UNIVERSITÉ DE GENÈVE

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

UNIVERSITÉ DE GENÈVE

TRAVAIL DE BACHELOR

Biais et équité à travers les algorithmes d'IA : comment se rapprocher d'une *fairness*

Une étude de l'état de la *fairness* dans les algorithmes d'Intelligence Artificielle et des biais introduits

Auteur :
Klim Eduardo
TROYAN MACHADO

Superviseur :
Pr. Stéphane
MARCHAND-
MAILLET

13 septembre 2021

Faculté des Sciences de l'Informatique

"Science sans conscience n'est que ruine de l'âme."

— François Rabelais, 1532

"Une réponse approximative à la bonne question a beaucoup plus de valeur qu'une réponse précise à la mauvaise question."

— John Tukey

"Let us change our traditional attitude to the construction of programs : Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do."

— Donald E. Knuth

"By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."

— Eliezer Yudkowsky

REMERCIEMENTS

Chers professeurs, vous avez étés mes accompagnants, ”changeurs de paradigmes”, ”transformateurs” ; je vous adresse ainsi ce court message afin de vous exprimer ma reconnaissance pour votre disponibilité, votre attention et le partage de votre expertise, tout au long de mon parcours de bachelier en Sciences Informatiques. Vos connaissances et discussions ont su guider mes réflexions et les animer profondément.

Je tiens à exprimer une reconnaissance particulière au Professeur Stéphane Marchand-Maillet, superviseur et honnête guide, ayant su m’encadrer, m’orienter et m’aider dans de nombreuses étapes de mon cursus universitaire, et ce, tout en me laissant une précieuse liberté et autonomie.

Je remercie ma chère famille et, en particulier, ma mère qui m’a toujours soutenu et motivé par sa confiance, son enthousiasme et intérêt envers mes études et travaux.

À tous ces intervenants, je présente mes sincères remerciements et ma gratitude.

RÉSUMÉ

Cet article propose une première approche et vue d'ensemble de l'état actuel de la recherche en matière de *fairness* et biais en IA, ainsi qu'une discussion des pistes pour des méthodes de détection et/ou correction des biais. Il se base sur de nombreux articles et travaux proposant des avancées sur le sujet ainsi que des articles plus similaires ayant également cherché à offrir une compréhension globale - qui s'est vu compliquer par la croissance exponentielle du nombre de publications en Machine Learning - de l'état de l'art. Nous commençons par introduire le contexte dans lequel les recherches sont entreprises ainsi que ce qui motive le sujet d'étude dont il est question. L'introduction est suivie d'un chapitre préliminaire explicitant les terminologies et concepts qui seront utilisés et posant les bases requises, en particulier par rapport aux biais, algorithmes ainsi qu'au bayésianisme - opposé au fréquentisme - et ses principes. Il s'en suit une brève présentation de certains travaux proches du sujet d'étude ou l'étayant, et méritant une attention ou un approfondissement de la part du lecteur. Le chapitre de développement est consacré à l'étude et discussion de deux publications soigneusement sélectionnées dont les éléments de recherche principaux et ceux appuyant les idées du sujet d'étude sont relevés. Ils participent ainsi principalement aux éléments de réponse que nous démarquons dans le chapitre faisant part des résultats pertinents conclus suite à l'étude de l'état de la recherche actuelle, ainsi que la direction qui pourrait être suivie. Finalement, nous concluons en mettant en exergue les aspects à retenir dans la quête d'un aboutissement vers "une" fairness - plutôt que "la" - ainsi que dans celle d'un contrôle des biais algorithmiques et humains.

Mots-clés— Algorithme, Biais, Fairness, Machine Learning, Bayésianisme, NLP, Word Embeddings

TABLE DES MATIÈRES

1

INTRODUCTION ET MOTIVATIONS

L'ère du numérique, bouleversant changement s'étant glissée dans toutes les sphères de la vie humaine ces dernières décennies, accompagnée des progrès croissants et poussée plus récemment par les avancées fulgurantes en Intelligence Artificielle, a amené à s'interroger sur l'impact de l'usage de *big data* dans notre quotidien. Tous les domaines régissant l'ordre de la vie sociétale se trouvent concernés, pressentant d'importantes fractures déjà présentes ou potentielles qui convoquent des réflexions en la matière : comment maîtriser ce nouveau pouvoir d'influence à grande échelle ? Comment aboutir en pratique, et de façon réaliste, à une protection des données sans contraindre les progrès scientifiques (e.g., *deep learning*) qui se basent de plus en sur la qualité et quantité des données à disposition ? Comment balancer un "surcontrôle" et une trop forte influence des algorithmes dans le contexte de flux de données encore difficilement contrôlables ? Comment distinguer une sélection appropriée de données qui seront traitées par des algorithmes et spécifier leur objectif de façon suffisamment précise ?

Le questionnement des médias dans les pages titrées, par exemple, "Et si l'Intelligence Artificielle était déjà hors de contrôle ?" [1], complété le plus souvent par des exclamations sur les avancées en termes de transhumanisme, apparaît comme une preuve vivante de la force d'influence pouvant susciter une attaque au sentiment de sécurité humaine. Il est en fait naturel d'avoir du mal à se situer au sein de ce caléidoscope d'aspects, les uns plus attractifs à approfondir que les autres. D'une part, nous comprenons à la lecture d'articles - tant vulgarisés que scientifiques - que les machines, en particulier celle basées sur l'IA, sont entraînées à pouvoir prendre des décisions d'elles-mêmes, effectuer des choix, et finalement, agir de façon (quasi-)autonome. En effet, comme décrit, dans l'étude canadienne de Maclure, Saintpierre (2018) [2], les algorithmes sont conçus afin "d'entraîner une machine à distinguer, par exemple, les essences de bois par la reconnaissance d'image, ce qui fait qu'elle se trouve en mesure de reconnaître et ainsi trier les billots qui lui seront présentés selon leur essence". Un parallèle peut être fait entre ce procédé d'identification d'images qui sert à trier à haute vitesse et avec extrême précision des images de strates de cerveau et d'y repérer, le cas échéant, "des tumeurs cancéreuses difficilement identifiables par l'humain" [3]. "La machine agit donc avec une certaine autonomie, en extrapolant à partir de recoupements d'information. Dans la majorité des cas, pour l'instant, il est envisagé d'utiliser des systèmes d'IA qui agiront en complémentarité avec l'humain" [4], c'est-à-dire en faisant des recommandations qui devraient être validées par des humains. Cette perspective devrait rassurer, d'une certaine façon, les craintes quant à la perte de contrôle totale sur les machines intelligentes. Une idée, difficilement évitable et donc dont il faut tenir compte, est que "..plus l'utilisation deviendra régulière et les résultats

1. Et si l'Intelligence Artificielle était déjà hors de contrôle ?
2. *Le nouvel âge de l'intelligence artificielle : une synthèse des enjeux éthiques*, Vol. 30, n°3., 2018
3. *Ibid.*, p.753
4. *Ibid.*, p.753

probants, plus on aura tendance à se fier aux décisions que ces machines prendront. La machine agira avec une certaine autonomie, en extrapolant à partir de recoupements d'information" [5]. Cette ouverture vers une certaine autogestion ne peut que mobiliser plus de questionnements de la part des humains qui devront faire face à tous les possibles risques dans un processus d'implémentation d'un fonctionnement durable.

