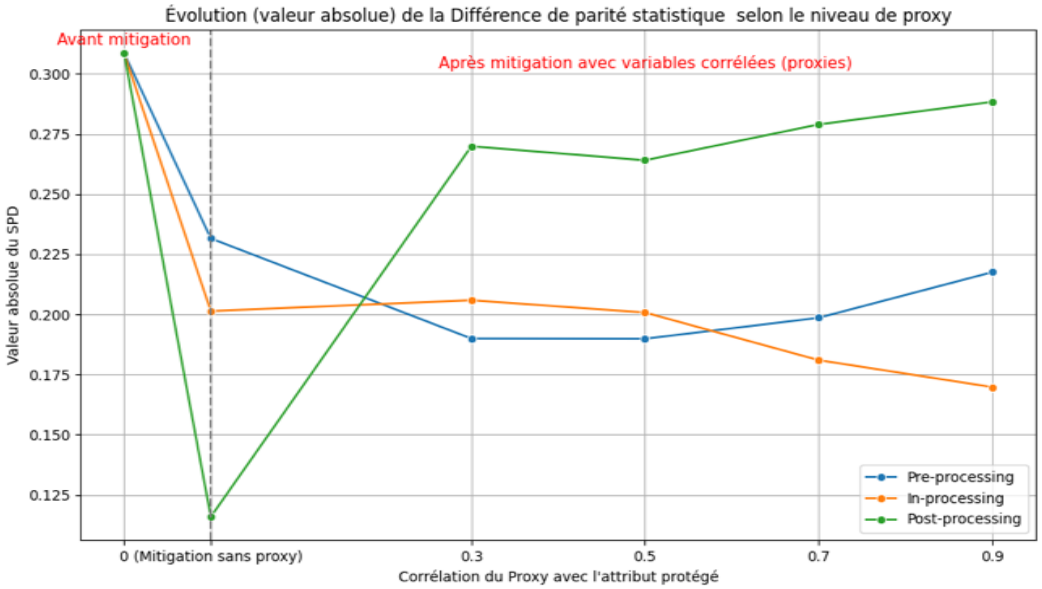

Figure 9: Évolution de la différence de parité statistique (SPD) selon le niveau de proxy et les méthodes de mitigation

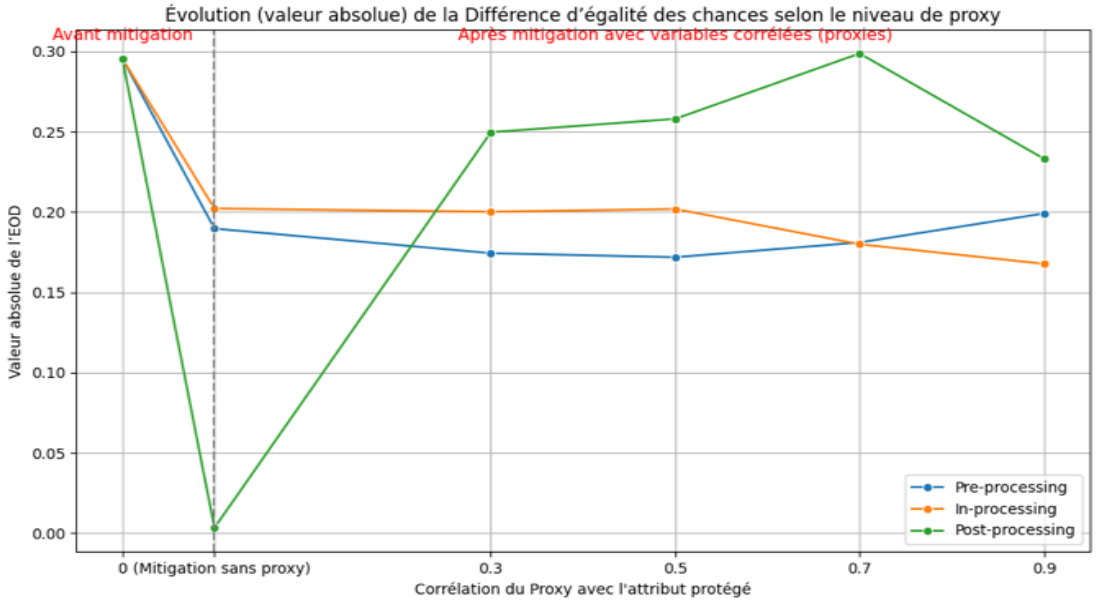

Figure 10: Évolution de la différence d'égalité des chances (EOD) selon le niveau de proxy et les méthodes de mitigation