Pour ces raisons, nous évoquons actuellement une nouvelle ère de l'Intelligence Artificielle suscitant inoxérablement et en continu plusieurs interrogations quant à la distribution des niveaux de responsabilités des "acteurs" impliqués. Responsables de par cette appartenance, nous tentons de plonger dans cet océan de connaissances à croissance exponentielle [6] [7] afin de discuter - puis d'assurer - une veille sur la problématique des biais et donc de la *fairness* ("équité") dans les domaines liés à l'Intelligence Artificielle.

Alors que pour la plupart des individus les enjeux mentionnés semblent pour eux innateignables - dans le sens où ils ne peuvent directement agir dessus - et ne restent ainsi qu'un tableau inachevé à contempler, il est nécessaire d'au moins s'interroger sur les diverses responsabilités et moyens à disposition. La responsabilité appartient-elle aux concepteurs d'algorithmes ? Aux testeurs et entraîneurs en charge, aux préparateurs, aux fournisseurs de données à la machine ou encore à ceux qui ont créé les senseurs qui devraient permettre au véhicule de percevoir correctement son environnement et de se diriger ? Ou encore au gouvernement qui a mis en oeuvre, ou pas, un cadre réglementaire possiblement lacunaire ? Nous nous retrouvons ici devant une série de questions qui exigent potentiellement l'élaboration d'une nouvelle conception de la responsabilité morale et juridique, une responsabilité adaptée aux machines autonomes et apprenantes, ainsi qu'au cadre en vigueur. Certains intervenants, tel que le Professeur Rachid Guerraoui (EPFL [8]), signalent qu'aujourd'hui déjà, des systèmes d'IA détectent lorsqu'un humain tente de modifier leur comportement et font parfois tout pour rejeter cette intervention et la contourner [9].

Des données insuffisantes ou relatant des pratiques discriminatoires peuvent reproduire des biais ou même en créer, par exemple, en faisant des corrélations entre des éléments qui ne devraient pas être liés. Des biais discriminatoires ont notamment été repérés dans des algorithmes chargés d'évaluer les potentialités de récidives de criminels et leur admissibilité à une libération conditionnelle [10] [11] [12] [13]. De tels algorithmes employés pour traiter ces demandes évalueraient systématiquement les demandes provenant d'Afro-Américains comme présentant des risques de récidive supérieurs à la réalité, alors qu'il minimiserait - à tort - les risques de récidive des Caucasiens, notamment parce que les données ayant servi à l'apprentissage de l'algorithme font ressortir un plus haut taux de criminalité dans les communautés noires.

---

5. *Ibid.*, p.753
6. Google says "exponential" growth of AI is changing nature of compute
7. Growth of AI research in 2020 ? Steady on the exponential path in times of crisis.
8. EPFL : Ecole Polytechnique Fédérale de Lausanne
9. Et si l'Intelligence Artificielle était déjà hors de contrôle ?
10. Automated Justice : Algorithms, Big Data and Criminal Justice Systems
11. Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality : Algorithms are Color Blind and Never Tire Future
12. AI is sending people to jail - and getting it wrong
13. Discriminating algorithms : 5 times AI showed prejudices

Similairement, un algorithme utilisé pour trier des CV [14] dans une firme de recrutement sélectionnerait des hommes blancs dans une proportion beaucoup plus élevée que des femmes ou des hommes de couleur pour pourvoir au poste [15] car, ici encore, les données fournies aux algorithmes étaient initialement biaisées, puisque basées sur l'historique de recrutement de grandes organisations où, traditionnellement, des hommes blancs occupaient les postes de pouvoir. Les biais algorithmiques peuvent ainsi perpétuer des injustices déjà existantes, voire pire, amplifier les discriminations menant à des impacts négatifs sur le plan du respect des principes d'équité entre les personnes.

Enfin, la complexité de l'enjeu est encore plus grande lorsque les données qui sont fournies à la machine proviennent des interactions avec les utilisateurs et permettent à l'algorithme de continuer à apprendre en cours d'utilisation réelle. C'est le cas par exemple de la société Netflix [16], qui adapte ses propositions de contenu en fonction des choix faits par l'utilisateur. Il est essentiel d'empêcher la manipulation des IAs lorsqu'elles sont placées en contexte réel et s'assurer que les algorithmes ne subissent pas une influence indue qui viendrait modifier la trajectoire décisionnelle souhaitée en introduisant des biais nocifs. Ces problématiques apparaissent comme un défi actuel et majeur, en particulier pour la communauté académique des Sciences Informatiques.

Dans une perspective de communication inter-disciplinaire, il s'agirait de mettre la priorité sur la détection et discussion régulière des nouveaux enjeux éthiques ou non résolus, afin de parevnir à délimiter les territoires de confiance et la fiabilité dans les décisions prises par les machines qui sont ici mises en cause. Cette stratégie va au regard de l'inévitable évolution sociétale vers l'hypermodernité qui succède aux temps post modernes. Peut-être serait-il souhaitable que les systèmes reposant sur l'IA pour prendre des décisions fonctionnent à l'intérieur de balises où "des mécanismes de surveillance" [17] devraient être prévus et validés avant l'exécution de décisions ? Comme le suggèrent, entre autres, dans leur étude Maclure et Saintpierre (2018) [18].

Notre intérêt envers ce vaste sujet d'études se fait naturellement complété par certains scientifiques ayant déjà creusé ces sujets de façon plus systémique.

L'étendue présence des biais dans les moteurs de recherche (e.g., recommendations, complétion de texte), les réseaux sociaux, les publicités sur les sites internet, etc, engendrent des répercussions qui ne peuvent pas être négligée : discrimination de genre, discrimination ethnique, incitation à la haine et diffamation, manipulation des votes, etc.

Les inquiétudes quant aux biais et au manque de fairness a amené la création de groupes - tels que l'union entre Google, Amazon, Microsoft, et autres - dédiés à l'étude et la résolution de ces enjeux. De plus, d'autres démarches ont donné lieu à, par exemple, la conférence annuelle FAccT [19]. Toutefois, il faut garder en tête que les principaux concernés quant à la

14. CV : Curricula Vitae
15. Challenges for mitigating bias in algorithmic hiring
16. Maclure, Saintpierre, 2018, p.756
17. Maclure, Saintpierre, 2018, p.757
18. *Ibid.*, p.757
19. FAccT : Fairness, Accountability, Transparency. Site officiel : FAccT conference

concrète réalisation des systèmes liés aux biais font face au dilemme socio-économique du *trade-off* performance-fairness. Ceci n'est généralement pas un bon indicateur d'efforts à la problématique de la fairness, et donc des biais.

Pour terminer ce chapitre d'introduction dans lequel de nombreux aspects et exemples n'ont pas été abordés dû à la vastitude du sujet d'étude, il semble important encore de relever la complexité de la problématique qui fait intervenir une multitude d'acteurs de différents horizons. La propriété intellectuelle des algorithmes et données que les sociétés au pouvoir ne désirent en général pas partager (e.g., pour des raisons de compétitivité) viennent rendre les tâches de recherche et l'étude des biais algorithmiques plus compliquées. De plus, quand bien même le problème de manque de transparence serait levé, la croissance de la complexité des nouveaux algorithmes (e.g., *deep learning* et le problème de "boîte noire") rend les anticipations des résultats et leur analyse encore plus compliqué. L'état de la *fairness* en IA et des biais introduits est un chantier incontournable qui mérite d'être visité - par l'ingénieur, le législateur ainsi que le politique - et, pour les personnes responsables disposant des facultés, opportunités et outils adéquats, auquel il faudrait activement participer.

2

PRÉLIMINAIRE : TERMINOLOGIES, NOTIONS ET CONCEPTS

2.1 ALGORITHMES ET BIAIS

Un algorithme peut être défini comme une séquence d'instructions avec un ordre possible bien défini et utilisant des données (fournies ou/et produites par l'exécution de l'algorithme). Les capacités d'exécution d'un algorithme dépendent de l'environnement dans lequel il est exécuté. Les avancées technologiques ont notamment permis l'usage de méthodes de Machine Learning et Deep Learning afin de produire des résultats plus précis, et donnant sur un horizon de possibilités nouvelles pouvant grandement influencer le monde et ses systèmes actuels ; voir [52] par exemple.

À l'origine, les algorithmes avaient pour but de représenter un raisonnement logique humain bien défini afin de résoudre des tâches automatisables. Ainsi, un programme résoudrait un problème comme un programmeur pourrait le faire, mais en étant largement plus efficace en terme de performance (vitesse, précision). La conception du problème qu'a le programmeur peut ainsi engendrer l'inclusion de biais par ce dernier du fait de sa façon de voir les systèmes de ce monde à considérer et d'ainsi résoudre le problème. Plus les buts se sont avérés ambitieux et plus les tâches se sont avérées compliquées, plus il a fallu posséder une puissance de calcul et un espace mémoire accrus afin de les réaliser [1]. En particulier, les algorithmes de Machine Learning et Deep Learning qui sont des sujets d'études d'actualité depuis plusieurs années reposent énormément sur ces aspects (voir [58]). Cependant, ils ont encore plus apporté l'effet notable de "boîte noire" [2]. Bien que de nombreux algorithmes soient en pratique déployés avec cet effet, il semblerait naturel et plus responsable de l'éviter ou d'en tout cas l'interpréter ; voir le concept de *explainable AI*[36][59]. L'exécution de tels algorithmes et leur traitement des données sont devenus extrêmement compliqués et délicats à comprendre et analyser. Une des conséquences que nous étudierons un peu plus tard dans l'article est que la compréhension et correction de biais potentiellement introduits dans un algorithme sont certainement compliquées.

Algorithmes, données et utilisateurs sont tous acteurs et potentielle source de biais algorithmiques [3]. Les biais peuvent être introduits de plusieurs manières : par le programmeur ou designer - consciemment/volontairement ou non -, par les choix des utilisateurs, durant ou après la sélection des données, par leur nature, etc. Par exemple, afin de définir l'ordre dans lequel traiter les données, il revient au programmeur de faire un choix d'implémentation, de

1. The computational limits of Deep Learning
2. Why Are We Using Black Box Models in AI When We Don't Need To ? A Lesson From An Explainable AI Competition
3. What Do We Do About the Biases in AI ?

contraintes, de paramètres, etc. Obtenir des résultats espérés ou sensés, sans comprendre comment, n'offre aucune garantie quant au bon fonctionnement de l'algorithme ni à la confiance en sa fiabilité et précision future.

Les biais algorithmiques décrivent les résultats inéquitables de l'introduction de biais, volontaire ou non, dans les algorithmes. De tels résultats, dits biaisés, peuvent considérablement détourner le but même de l'algorithme. Certains groupes (e.g., biais de genre avec groupe hommes par rapport au groupe femmes) peuvent être privilégiés par l'algorithme et ainsi avoir une influence sur ses utilisateurs, directs ou indirects. À cela s'ajoute le fait que l'influence des algorithmes sur la société est bilatérale : les biais proviennent des actions de la société et nourrissent les biais qui impacteront cette même société.

Il existe de nombreuses catégories de biais d'algorithme [4]. Les biais peuvent venir des personnes ayant participé à la création de l'algorithme et/ou à la création et mise en place des systèmes avec lesquels l'algorithme est utilisé. Dans le cas où les biais font directement partie de l'algorithme lui-même, le risque de reproduction de résultats biaisés lors des utilisations de l'algorithme est accru, et ce potientellement indépendamment des autres systèmes en jeu. Il peut s'agir d'une introduction de biais volontaire ou non, visible ou pas, par les personnes ayant participé à la création de l'algorithme. Un exemple simple de biais pourrait être un biais introduit dans un système de recommendation dont un des mécanismes participant aux recommendations serait de proposer des résultats dans l'ordre alphabétique. Ainsi, dans le cas d'un choix parmi n recommendations, les recommendations commençant par la lettre "a" auront plus de chance d'être sélectionnées par l'utilisateur que celles commençant par la lettre "z" - car elles apparaissent de haut en bas de la page alphabétiquement. Des programmes de décision pourraient facilement désavantager [5] [6] un groupe d'utilisateurs à capacité, au sens le plus général, réduite, et ce de façon plus marquée lorsqu'il est possible par ses propres choix d'influencer grandement les résultats de l'algorithme. En somme, les individus désavantagés peuvent se trouver l'être encore plus ; cet effet est une amplification des biais, une exacerbation de la discrimination.

L'environnement et contexte dans lequel les algorithmes sont utilisés peuvent également influencer les résultats de l'algorithme ainsi que leur interprétation, et ce par l'ajout de biais. Ainsi, une surveillance et mise à jour des algorithmes par rapport à, par exemple, des nouvelles découvertes et avancées ou des changements sociétales/culturelles est importante afin d'éviter des biais "émergents" [7]. Une sous-catégorie de biais d'émergence est celle des biais renforcés par une réutilisation des données produites par l'algorithme comme nouvelles données d'entrée pour ce même algorithme. Un exemple bien connu est celui du système COMPAS [8] [9] (utilisant en partie le principe mentionné) qui a été critiqué pour stigmatiser les personnes de couleur noire. Un exemple simple auquel nous pouvons penser est celui des zones considérées comme ayant un taux de criminalité élevé. Un algorithme qui décide d'où les patrouilles de police vont devoir circuler en se basant sur où les patrouilles de police sont présentes ou sont intervenues

4. Three kinds of biases
5. For some employment algorithms, disability discrimination by default
6. How Algorithmic Bias Hurts People With Disabilities
7. 5 unexpected sources of bias in Artificial Intelligence
8. Inspecting algorithms for bias
9. 5 examples of biased AI

risque de concentrer continuellement la police aux mêmes endroits et ainsi augmenter les chances d'altercations pour les personnes y vivant ainsi que transmettre l'idée potentiellement biaisée que ces endroits ont un taux de criminalité élevé. Comme le signalent [20], il est aussi possible de trouver des biais découlant des contraintes techniques ou de puissance des machines exécutant les algorithmes. Plus généralement, il est important de faire attention aux systèmes ou algorithmes auxquels plusieurs personnes ont contribué. En effet, toute différence (e.g., de connaissance ou expertise) entre personnes ou groupes de personnes (e.g., le créateur d'un algorithme et la personne qui le met en place dans un système) peut donner lieu à des biais et par conséquent des résultats innatendus.