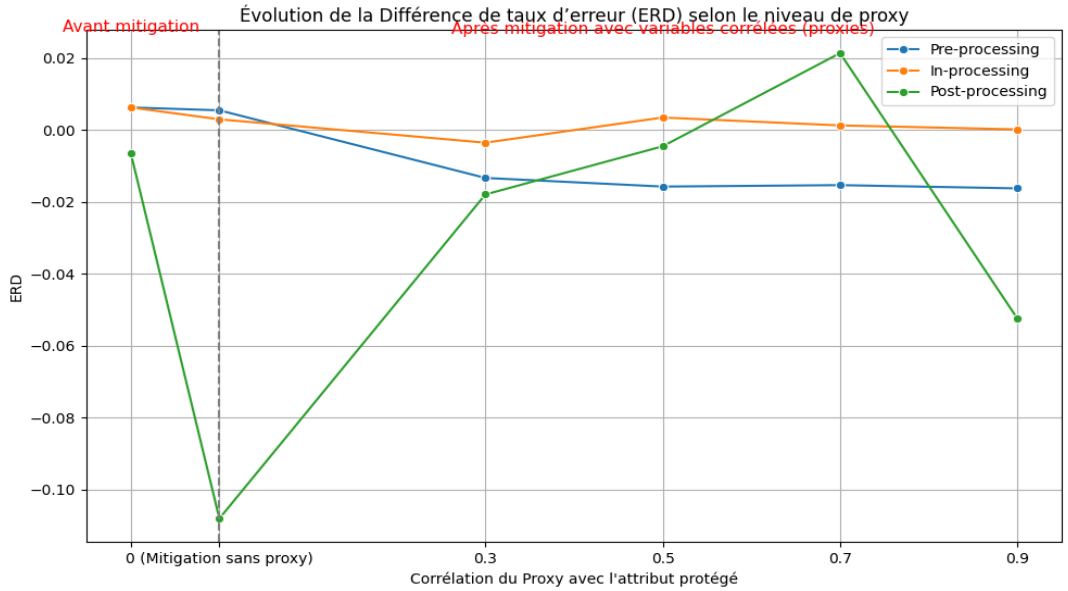

Figure 11: Évolution de la différence de taux d'erreur (ERD) selon le niveau de proxy et les méthodes de mitigation

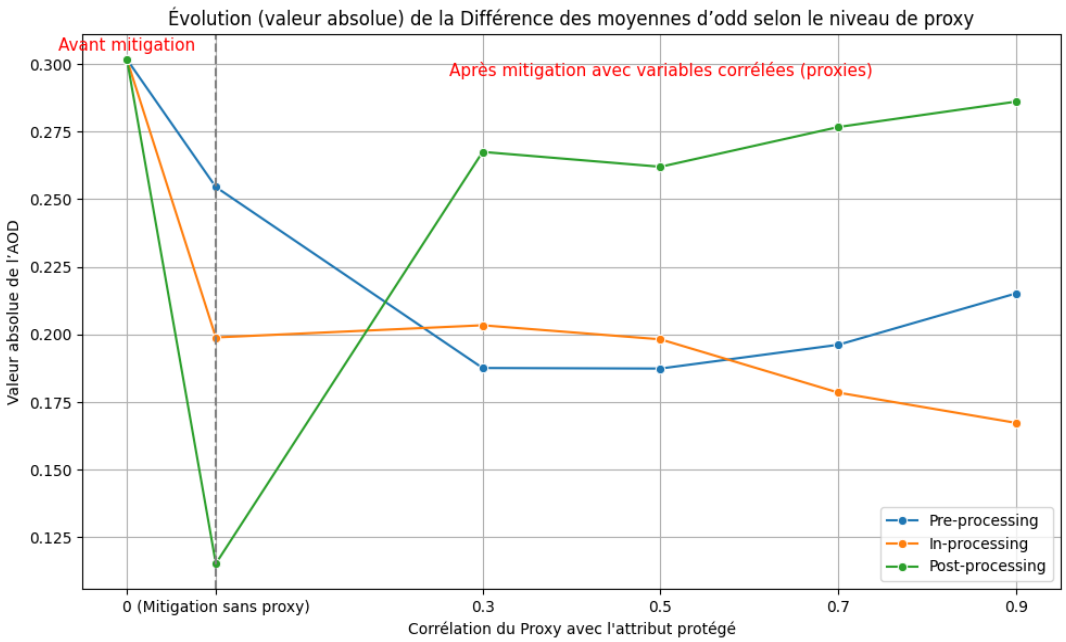

Figure 12: Évolution de la différence des moyennes des chances (AOD) selon le niveau de proxy et les méthodes de mitigation

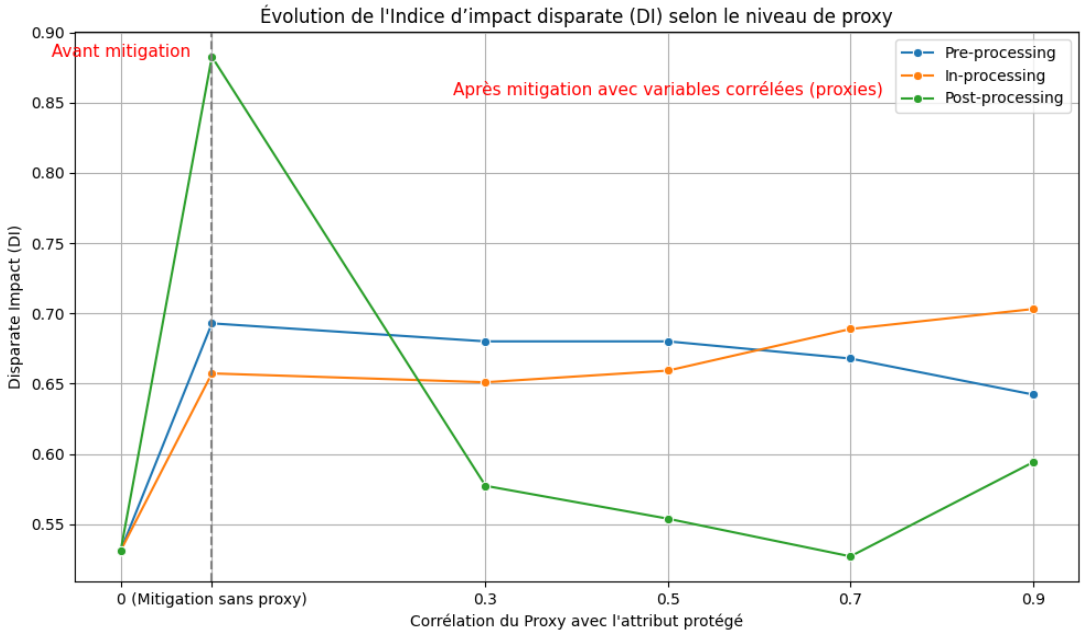

Figure 13: Évolution de la différence de l'indice d'impact disparate (DI) selon le niveau de proxy et les méthodes de mitigation