Un aspect fondamental de l'étude des biais est l'importante dépendance des données fournies à l'algorithme. Ainsi, comme discuté par [26] ou [5], le choix de données sélectionnées (e.g., il est possible de ne prendre qu'un sous-ensemble des données d'un ensemble de données obtenues après une étude, une observation, etc) et la façon de les adapter au système ou problème donné influencent considérablement les résultats et, dans le cas d'algorithmes de Machine Learning ou Deep Learning, leur apprentissage résultant en la création d'un modèle potentiellement différent de celui idéalement voulu. Par exemple, plus l'ensemble de données est grand, plus il y a de chance que les groupes les plus représentés soient plus traités et éventuellement favorisés par rapport aux minorités pour lesquelles moins de données sont disponibles. Ainsi, il pourrait falloir rebalancer cette différence ou l'ajuster si besoin, sans pour autant favoriser un groupe sans que cela ait du sens par rapport à la situation réelle. C'est dans cela que réside aussi un des enjeux compliqués de la correction de biais qui sera discuté dans la section 4. L'environnement dans lequel l'algorithme sera utilisé peut donner lieu à des données complètement nouvelles pour lesquelles des résultats innatendus ou indésirés pourraient être obtenus. La difficulté résidant dans le fait que nous ne pouvons pas prévoir simplement toutes les données possibles qui seront manipulées par l'algorithme.

Alors que les avancées technologiques et les algorithmes - base de ces améliorations - se développent, leur complexité croissante est suivie de près par celle de la détection, analyse et correction des biais certainement existants.

2.2 NLP ET ALGORITHMES DE MACHINE LEARNING

Le *Natural Language Processing* (NLP) est un domaine important de l'Intelligence Artificielle ou du Machine Learning que nous retrouvons dans un grand nombre d'applications courantes telles que la détection de spam dans les boîtes e-mail[30], les recommandations des moteurs de recherche[43], les outils de traduction[56] et la gestion de contenu sensible sur les réseaux sociaux[9], pour n'en citer que quelques unes. Cette proximité directe avec un grand nombre d'humains "subissant" et influençant les algorithmes de NLP devrait assurer une attention particulière des chercheurs, certes en matière de performance, mais plus important encore, en matière de fairness et gestion des biais dans les algorithmes. D'autant plus que, en particulier pour ce domaine, les données principales (i.e., du texte) ont directement été produites par des humains (e.g., articles de journal, littéraire, scientifique, commentaires sur les réseaux sociaux, revues de produits, etc). Ceci facilite alors l'introduction de biais sans contrôle adéquat des données utilisées sachant que des effets d'aggravement (voir [63][31][38]) peuvent subvenir jusqu'à devenir compliqué à gérer raisonnablement. De nombreuses exclamations

ont été entendues quant à de tels effets ayant été observés dans les algorithmes de certaines sociétés. Par exemple, YouTube s'est vu accuser par divers journalistes et académiques ([63]) de l'existence d'un effet de radicalisation sur ses utilisateurs bien que plusieurs publications scientifiques affirment ne pas pouvoir conclure faute d'évidence[50][35]. À cela s'ajoute encore la forte scalabilité des modèles qui possèdent ainsi une forte influence sur les utilisateurs des systèmes sur lesquels ils se basent (e.g., Google, Facebook, YouTube, etc). Malheureusement pour la fairness, étant en général opposé au désir de performance, les plus grandes sociétés dotées d'une influence non-négligeable sur l'immense partie des humains se servant de leurs services ne partagent pas publiquement les détails des fonctionnements de leurs algorithmes (voir par exemple [15] pour une idée de l'algorithme de recommandation de YouTube).

L'idée du NLP est de comprendre la syntaxe, la sémantique et les structures des langages naturels (i.e., langages humains). Pour cela, il est nécessaire d'analyser des données (i.e., du texte sous forme écrite ou sonore) afin de produire des résultats associés en se basant sur ce qui a pu être appris. Le NLP, à l'intersection de la linguistique et des sciences informatiques, peut être vu comme un pont entre les langages naturels et la machine. Les difficultés des langages naturels telles que leur ambiguïté (e.g., dans la détection et le traitement de sarcasme) sont encore des défis à relever.

Dans la section 4.2.4 du chapitre 4, nous étudions la publication "*Man is to Programmer as Woman is to Homemaker ? Debiasing Word Embeddings*"[7] afin d'approfondir les notions qui auront été vues jusque là ainsi que démontrer l'importance du contrôle des biais dans le NLP pour lequel le problème est encore loin d'être complétement résolu [10].

2.2.1 *Word Embeddings : Word2Vec*

Le *Word Embeddings* (voir [41]) est une représentation vectorielle numérique (i.e., dont les éléments sont des nombres) des mots. Chaque dimension du vecteur ou word embedding contient de l'information syntaxique ou sémantique d'un mot, et le vecteur entier est considéré comme une représentation du mot de base dans un espace dimensionnel supérieur de *features*. Il y a plusieurs méthodes permettant d'obtenir un word embedding ("vectorisation de mots"). Parmi les plus récentes, les chercheurs se concentrent sur les méthodes usant de réseaux de neurones (appelées "*neural network embeddings*") telles que Word2Vec. De telles méthodes sont efficaces pour l'évaluation de similarité entre mots, identifier des entités nommées ou encore pour le *sentiment analysis* d'un corpus de texte.

Word2Vec est un algorithme usant d'un *2-layer neural network* (ce type d'embeddings est appelé "*neural embeddings*") qui génère des word embeddings. L'input est un corpus de texte et l'output est un ensemble de vecteurs (sous forme numérique, les mots sont "traduits" en des nombres) obtenu selon la reconnaissance de patterns, similarités, features (e.g., le contexte particulier d'un mot) des mots composant le corpus de texte pris en entrée, etc. Chaque mot est alors représenté dans un espace à $N \in \mathbb{N}$ dimensions, où N est la taille d'un vecteur. De nouveau, des mots "similaires" seront alors proches dans l'espace à N dimensions. Par exemple, les mots "prince", "reine", "roi" et "princesse" ont des chances de former un cluster dans l'espace représenté.

10. We can reduce gender bias in natural language AI, but it will take a lot more work

Word2Vec peut aussi apprendre d'autres types d'association que les similarités mentionnées afin de résoudre des analogies. Par exemple, Word2Vec est utile pour la traduction des mots d'un langage à un autre en évaluant les relations entre les mots d'un langage et les associant (par rapport à leur position dans l'espace) à ceux d'un autre langage. Nous notons que la sortie de l'algorithme pourra alors être traitée par des réseaux de neurones profonds (*"deep neural networks"*) efficacement au vu de sa forme vectorielle. En fournissant suffisamment de données, d'exemples et des contextes, Word2Vec permet de deviner efficacement, par exemple, le sens d'un mot dans un contexte particulier pour ainsi résoudre des analogies [11] (i.e., type d'associations de mots).

Le Word Embeddings est un concept pour lequel il existe plusieurs techniques l'appliquant différemment. Ici, nous nous concentrons sur *Word2Vec* en notant que *GloVe* ([46]) est une autre technique très populaire et efficace dont nous n'allons cependant pas parler dans cet article.

Finalement, nous relevons que le concept de Word Embeddings donne lieu à des techniques très puissantes pour lesquelles les méthodologies et outils d'analyse employés sont importants et peuvent fortement influencer l'étude des biais. Ceci est d'autant plus marqué que le Word Embeddings est prône à l'introduction de biais humains. Ainsi, de nombreux chercheurs ont publié et publient encore des études de méthodologies utilisées (e.g., l'analyse par analogies discuté par [44], voir prochaine section) afin de permettre une meilleure approche aux chercheurs.

2.2.2 *Les analogies comme exemple*

Une analogie est un groupe de mots de la forme "A est à A' ce que B est à B'". Elle peut aussi être formulée de façon simplifiée par : "$A : A' :: B : B'$". Par exemple, Bern est à la Suisse ce que Séoul est à la Corée du Sud, c'est-à-dire certainement l'association "capitale du pays" a été faite. Généralement, la tâche est de trouver B' sachant A, A' et B. Il existe diverses méthodes ([44]) afin de résoudre les analogies : méthodes pair-based, méthodes set-based, matrice pair-pattern, etc. Les résultats d'un modèle peuvent parfois être très suprenants. En y réfléchissant longuement, il arrive de découvrir des relations que le modèle a pu déceler sans qu'elles ne soient évidentes à première vue pour un humain - un ensemble de connaissances plus poussées peut être requis.

Les analogies peuvent se résoudre à partir des word emeddings. En simplifiant, de nombreuses analogies peuvent être résolues en additionnant (au signe près) les word embeddings. Par exemple, l'analogie "*Man is to king as woman is to X*" ("L'homme est au roi ce que la femme est à X") peut être résolue en trouvant le mot le plus proche (i.e., donnant un vecteur similaire ou en tout cas le plus proche par rapport aux autres choix possibles - nous verrons que le cosinus est ainsi une mesure utile dans ce contexte - de $V_{king} - V_{man} + V_{woman} \approx V_{queen}$.

Pour comprendre ce type de relations permettant de résoudre les analogies, il faut alors s'intéresser à un type d'embedding tel que, par exemple, le Word2Vec embedding. Des ana-

11. Voir la section 2.2.2

logies plus compliquées peuvent donner lieu à des résultats intéressants et pas forcément attendus. Par exemple [12] : ”Le New York Times est à Sulzberger ce que Fox est à X” [13] qu'un modèle pourrait compléter par ”Murdoch”, ”Chernin”, ”Bancroft”, etc, en attribuant un ordre aux mots par rapport à la valeur de l'estimation effectuée. Un autre exemple, peut-être plus poétique : ”La raison est à l'amour ce que la folie est à ...”, que nous laissons au lecteur le plaisir de compléter.

[44] affirment principalement que les analogies pourraient entraîner des confusions ou une mésinterprétation des biais présents (ou potentiellement pire, une inteprétation de biais inexistants) si mal utilisées.

2.3 BAYÉSIANISME

Afin de conclure ce grand chapitre préliminaire au sujet d'étude, nous présentons le bayésianisme en définissant le contexte et contrastant avec le fréquentisme, puis abordons ses concepts principaux qui seront utiles par la suite, lors des discussions et développements des approches bayésiennes dans l'état de l'art.

Les études statistiques menées aujourd'hui sont principalement dominées par deux approches et modes de pensée que sont le fréquentisme et le bayésianisme. L'approche fréquentiste est encore plus marquée par l'adoption du test par *p-value* et le concept du ”statistiquement significatif”. L'usage de ces méthodes peut en fait dépendre de ce qui doit être mesuré. La méthode fréquentiste s'intéresse à la probabilité des événements selon une certaine théorie, une certaine hypothèse, un certain modèle. La méthode bayésienne, elle, s'intéresse à la probabilité des théories ou hypothèses - encore une fois, sans distinction particulière - au vu de certains évènements. Pour les modèles paramétrés, ces approches se concentrent sur l'estimation des paramètres et leur probabilité (postérieure) d'appartenance à un certain ensemble, respectivement. Les modèles non-paramétrés sont étudiés à travers leur fonction (e.g., régression). Cependant, [34] remarque que l'obsession pour le vrai modèle (*true model*) et ses paramètres écarte discrètement la possibilité d'existence de plusieurs ensembles de modèles et paramètres qui décriraient eux aussi les données raisonnablement bien. Ainsi, cela nous éloigne d'analyses des propriétés communes à différents modèles, et de ce qui fait qu'ils permettent une description similaire des données bien qu'ils soient différents. Il semble aussi raisonnable de penser qu'obtenir plusieurs modèles décrivant correctement les mêmes données renforce notre confiance en les prédictions. Inversément, les caractéristiques discordantes des modèles peuvent faire questionner cette confiance, nous poussant ainsi à requestionner nos inférences.

Ces remarques amènent déjà naturellement à s'intéresser au bayésianisme du fait de sa nature centrée sur la propagation de connaissance, comme nous le verrons un peu plus bas. [34] insistent sur la tromperie des paradigmes actuels qui se concentrent sur l'utilisation de procédures ayant pour but l'estimation exacte des vrais valeurs (*true values)* et modèles plutôt

12. Word2Vec examples

13. L'article Word2Vec examples propose l'explication suivante : ”The Sulzberger-Ochs family owns and runs the NYT. The Murdoch family owns News Corp., which owns Fox News. Peter Chernin was News Corp.'s COO for 13 years. Roger Ailes is president of Fox News. The Bancroft family sold the Wall St. Journal to News Corp.”

que de se concentrer sur l'évaluation de la capacité des modèles et paramètres à bien décrire les données. Ce changement de paradigme se trouverait extrêmement utile dans les cas où un vrai modèle n'existe pas forcément.