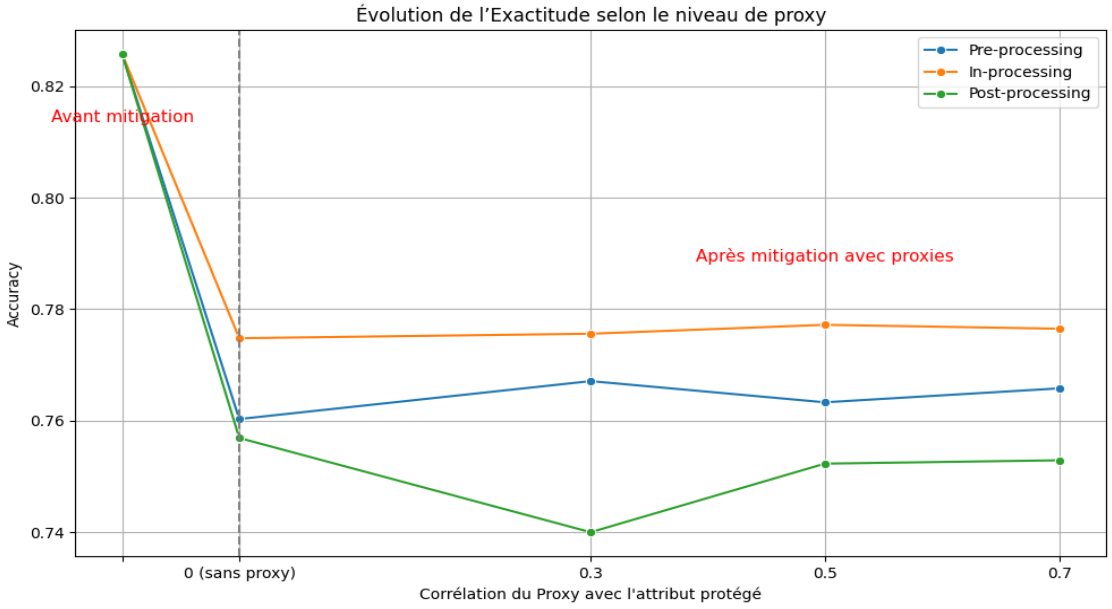

Figure 14: Évolution de l'exactitude selon le niveau de proxy et les méthodes de mitigation

Les figures 9 à 14 montrent que l'effet de l'introduction de variables proxies varie selon la métrique et selon le niveau de corrélation (0,3, 0,5, 0,7, 0,9). On observe cependant un schéma constant. Pour chaque métrique et à chaque niveau de corrélation, les méthodes de pré-traitement, in-traitement et post-traitement conduisent à des résultats meilleurs que la situation initiale avant mitigation.

Pour la parité statistique (SPD), la valeur absolue passe d'environ 0,31 avant mitigation à des valeurs toutes inférieures à cela, peu importe le niveau de corrélation. L'égalité des chances (EOD), qui était proche de –0,30, s'améliore nettement et descend à des valeurs bien plus faibles et ne dépassant plus ce niveau, même avec les proxies. La différence des taux d'erreur (ERD) reste proche de zéro après mitigation, même avec les proxies, avec une seule valeur extrême à la corrélation 0,7. La différence moyenne

des chances (AOD), initialement à 0,30, est réduite jusqu'à 0,18 et ne dépasse pas la valeur avant mitigation, même avec proxies. L'indice d'impact disparate (DI), tombé à 0,53 avant correction, remonte jusqu'à 0,70 après application des méthodes. Même si elle ne franchit pas le seuil de 0,80, on note une amélioration dépendamment du niveau de corrélation. Bien que l'exactitude chute après mitigation, un phénomène attendu à cause du compromis équité–performance[13] (*fairness–performance trade-off*), elle reste globalement stable et chute davantage avec le post-traitement.

Ces résultats mettent en évidence que, malgré la dégradation introduite par les proxies, les trois approches conservent une certaine robustesse et permettent d'atténuer significativement les écarts entre groupes protégés et non protégés.

[13] **Compromis équité–performance** : La communauté scientifique reconnaît de façon générale que les méthodes de mitigation des biais tendent à améliorer l'équité, mais souvent au détriment de la performance des modèles d'apprentissage automatique, notamment en termes d'exactitude. Ce phénomène est couramment désigné comme le compromis équité–performance (*fairness–performance trade-off*) (Chen et al., 2023).

Chapitre 5 : Discussion et limites

5.1 Discussion

Ce chapitre présente la discussion de notre expérience présentée au chapitre 3 ainsi que nos résultats du chapitre 4. Nous abordons également les limites identifiées.

Les résultats obtenus montrent qu'aucune méthode n'est parfaite. Chaque approche permet de réduire certains biais, mais avec des effets variables selon la métrique considérée. Notre objectif n'est pas d'élire une « meilleure méthode » de manière absolue, mais de comprendre les forces et les limites de chacune. En effet, il n'existe pas de consensus sur une métrique unique pour évaluer l'équité algorithmique. L'efficacité d'une méthode de mitigation dépend largement du contexte d'utilisation, des objectifs poursuivis, de la notion d'équité que l'on choisit d'adopter, ainsi que des contraintes techniques propres au système.

Comme l'ont souligné plusieurs travaux (Barocas et al., 2023; Pleiss et al., 2017; Verma et Rubin, 2018), la question de savoir si un classifieur est équitable n'a pas de réponse universelle. Elle repose sur les valeurs que l'on privilégie dans un système : souhaite-t-on avant tout préserver la performance prédictive, garantir une stricte égalité d'accès, ou corriger des inégalités historiques ? À chaque conception de l'équité correspond une métrique, mais celles-ci sont souvent incompatibles entre elles (Pleiss et al., 2017).

Cependant, nos constats confirment que *les méthodes de mitigation sont efficaces pour réduire les biais dans les systèmes d'apprentissage automatique*, même face à l'introduction progressive de biais implicites via des variables proxies. En effet, pour la majorité des métriques d'équité, les scores obtenus après mitigation (même en présence de proxies fortement corrélés) demeurent meilleurs que ceux mesurés sur le jeu de données initial, sans aucune correction.