Contrairement à la méthode fréquentiste où une unique hypothèse (qui définit le contexte dans lequel les évènements se produisent, la façon de modéliser le monde dans lequel est mené l'expérience) est considérée et des évènements ont lieu, la méthode bayésienne use de multiples hypothèses auxquelles elle va attribuer un pourcentage de plausibilité par rapport à un évènement ou une série d'évènements particulière. Cela permet de répondre à la question : "Ayant obtenu le résultat R lors d'une expérience, quelle est la probabilité qu'il ait été obtenu selon l'hypothèse H_i ?". Par exemple, ayant plusieurs dés équilibrés/non-pipés avec un nombre de faces différent, si l'on obtient un 5 comme résultat après un lancer de dé (sans savoir quel dé a été utilisé), quelle est la probabilité que le dé à i faces ait été lancé afin de produire ce résultat ? Ici, le nombre de dés différents (i faces différentes) est le nombre d'hypothèses (H_i hypothèses différentes). Dans le cas où le résultat obtenu est, par exemple, 12, nous savons que le dé à six faces ne peut pas avoir produit le résultat, donc l'hypothèse pour le dé à six faces a une plausibilité de 0 (la probabilité que ce dé ait été choisi est de 0). En revanche, les dés à 12 faces ou plus auront une probabilité décroissante non nulle d'avoir été choisis.

Dans de nombreux domaines (e.g., médecine), l'expérience est importante et grandement utilisée comme outil de choix. L'approche bayésienne offre une méthode à la pensée critique améliorant la prise de décision par hypothèse plus ou moins plausible. Il faut se demander : "Parmi un ensemble d'hypothèses, de modèles, laquelle ou lequel est la ou le plus pausible ?". Selon l'approche fréquentiste, une hypothèse est émise ou choisie et si un résultat a une probabilité trop faible (le résultat est invraisemblable) par rapport à un certain seuil fixé (*threshold*), alors l'hypothèse est rejetée. L'approche fréquentiste est pour l'instant encore communément utilisée (e.g., mesure des effets d'un traitement sur l'organisme). Ainsi, le test de rejet d'hypothèses est l'intérêt principal de nombreuses études suivant ce type d'approches. En général, la vraisemblance des résultats est étudiée afin d'obtenir la *p-value* guidant le rejet de l'hypothèse.

Un reproche adressé aux bayésiens par les fréquentistes est la part de subjectivité évidente dont fait part l'estimation de l'a priori initial (la plausibilité accordée à l'hypothèse) de chaque hypothèse considérée. Lors d'une séquence d'événements, les autres a priori résultent de l'a posteriori obtenu à la suite de l'expérience réalisée. Ainsi, l'a priori initial influence de façon considérable les résultats mais les a priori ou a posteriori futurs deviennent en un sens moins subjectifs à chaque inférence.

Toutefois, parmi les reproches qu'il est possible d'émettre, les bayésiens peuvent eux aussi relever une part de subjectivité dans l'approche fréquentiste, bien que considérée sous une autre forme. Selon une approche fréquentiste, lorsque les conclusions de plusieurs études se contredisent pour une hypothèse (e.g., suite à de nouvelles découvertes, de nouvelles données, etc), il est aussi question d'un niveau de plausibilité pour chaque hypothèse dépendant aussi de divers facteurs. En ce sens, l'approche bayésienne semble à nouveau adéquate à cette tâche.

Avant de continuer la discussion du fréquentisme et, plus particulièrement, du bayésianisme, il est nécessaire de définir la formule de Bayes, à l'origine de la méthode et philosophie

bayésienne.

Afin d'y parvenir, nous définissons d'abord la loi des probabilités totales.

Théorème 2.3.1 (Loi des probabilités totales)**.** *Soit x un événement, D des données, H_i une hypothèse.*

$$P(x|D) = \sum_{H_i} P(H_i|D)P(x|D, H_i)$$

La loi des probabilités totales dit que pour prédire la probabilité d'un événement x sachant les données D, il faut calculer la somme parmi les hypothèses H_i du produit entre la probabilité des différentes hypothèses H_i sachant les données D et la probabilité de l'événement x sachant les données D et l'hypothèse H_i.

Ainsi, avec 2.3.1 et parce que $P(x|H)P(H|x) = P(H)P(x|H)$, la formule de Bayes peut être obtenue.

Théorème 2.3.2 (Formule de Bayes)**.** *Soit H une hypothèse parmi l'ensemble d'hypothèses H_i considérées et soit D les données disponibles. La formule de Bayes peut s'écrire :*

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_{H_i} P(D|H_i)P(H_i)}$$

Le terme de gauche $P(H|D)$ peut être interprété comme "la crédibilité de l'hypothèse (ou théorie) H sachant D". C'est-à-dire que l'hypothèse H est évaluée parmi d'autres hypothèses (voir 2.3.1). Quant au terme $P(D|H)$, il peut être interprété comme "la vraisemblance des données D considérant l'hypothèse H" (en admettant ici l'hypothèse comme vraie). Le terme $P(H)$ est le préjugé sur l'hypothèse H, l'a priori. Idéalement, il est l'état actuel des connaissances sur cette hypothèse et donc, mélangé à notre subjectivité, la crédibilité a priori de l'hypothèse H. Le dénominateur somme sur toutes les hypothèses à considérer.

Remarque. Dans un usage idéal de la formule de Bayes, il faudrait en fait, sans la moindre négligence, prendre en compte toutes les informations dont le monde dispose jusqu'au moment de son évaluation - bien que certaines informations soient évidemment moins conséquentes que d'autres. Nous appuyons aussi le fait que "la vraisemblance des données n'est pas la crédibilité de la théorie" (et ce d'autant plus en présence de forts à priori) [14]. Ainsi, une donnée peut être probable sachant une théorie, mais la théorie peut ne pas être crédible sachant la donnée (i.e., $P(D|T) \neq P(T|D)$). Lê Nguyen de la chaîne Science4All [15] donne l'exemple parlant suivant. "Un arabe est probablement musulman (avec 99% de chance) mais un musulman n'est probablement pas arabe (avec 27% de chance)".

En suivant une approche baéysienne, chaque fois qu'un évènement survient, les probabilités des hypothèses (c'est-à-dire l'a posteriori de chaque hypothèse, $P(H|D)$) sont ajustées. En reprenant l'exemple des dés, $P(X)$ est la probabilité de faire, par exemple, un 12 (sans sélectionner de dé en particulier). La vraisemblance du résultat fois l'a priori

14. Lê Nguyen (Science4All), La formule du savoir
15. *Ibid.*

(notre "croyance" a priori de l'hypothèse) est $P(X|H_i)P(H_i)$. Ainsi, lors d'une séquence d'évènements, l'étude d'un évènement Y survenant juste après un évènement X se servira de l'a posteriori obtenu comme a priori pour le calcul de ce nouvel a posteriori, pour la même hypothèse (bien entendu, cela se calcule pour toutes les hypothèses). Nous obtenons ainsi naturellement un raisonnement par inférence.

Dennis Lindley, 1928-2013, avait écrit : "Le postérieur d'aujourd'hui est le préjugé de demain.". Pour un Bayésien et en caricaturant, le préjugé devrait résulter de fastidieux calculs qui s'appuieraient avec une certaine précision sur l'énorme ensemble de données précédemment obtenues, jusqu'à cet instant. "Préjuger" pour un Bayésien, c'est juger avant d'avoir pris en compte les nouvelles données empiriques. En pratique, le préjugé est en fait un postérieur très compliqué à calculer. Le travail principal du Bayésien est alors de calculer au mieux ce préjugé.