Dans la question de recherche RQ1, nous avions observé que les trois grandes approches de mitigation parvenaient toutes à réduire les inégalités de traitement dans un contexte de biais explicite. Cette tendance se maintient même en présence de proxy, bien que de façon plus nuancée. Les valeurs tendent à se détériorer lorsqu'on ajoute les variables corrélées à l'attribut sensible. Cependant, la dégradation n'est ni linéaire ni uniforme et l'impact du biais implicite dépend à la fois de la métrique considérée et de la méthode appliquée. Les résultats obtenus ne suivent pas toujours un schéma simple ou prévisible (Friedler et al., 2018; Mehrabi et al., 2022) . Nos résultats empiriques confirment plusieurs tendances observées dans la littérature concernant l'efficacité relative des différentes approches de mitigation des biais.

Premièrement, la méthode de pré-traitement apparaît comme un équilibre pertinent entre équité et performance, que ce soit en présence de biais explicites ou implicites. Pour toutes les métriques analysées, les scores après mitigation demeurent meilleurs que ceux obtenus sans intervention corrective, même lorsque des variables corrélées (proxies) sont introduites dans les données. Ce type d'approche est particulièrement pertinent dans les contextes opérationnels où il est essentiel de réduire les biais sans compromettre la

précision des prédictions. Contrairement aux méthodes plus intrusives, Reweighing agit en amont avant l'entraînement du modèle, ce qui facilite son intégration dans les pipelines existants. Selon Bellamy et al. (2018), ce type de méthode constitue souvent une solution pragmatique pour les organisations, car elle permet d'atteindre un meilleur équilibre entre équité et performance, surtout dans les environnements où la fiabilité des modèles est aussi cruciale que leur équité. Par ailleurs, un phénomène particulièrement surprenant émerge. Pour certains niveaux de corrélation (notamment faibles à modérés) les performances post-mitigation avec proxy surpassent parfois même celles obtenues sans proxy. En d'autres termes, ajouter un biais implicite léger semble améliorer l'équité après traitement. Un paradoxe en apparence, mais qui s'explique. En effet, l'introduction de variables corrélées à l'attribut protégé peut agir comme un signal faible mais exploitable. Ce signal rend les disparités plus visibles aux yeux du modèle, facilitant ainsi l'identification des biais et leur correction. Ce constat rejoint les travaux de Gupta et al. (2018) sur le concept de « proxy fairness ». Les auteurs démontrent que, dans certaines conditions, des groupes proxy (bien que distincts des groupes protégés initiaux) peuvent favoriser l'équité en reproduisant des contraintes similaires à celles des groupes protégés réels. Cette visibilité accrue permet au modèle de mieux ajuster ses paramètres en fonction des inégalités sous-jacentes, surtout lorsque ces biais étaient latents ou faiblement exprimés dans les données originales.

La méthode in-processing se distingue par sa stabilité. Pour l'ensemble des métriques sensibles à l'équité, ses résultats varient peu après mitigation que ce soit en présence ou en absence de proxy. Elle fait preuve d'une résilience constante, même lorsque la corrélation entre le proxy et l'attribut protégé devient élevée. Contrairement aux

approches de pré et post-traitement, ses performances en matière d'équité se dégradent très peu. Cette robustesse s'explique par la conception même de la méthode car le modèle apprend à prédire tout en empêchant la déduction de l'attribut protégé à partir de ses prédictions. Il corrige ainsi le biais à la source, dès la phase d'apprentissage (Zhang et al., 2018). On observe aussi que lorsque le niveau de corrélation est très élevé (par exemple p = 0,9), certaines métriques d'équité s'améliorent. Ce phénomène, bien que contre-intuitif, peut s'expliquer par une visibilité accrue du biais. En effet, plus le signal de discrimination est fort, plus le modèle (et notamment le classifieur adversaire dans ce cas) est en mesure de le détecter et de le neutraliser. À l'inverse, un proxy faiblement corrélé peut brouiller ce signal, rendant le biais plus difficile à atténuer.

La méthode post-processing se révèle être la plus sensible à l'introduction de biais implicites. Avant l'introduction de variable proxies, dans le RQ1, on observe qu'elle corrige les biais de manière particulièrement efficace. Lorsque le biais est explicite, elle affiche des scores d'équité très faibles, bien qu'au prix d'un compromis notable sur la performance. Ce constat rejoint les conclusions de Barocas et al., (2023) qui soulignent la capacité des approches post-traitement à imposer des contraintes d'équité sur le résultat, permettant une parité stricte au niveau du groupe. Toutefois, cette efficacité se fait au détriment de la performance globale et une chute considérable de l'exactitude. Ainsi, bien qu'il améliore considérablement l'équité, il peut entraîner une réduction significative de la précision. Les méthodes post-traitement comme le ROC sont donc souvent inadaptées aux environnements où la précision du modèle est prioritaire (Bellamy et al., 2018). De surcroît, l'efficacité de la méthode ROC s'effondre dès que des variables corrélées sont introduites. Cette fragilité est particulièrement visible dans nos résultats. Plusieurs

métriques d'équité se détériorent rapidement dès une corrélation modérée (p = 0,3), et les écarts continuent de s'amplifier à mesure que la corrélation augmente. La correction effectuée a posteriori ne suffit manifestement plus à compenser l'impact des biais implicites, surtout lorsqu'ils émergent de liens complexes et indirects au sein des données. Cela s'explique en grande partie par la nature même de cette méthode. Le post-traitement agit exclusivement sur les sorties du modèle, sans accéder ni à son architecture interne, ni aux dynamiques de son entraînement. Il ajuste les prédictions finales, mais reste aveugle à la manière dont les biais ont été intégrés ou reproduits en amont. Autrement dit, il corrige les symptômes sans traiter les causes. Dès lors, plus les biais deviennent subtils, ou masqués (comme c'est le cas ici avec les proxies), plus cette approche révèle ses limites. Elle fonctionne dans des contextes où les discriminations sont visibles et explicites, mais perd en robustesse dès que la complexité structurelle des données augmente. Ces constats rejoignent les observations de Bellamy et al. (2018), qui soulignent que les méthodes de post-traitement, bien que simples à mettre en œuvre et utiles en contexte de modèle "boîte noire", s'attaquent essentiellement aux prédictions biaisées, sans intervenir sur les causes profondes du biais. En d'autres termes, elles corrigent les symptômes, mais laissent intacte la structure sous-jacente des données et du modèle. Cela les rend moins adaptées à des contextes où les biais sont diffus, implicites ou enracinés dans la corrélation entre variables. À contrario, les méthodes qui interviennent plus en amont (comme le pré-traitement) permettent d'agir à la source du déséquilibre, et sont considérées comme des options plus robustes lorsque les données sont accessibles.