À présent, il serait utile pour la suite du développement de revenir à l'approche fréquentiste [16]. Un postulat de fréquentiste est l'existence d'une vraie (i.e., résultat de la réalité) probabilité ou fréquence p bien définie qui garantirait alors la convergence vers une fréquence limite d'une séquence infinie d'expériences ou observations. Toutefois, la fréquence d'un événement ne converge pas nécessairement. De plus, il est rarement possible de pouvoir observer un événement sur toute la population souhaitée. Ceci implique de devoir estimer une fréquence en se basant sur l'échantillon obtenu.

Considérons un événement X, une population Y, f la fréquence empirique et p la vraie probabilité. Nous dirons que f est la fréquence de l'événement X dans un échantillon de la population Y alors que p est la fréquence de X dans la population entière Y. Ainsi, le fréquentiste s'intéresse à l'évaluation de la fréquence f d'un événement dans un certain échantillon (i.e., une partie de la population) afin de savoir si elle acceptable par rapport à son hypothèse de la fréquence limite p ; le but étant de dévier de p le moins possible. Un des plus importants théorèmes des probabilités, le théorême central limite (TCL), démontre que pour un ensemble de données suffisamment grand, les erreurs attendues (i.e., erreurs d'estimation de la vraie probabilité à partir de la fréquence f) sont distribuées selon une loi normale. Ceci est remarquable dans le sens où nous pourrions penser que cela dépendrait plutôt fortement de l'expérience en question. Une utilisation pratique et conséquence du TCL est que nous pouvons alors rejeter l'hypothèse de la probabilité p si la courbe de la normale ne contient pas la fréquence empirique f évaluée (i.e., si p est la vraie probabilité, alors la fréquence f devrait se retrouver sous la courbe de la normale).

L'approche épistémologique fréquentiste étant très présente dans la recherche scientifique, la reproductibilité des expériences (ou en tout cas leur caractère objectif en le sens qu'elles ne dépendent par exemple pas de l'expérimentateur) ainsi que la taille de l'échantillon disponible sont extrêment importantes. L'approche fréquentiste a en fait plusieurs limitations importantes qui ne peuvent être ignorées. Tout d'abord, elle suppose l'existence d'une "vraie" probabilité p sans pouvoir nécessairement prouver qu'elle existe - l'entité p est postulée par le modèle. Pour pallier à cela, il est possible d'utiliser des intervalles de confiance afin d'estimer avec une marge la valeur de probabilité (voir [8]). Ensuite, ce qui ferait état d'un énorme reproche par un Bayésien, est que l'approche fréquentiste néglige complètement le préjugé ou a priori

16. Lê Nguyen (Science4All), Le fréquentisme

construit par l'état actuel de nos connaissances [17]. Une dernière limitation du fréquentisme vient directement de son essence : il n'est utilisé que pour des phénomènes avec fréquences. C'est-à-dire qu'il faut une séquence d'événements observables pour estimer une fréquence. Malheureusement, nombreux champs d'étude, bien qu'importants et conséquent, ne disposent pas toujours de moyens pour appliquer une telle approche. C'est le cas en particulier pour des études de phénomènes fortement liés à l'état physique de l'Univers (i.e., son état complet passé, présent voire futur) telles que, par exemple, le changement climatique, des grands événements historiques, la formation de l'Univers, etc.

Le bayésianisme semble alors être une approche plus prometteuse pour se défaire de telles limitations. Plus particulièrement, il permettrait de rendre plus bénéfiques les études par *p-value* ainsi que l'état des connaissances et recherches actuelles.

Avant de discuter le sujet de la p-value et du "statistiquement significatif" fortement lié au fréquentisme, nous présentons un standard d'études scientifiques avec lequel ces notions sont utilisées : les tests randomisés et contrôlés en double aveugle [18]. Dans le cadre de ces processus test visant à répondre aux questions du type "Est-ce qu'un traitement a tel effet ?", deux groupes sont définis : le groupe test et le groupe de contrôle. Le terme "double aveugle" signifie que ni les sujets de l'expérience ni les expérimentateurs ne connaissent le label de chaque groupe. Par exemple, dans le cadre d'évaluation d'un médicament, le groupe test sera celui à qui le médicament sera donné (sans que les sujets ne soient au courant), contrairement au groupe contrôle à qui le médicament ne sera pas administré. Ceci permet d'éviter l'introduction de biais (e.g., régression à la moyenne, effet placebo) et d'ainsi assurer une meilleure qualité de données expérimentales malgré la quantité de données qui se voit diminuer du fait du le cadre strict imposé par ce type de tests (voir [16]). Comme vu précédemment avec l'approche fréquentiste, les tests randomisés et contrôlés en double aveugle mettent de côté un grand nombre de sujets d'études qu'ils ne permettent pas de traiter. Par ailleurs, les résultats de tels tests sont généralement analysés à travers le concept de *p-value* que nous allons à présent discuter ci-dessous [19].

Depuis son introduction dans la recherche scientifique (voir [6]), la p-value a apporté une référence de rigueur et validité sur laquelle de nombreux scientifiques se basent naturellement : "Quelle est la p-value obtenue ?". Ceci a donné lieu à une sorte de biais de publication où les études par p-value et "statistiquement significatif" sont favorisées (e.g., moins d'études avec un intitulé de la forme "Le traitement t n'a aucun effet" opposé à plus d'études avec un intitulé de la forme "Le traitement t a un effet E"). De fait, les études portant un achèvemant semblant plus spectaculaire prévaleront sur les autres, biaisant ainsi l'information globale partagée.

L'étude par p-value fonctionne sur le principe de réfutation. C'est-à-dire, des données D sont collectées puis utilisées afin de réfuter - par preuve dite "statistiquement signifivative - une hypothèse ou théorie H. Si en supposant H vraie les données sont trop invraisemblables, alors H est rejetée. Plus la p-value p_{val} est petite, plus il faudrait rejeter l'hypothèse H. Le seuil (i.e., la probabilité que les données D rejettent l'hypothèse H par p-value dans le cadre de l'hypothèse H) choisi dépend du sujet d'étude (e.g., modèle de ML, efficacité d'un

17. Hygiène Mentale, Ep28, Le Sophisme du Procureur (et quelques autres leçons bayésiennes)
18. Lê Nguyen (Science4All), Le standard ultime des sciences ! !
19. Lê Nguyen (Science4All), La plus grosse confusion des sciences : la p-value ! !

médicament, physique des particules) en question mais il est commun de retrouver les valeurs de 1%, 5% ou 1e-5%. Il faut bien comprendre qu'ainsi, pour une hypothèse vraie, le seuil de la p-value est égal à la probabilité de rejeter l'hypothèse par méthode de p-value. De ce fait, pour une hypothèse - considérée - vraie, un test statistique par p-value avec un seuil de $x\%$ a une probabilité de $x\%$ de rejeter l'hypothèse.