En conclusion, notre étude montre clairement que les méthodes de mitigation permettent de réduire les écarts entre groupes, même si leur efficacité varie selon les métriques et le contexte. Ils révèlent également que, si certaines approches apparaissent plus robustes face aux biais implicites, aucune ne garantit une équité parfaite. Cette observation rejoint la littérature, qui rappelle que l'équité algorithmique n'est pas un état absolu mais une construction dépendante des métriques retenues et des valeurs privilégiées dans un système.

Cette étude contribue ainsi à la recherche en apportant une validation empirique dans un cadre simulé de recrutement, en soulignant notamment le rôle des variables proxy dans la persistance des biais.

Ces résultats renforcent enfin l'importance d'intégrer ces techniques qui donnent résultats concrets, combinant cadres théoriques et expérimentations concrètes, afin de développer des systèmes d'IA plus responsables et équitables.

5.2 Limites de l'étude

Cette étude présente certaines limites. La première concerne l'inaccessibilité aux systèmes décisionnels réels basés sur l'intelligence artificielle. En effet, notre objectif principal était d'analyser les biais présents dans ce type de systèmes et d'en évaluer l'efficacité des méthodes de mitigation. Toutefois, les codes sources et les données utilisées par ces algorithmes sont rarement accessibles, comme le remarquent également plusieurs autres travaux (Chen et al., 2023 ; Semmelrock et al., 2025). Cette limite découle de contraintes juridiques, éthiques, de confidentialité, mais aussi de propriété

intellectuelle. Il convient également de prendre en compte que les entreprises concernées sont réticentes à ouvrir leurs systèmes à l'analyse, de crainte d'exposer d'éventuelles failles ou pratiques discriminatoires. Face à ces contraintes, nous avons donc dû nous appuyer sur des environnements simulés et sur des bibliothèques spécialisées qui nous ont permis de reproduire certains comportements des systèmes d'IA, mais sans en refléter pleinement la complexité.

Une deuxième limite découle également de ce manque d'accès. Quoique la base FairCVdb ait été conçue pour refléter des situations crédibles de recrutement, elle repose sur des données synthétiques. Ainsi, le jeu de données ne reflète pas totalement les subtilités du monde réel, notamment les dynamiques organisationnelles, les comportements humains ou les biais contextuels qui émergent dans des processus d'embauche authentiques. Par conséquent, bien que ces choix favorisent la reproductibilité des expériences, ils peuvent toutefois limiter la représentativité des résultats.

La troisième limite a trait aux méthodes testées. Nous avons testé une seule méthode de mitigation par type (prétraitement, en cours d'apprentissage, post-traitement). Ce choix a été motivé par une volonté de clarté et de comparabilité. Même s'il nous a permis de structurer et d'effectuer une analyse comparative de différents types de méthodes, il restreint la généralisation des résultats. D'autres techniques, même appartenant à la même catégorie, pourraient réagir différemment face à certains types de biais. De plus, le nombre de scénarios testés reste limité : seule une forme de biais indirect a été introduite, via des variables artificiellement corrélées aux attributs protégés.

Par ailleurs, le choix des métriques d'équité utilisées dans cette étude (comme la différence de parité statistique, la différence d'égalité des chances, la différence de taux d'erreur, la différence moyenne des chances et l'impact disparate) repose sur une certaine vision de ce qu'est une décision juste en intelligence artificielle. Ces métriques sont largement reconnues dans la littérature, mais elles ne couvrent pas toutes les dimensions possibles de l'équité. Par exemple, elles ne prennent pas en compte des facteurs plus complexes comme les discriminations croisées (liées à plusieurs critères à la fois), les inégalités de pouvoir ou les biais ancrés dans les institutions. D'autres indicateurs, notamment ceux prenant en compte les dynamiques intersectionnelles, les asymétries de pouvoir ou les biais institutionnels, pourraient compléter l'analyse et offrir une compréhension plus nuancée des biais (Corbett-Davies et al., 2023).

Enfin, la dernière limite concerne le caractère ponctuel de l'évaluation, c'est-à-dire qu'elle a été réalisée à un moment donné, sans tenir compte de l'évolution possible des systèmes dans le temps. En réalité, les biais peuvent apparaître progressivement ou s'aggraver avec l'accumulation de données et l'adaptation des modèles (Davis et al., 2025) .Une approche longitudinale aurait permis de mieux capter ces dérives invisibles à court terme.

Chapitre 6 : Conclusion et futures directions

Notre recherche s'est penchée sur les enjeux d'équité et de biais algorithmiques, en particulier dans les systèmes d'intelligence artificielle appliqués au recrutement. Nous avons tenté de répondre à une question simple : est-il possible de rendre les outils de décision automatisés par l'apprentissage automatique plus équitables ?

En nous concentrant sur les biais liés au genre et à l'origine, l'étude cherchait à y voir plus clair. L'investigation s'est nourrie d'allers-retours constants entre hypothèses de travail et vérifications empiriques. Nous avons exploré des pistes, testé des méthodes et confronté nos intuitions aux résultats. Pour ce faire, une approche hybride a été retenue, combinant une revue approfondie des biais genrés et raciaux avec une évaluation empirique de plusieurs méthodes de mitigation. L'objectif n'était pas seulement d'identifier les dérives possibles, mais aussi de tester, concrètement, des façons d'y remédier.

Mais au fond, ce travail n'était pas qu'un exercice technique. Il s'agissait aussi de comprendre comment les outils, parfois à leur insu, peuvent reproduire ou même renforcer des inégalités déjà profondément ancrées dans nos sociétés. L'intérêt s'est ensuite orienté vers les moyens possibles de les réduire. Au-delà des solutions algorithmiques, l'analyse s'est élargie à des cadres conceptuels permettant de mieux comprendre les mécanismes de reproduction des inégalités. En croisant théorie et expérimentation, nous avons cherché à élaborer une démarche rigoureuse tout en l'ancrant dans des réalités concrètes.

Malgré notre volonté d'adopter une approche interdisciplinaire, mêlant technologie, société, éthique et humains, notre étude reste perfectible. Une recherche menée en collaboration avec des experts aux profils variés (en data science, droit, sciences sociales etc.) permettrait d'aller plus loin. Il serait également pertinent d'élargir les méthodes testées, afin de mieux comprendre leurs limites et complémentarités. Et l'accès, même partiel ou anonymisé, à des données réelles renforcerait la validité et l'applicabilité des résultats.

Autre élément essentiel : le temps. Certains biais ne sont visibles qu'après coup. Ils s'installent lentement et évoluent avec les modèles. Intégrer une dimension temporelle, via des analyses longitudinales, serait donc fondamental pour évaluer la durabilité des efforts de mitigation.

En conclusion, notre étude montre qu'il est possible de détecter et de corriger certains biais, même implicites. Mais la route vers des systèmes réellement équitables est semée d'embûches. Chaque méthode a ses forces et ses limites, et des arbitrages doivent être faits entre performance, équité et faisabilité. Aucun outil ne peut, à lui seul, garantir une équité parfaite. Mais en continuant à questionner, à expérimenter, et à imaginer d'autres façons de faire, nous faisons un pas vers des technologies plus justes. Car vouloir une IA équitable, ce n'est pas seulement une ambition technique, c'est un choix de société. Ce n'est donc pas le progrès qu'il faut craindre, mais l'absence de vigilance éthique dans son déploiement.

« Tout notre progrès technologique dont on chante les louanges est comme une hache dans la main d'un criminel. »

— Albert Einstein

Bibliographie

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudik, John Langford, et Hanna Wallach. 2018. « A Reductions Approach to Fair Classification ». *Proceedings of the 35th International Conference on Machine Learning*, p 60-69.

Barocas, Solon, et Andrew D. Selbst. 2016. « Big Data's Disparate Impact ». SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Barocas, Solon, Moritz Hardt, et Arvind Narayanan. 2023. Fairness and Machine Learning.*MIT Press*.

Bashkirova, Anna, et Dario Krpan. 2024. « Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance ». *Computers in Human Behavior: Artificial Humans*, vol, issue 1.

Beattie, Geoffrey, et Patrick Johnson. 2012. « Possible unconscious bias in recruitment and promotion and the need to promote equality ». *Perspectives: Policy and Practice in Higher Education* vol 16, no 1, p 7-13.

Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, et al. 2018. « AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias ». *arXiv:1810.01943*.

Beneduce, Giusy.2020. « Artificial intelligence in recruitment: just because it's biased, does it mean it's bad? ». *NOVA—School of Business and Economics*.

Bertrand, Marianne, et Sendhil Mullainathan. 2004. « Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination ». *American Economic Review*, vol 94, no 4.

Bogen, Miranda et Rieke, Aaron.2018. «Help wanted: an examination of hiring algorithms, equity, and bias». *Upturn.*

Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, et Percy Liang. 2023. « The Foundation Model Transparency Index ». arXiv.

Breiman, Leo. 2001. « Random Forests ». *Machine Learning* vol 45, no 1, p 5-32.

Buolamwini, Joy, et Timnit Gebru. 2018. « Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification ».

Chang, Xinyu. 2023. «Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm». *Advances in Economics, Management and Political Sciences*, vol. 23, no 1, 2023, p. 134–140.

Chen, Zhenpeng, Jie M. Zhang, Federica Sarro, et Mark Harman. 2023. « A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers ». *arXiv:2207.03277*.

Chen, Zhisheng. 2023. « Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices ». *Humanities and Social Sciences Communications* vol. 10, no 1, p. 567.

Chong, Leah, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, et Jonathan Cagan. 2022. « Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice ». *Computers in Human Behavior, vol 127*.

Commission ontarienne des droits de la personne (CODP). 2024.« Évaluation de l'impact de l'intelligence artificielle sur les droits de la personne ». En ligne : https://www3.ohrc.on.ca/fr/evaluation-de-limpact-de-lintelligence-artificielle-sur-les-droits-de-la-personne.

Corbett-Davies, Sam, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, et Sharad Goel. 2023. « The Measure and Mismeasure of Fairness ». *arXiv:1808.00023*.
Creswell, John W, et J David Creswell.2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, chapter 8.

Dastin, Jeffrey. 2018. « Insight - Amazon scraps secret AI recruiting tool that showed bias against women». *Reuters.* En ligne : https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-922showed-bias-against-women-idUSKCN1MK08G/.

Davis, Sharon E, Chad Dorn, Daniel J Park, et Michael E Matheny. 2025. « Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability ». *Journal of the American Medical Informatics Association* vol 32, issue 5, p 845-54.

Feldman, Michael, Friedler, John Moeller, Carlos Scheidegger, et Suresh Venkatasubramanian. 2015. « Certifying and removing disparate impact ». *arXiv:1412.3756*.

Ferrara, Emilio. 2024. « Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies ». vol 6, no.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, et Derek Roth. 2018. « A comparative study of fairness-enhancing interventions in machine learning ». *arXiv:1802.04422.*

Ganesh, Prakhar, Usman Gohar, Lu Cheng, et Golnoosh Farnadi. 2024. « Different Horses for Different Courses: Comparing Bias Mitigation Algorithms in ML ». *arXiv:2411.11101*.

Gartner (2019). « Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form ». https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have.

Goddard, Kate, Abdul Roudsari, et Jeremy C. Wyatt. 2012. « Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators ». *Journal of the American Medical Informatics Association.*

Gonzalez, Manuel, John Capman, Frederick Oswald, Evan Theys, et David Tomczak. 2019. « "Where's the I-O?" Artificial Intelligence and Machine Learning in Talent Management Systems ». *Personnel Assessment and Decisions*, vol. 5, no 3.

Goodfellow, Ian, Bengio, Yoshua et Courville, Aaron. 2016. Deep learning. *Cambridge: MIT press, 2016.*

Gouvernement du Canada, Statistique Canada. 2024. « Analyse de l'utilisation de l'intelligence artificielle par les entreprises au Canada, deuxième trimestre de 2024 ». 20 juin 2024. https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2024008-fra.htm.

Gouvernement du Canada. (2023). « Rights in the workplace». *Canada.ca.* En ligne : https://www.canada.ca/en/canadian-heritage/services/rights-workplace.html .

Gouvernement du Québec. (s.d.). « Charte des droits et libertés de la personne, RLRQ, c. C-12». Légis Québec. En ligne : https://www.legisquebec.gouv.qc.ca/fr/document/lc/c-12

Grand View Research**.** 2024. *Artificial Intelligence Market Size, Share et Trends Analysis Report By Solution, By Technology (Deep Learning, Machine Learning, NLP, Machine Vision, Generative AI), By Function, By End-Use, By Region, And Segment Forecasts, 2025–2030*. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market.

Graves, Alex. 2012. « Long Short-Term Memory ». In *Supervised Sequence Labelling with Recurrent Neural Networks*.

Gupta, Maya, Andrew Cotter, Mahdi Milani Fard, et Serena Wang. 2018. « Proxy Fairness ». *arXiv:1806.11212.*

Hamida, Sayda Umma, Mohammad Jabed Morshed Chowdhury, Narayan Ranjan Chakraborty, Kamanashis Biswas, et Shahrab Khan Sami. 2024. « Exploring the Landscape of Explainable Artificial Intelligence (XAI): A Systematic Review of Techniques and Applications ». *Big Data and Cognitive Computing* vol 8, no 11, 149.

Hunkenschroer, Anna Lena, et Christoph Luetge. 2022. « Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda ». *Journal of Business Ethics*, vol 178, p. 977-1007.

IEEE Standards Association. « The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ». 2016. En ligne : https://standards.ieee.org/wp-content/uploads/import/documents/other/ec_about_us.pdf

IEEE Standards Association.« The IEEE Standard for Algorithmic Bias Considerations ». 2024.. En ligne : https://standards.ieee.org/ieee/7003/11357/.

Kahneman, Daniel. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux.*

Kamiran, Faisal, Asim Karim, et Xiangliang Zhang. 2012. « Decision Theory for Discrimination-Aware Classification ». *2012 IEEE 12th International Conference on Data Mining*.

Kamiran, Faisal, et Toon Calders. 2012. « Data Preprocessing Techniques for Classification without Discrimination ». *Knowledge and Information Systems*, vol 3, p 1-33.

Kleinberg, Jon, Sendhil Mullainathan, et Manish Raghavan. 2016. « Inherent Trade-Offs in the Fair Determination of Risk Scores ». arXiv.

Kline, Patrick M., Rose E.van K. et Walters, Christopher R. (2022). "Systemic Discrimination Among Large U.S. Employers". *National Bureau of Economic Research*, Working Paper n° 29053.

Krügel, Sebastian, Andreas Ostermaier, et Matthias Uhl. 2021. « Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions ». *arXiv:2106.16122*.

Li, Jingshu, Yitian Yang, Renwen Zhang, et Yi-chieh Lee. 2024. « Overconfident and Unconfident AI Hinder Human-AI Collaboration ». *arXiv:2402.07632*.

Lim, Weng Marc. 2024. « What Is Quantitative Research? An Overview and Guidelines ». *Australasian Marketing Journal*.

Mavrogiorgos, Konstantinos, Athanasios Kiourtis, Argyro Mavrogiorgou, Andreas Menychtas, et Dimosthenis Kyriazis. 2024. « Bias in Machine Learning: A Literature Review ». *Applied Sciences*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, et Aram Galstyan. 2022. « A Survey on Bias and Fairness in Machine Learning ». arXiv.

Mielke, Sabrina J., Zaid Alyafeai, Elizabeth Salesky, et al. 2021. « Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP ». *arXiv:2112.10508*.

Ng, Alfred. 2023. « 'Wholly ineffective and pretty obviously racist': Inside New Orleans' struggle with facial-recognition policing ».*Politico.* En ligne : https://www.politico.com/news/2023/10/31/new-orleans-police-facial-recognition-00121427.

Nicoletti, Leonardo et Bass, Dina**.** 2023. « Humans Are Biased. Generative AI Is Even Worse ». *Bloomberg Technology.* En ligne: https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Organisation de coopération et de développement économiques. « Principes de l'IA *».*2019. En ligne : https://www.oecd.org/fr/themes/principes-de-l-ia.html.

Organisation internationale de normalisation (ISO). 2021. « *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making* ». En ligne : https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:24027:ed-1:v1:en.

Oyeniran, Oyekunle, Adebunmi Adewusi, Adams Adeleke, Lucy Akwawa, et Chidimma Azubuko. 2022. « Ethical AI: Addressing bias in machine learning models and software applications ». *Computer Science & IT Research* vol 3, issue 3, p 115-26.

Pangambam, S. 2018. « Andrew Ng: Artificial Intelligence Is the New Electricity at Stanford GSB (Transcript) ». *The Singju Post* (blog). 23 janvier. https://singjupost.com/andrew-ng-artificial-intelligence-is-the-new-electricity-at-stanford-gsb-transcript/.

Peña, Alejandro, Ignacio Serna, Aythami Morales, et Julian Fierrez. 2020. « Bias in Multimodal AI: Testbed for Fair Automatic Recruitment ». *arXiv:2004.07173*.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, et Kilian Q Weinberger. 2017. « On Fairness and Calibration ». *Advances in Neural Information Processing Systems* 30.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, et Karen Levy. 2019. « Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices ». *arXiv:1906.09208*.

Ramesh, Krithika, Sunayana Sitaram, et Monojit Choudhury. 2023. « Fairness in Language Models Beyond English: Gaps and Challenges ». arXiv.

ReportLinker. 2023. « Information Technology Global Market Report 2023 ». GlobeNewswire News Room. 21 avril. https://www.globenewswire.com/news-release/2023/04/21/2652000/0/en/Information-Technology-Global-Market-Report-2023.html.

Rivera, Lauren A. 2012. « Hiring as Cultural Matching: The Case of Elite Professional Service Firms ». *American Sociological Review* , vol 77, issue 6.

Samiksha Chaudhuri et Ipsita Mohanty. 2023. « The Importance of Bias Mitigation in AI: Strategies for Fair, Ethical AI Systems ». *UXmatters.*

Schwing, Alexander G., et Raquel Urtasun. 2015. « Fully Connected Deep Structured Networks ». *arXiv:1503.02351*.

Semmelrock, Harald, Tony Ross-Hellauer, Simone Kopeinik, et al. 2025. « Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers ». *arXiv:2406.14325*.

Stanford Graduate School of Business (2017). Andrew Ng: Artificial Intelligence is the New Electricity. [Vidéo]. YouTube. https://www.youtube.com/watch?v=21EiKfQYZXc.

Suresh, Harini, et John V. Guttag. 2021. « A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle ». *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Université de Montréal. 2018.« La Déclaration de Montréal IA responsable ». En ligne : https://declarationmontreal-iaresponsable.com/la-declaration/.

Verma, Sahil, et Julia Rubin. 2018. « Fairness Definitions Explained ». *Proceedings of the International Workshop on Software Fairness*.

Villemure, René. 2019. *L'éthique pour tous ...même vous! Petit traité pour mieux vivre ensemble*. Les Éditions de l'Homme.

Walter, Yoshija. 2024. « Managing the Race to the Moon: Global Policy and Governance in Artificial Intelligence Regulation—A Contemporary Overview and an Analysis of Socioeconomic Consequences ». *Discover Artificial Intelligence*, vol 4, no 1.

Watson, Matthew, Divyashree Shivakumar Sreepathihalli, Francois Chollet, et al. 2024. « KerasCV and KerasNLP: Vision and Language Power-Ups ». *arXiv:2405.20247.*

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. « The FAIR Guiding Principles for Scientific Data Management and Stewardship ». *Scientific Data* vol 3, article num 160018.

Ying, Xue. 2019. « An Overview of Overfitting and Its Solutions ». *Journal of Physics: Conference Series, 1168, 022022.*

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, et Krishna P. Gummadi. 2017. « Fairness Constraints: Mechanisms for Fair Classification ». *arXiv:1507.05259*.

Zajko, Mike. 2021. « Conservative AI and social inequality: Conceptualizing alternatives to bias through social theory ». *AI & SOCIETY*.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, et Cynthia Dwork. 2013. « Learning Fair Representations ». *Proceedings of the 30th International Conference on Machine Learning*.

Zhang, Brian Hu, Blake Lemoine, et Margaret Mitchell. 2018. « Mitigating Unwanted Biases with Adversarial Learning ». *arXiv:1801.07593*.

Zhang, Jie, et Zhisheng Chen. 2023. « Exploring Human Resource Management Digital Transformation in the Digital Age ». *Journal of the Knowledge Economy*, vol. 15, no 1.

Annexe A : Déclaration sur l'utilisation de l'intelligence artificielle générative (IAG)

Je déclare avoir pris une entente avec mes directeurs de recherche dans le cadre de ce travail quant aux types d'utilisation faite de l'intelligence artificielle générative (IAG).

Usages permis de l'IAG

Je m'engage à respecter les principes d'intégrité académique et à utiliser l'IAG uniquement dans les limites suivantes :

- **Compréhension des concepts**

  L'IAG a pu être utilisée pour m'aider à mieux comprendre un modèle, un article de recherche ou une méthodologie scientifique. Toutefois, l'analyse critique et l'interprétation des résultats relèvent exclusivement de ma responsabilité.

- **Collecte de la littérature**

  L'IAG a pu être sollicitée comme outil d'assistance pour identifier des articles et ouvrages pertinents en lien avec ma problématique. La sélection finale et l'analyse de ces sources ont été effectuées de manière critique et personnelle.

- **Correction linguistique**

  L'IAG a pu être utilisée pour corriger l'orthographe, la grammaire et la syntaxe. Mon argumentation et mes idées demeurent toutefois personnelles et conformes aux exigences académiques.

- **Correction de code informatique**

  Dans le cas de l'utilisation de code informatique dans ce mémoire, l'IAG a pu m'aider à identifier et corriger des erreurs techniques. La conception,

l'implémentation et l'explication du code reflètent néanmoins mon propre travail et ma compréhension.

- **Recherche de données**

  L'IAG a pu être mobilisée pour identifier des sources de données pertinentes. La vérification de leur fiabilité et validité ainsi que leur analyse ont été faites par moi-même.

Usages exclus de l'IAG

Je déclare ne pas avoir utilisé l'IAG pour :

- Générer automatiquement des portions de mon mémoire sans analyse critique de ma part.
- Produire du contenu sans vérification et reformulation personnelles.
- Réaliser une analyse ou une discussion artificiellement construite sans intervention réelle de ma part.