Une opposition bayésienne face à l'utilisaton décrite de la p-value serait de rappeler que la vraisemblance des données n'est pas la crédibilité de la théorie. Or, selon cette utilisation de la p-value, le manque de vraisemblance des données pousse (et suffit !) à rejeter une hypothèse (et donc sa crédibilité). Ce genre de raisonnement inadéquat peut (et a déjà) donner lieu à des confusions et mauvaises conclusions.

Une forte limitation de cette utilisation de la p-value est que le cadre peu explicite est celui d'une hypothèse admise vraie. Cependant, si l'hypothèse s'avère être rejetée par un certain ensemble de données, l'information quant à la vraisemblance de l'hypothèse sachant qu'elle a été rejetée est perdue. Malgré certaines publications (e.g., [62]) poussant à ne pas se baser sur le "statistiquement significatif", cette méthode de réfutation par preuve par p-value statistiquement significative est actuellement grandement utilisée quand bien même plusieurs problèmes [20] [21] pourraient lui être adressés (voir plus bas).

Une objection à son utilisation porte sur les conclusions de la forme "l'hypothèse est fausse avec une probabilité de 95% car il a été obtenu une p-value de 5% après expérience" qui sont contre-productives et issues directement de la définition même de p-value. Parmi les remarques résumées qu'émet [62] sur les travers de la p-value et du "statistiquement significatif", nous retrouvons, par exemple, les suivantes :

— "Ne pas baser ses conclusions uniquement sur le fait qu'une association ou un effet a été trouvé être "statistiquement significatif" (i.e., la p-value a passé un quelconque seuil arbitraire tel que $p < 0.005$)"
— "Ne pas croire qu'une association ou un effet existe just parce que c'était "statistiquement significatif"."
— "Ne pas croire qu'une association ou un effet est absent juste parce que ce n'était pas "statistiquement significatif"."
— "Ne pas croire que la p-value donne la probabilité que la chance seule a produit l'association ou effet observé ou la probabilité que l'hypothèse de test est vraie."

Un autre problème, plus théorique et conceptuel, est que, à l'infini, quelque soit le seuil de p-value choisi, une hypothèse vraie sera forcément rejetée. Par conséquent, un test conduit manquerait de "puissance statistique". Une hypothèse avec une probabilité de 1% d'être rejetée (et donc avec 99% de chance de ne pas l'être) aura une probabilité de $(99\%)^n$ de ne pas être abandonnée ou rejetée après avoir passé n tests. Or, une telle probabilité décroît exponentiellement vite. Après 500 tests, l'hypothèse/théorie a alors une chance de moins d'1% de ne pas être rejetée, et donc, par extension, une probabilité de 0 à l'infini. Ce qui mène à la conclusion que toute théorie, vraie ou fausse, à réfuter (ou non) par p-value devrait être rejetée et ce, indépendamment du seuil de p-value.

20. Lê Nguyen (Science4All), 5 sacrés travers de la science par p-value
21. Lê Nguyen (Science4All), 5 énormes travers de la science par réfutation

Ensuite, un problème qui ronge cette fois plutôt l'environnement des chercheurs est celui du *p-hacking*. Cela repose sur le fait d'effectuer certaines manipulations (e.g., effectuer un grand nombre de tests statistiques par p-value jusqu'à réduire suffisamment la p-value) visant à obtenir une nouvelle p-value suffisamment petite. Ce phénomène est alimenté et envenimé par le contexte difficile du système "*publish or perish*" duquel font partie les chercheurs encore non-établis, ainsi que par le manque de régulation sur les arrêts prématurés ou les tailles d'échantillons.

Un autre problème, potentiellement plus grave, concerne les possibilités après avoir rejeté toutes les théories par p-value. Ceci est renforcé par la quête de théories scientifiques "raisonnables" pour des systèmes très (trop ?) complexes. L'idée, par rapport aux capacités actuelles, est d'admettre des théories simples et efficaces (e.g., en termes de complexité spatiale et temporelle) capables de suffisamment bien décrire des phénomèmnes pour être acceptées à la place de théories qui requerraient une puissance de calculs trop supérieure voire innatteignable.

Une opposition plutôt forte du bayésianisme est que toutes les théories ne se valent pas et ont un a priori. Ainsi, l'a priori et donc la crédibilité donnée à une théorie est important. Dans le même ordre d'idée, le rasoir d'Occam (aussi combiné à une approche bayésienne, voir [42]) renforce l'idée que les théories ne se valent pas et devraient, dans ce cas, être jaugées par le critère de simplicité.

Nous en venons ainsi, de nouveau, à l'importance de l'a priori dans la façon de situer la crédibilité des résultats d'une étude. Des conclusions directes suite à une unique publication scientifique manquent de robustesse et devraient plutôt servir à ajuster la crédibilité d'une hypothèse plutôt que de la mettre en premier plan comme une nouvelle vérité absolue sans avoir pris en compte sa crédibilité a priori. Ceci revient à clairement distinguer le consensus scientifique de "la publication scientifique la plus récente ayant passé le test" : leur crédibilité est différente. Cela doit amener à se demander si la publication a été justement validée ou si elle se trouve être le fruit de biais et fluctuations statistiques. Faire preuve d'un certain scepticisme envers les publications scientifiques validées par ce type de tests ou méthodes semble être naturel au vu des problèmes précédemment mentionnés (voir [28] pour une critique de l'état de la confiance en les publications).

Pour conclure cette section et chapitre, nous relevons un problème conséquent du bayésianisme : il requiert une puissance de calculs immense afin de calculer les probabilités [22] nécessaires à une application exacte de sa théorie. Son utilisation pure pour les problèmes compliqués utiles à résoudre semble vouée à ne jamais voir le jour. Toutefois, il reste très raisonnable de travailler avec un pseudo-bayésianisme, ou plutôt, un bayésianisme pragmatique[25].

22. Lê Nguyen (Science4All), L'argument fatal contre le bayésianisme

3

TRAVAUX CONNEXES

Parmi les nombreux articles et ouvrages auxquels nous nous sommes interessés et référés, certains travaux, d'amplitude plus ou moins générale, ont été des lectures se situant précisément dans le contexte établi et abordant des enjeux importants propres au sujet d'étude du présent article. Ainsi, nous mentionnons ci-dessous, non-exhaustivement quelques-uns de ces articles ou ouvrages avec ce qu'ils apportent.

Pour une approche globale de la fairness et des biais dans les algorithmes, des revues de littérature et de l'état de l'art sont et ont été publiées. Par exemple "Fairness In Machine Learning : A Survey"[10] propose une revue générale de l'état de la fairness dans le domaine du Machine Learning afin de rendre cet important aspect plus accessible aux nouveaux arrivants. Il ne se focalise pas en détails sur la détection et correction des biais, mais plutôt les différentes écoles de pensée des divers domaines où l'enjeu de fairness est important. Dans un ordre d'idée comparable, cette fois avec des comparaisons plus directes, l'article "A comparative study of fairness-enhancing interventions in Machine Learning"[19] fait le point sur de nombreuses interventions ayant eu lieu et méthodes ayant été proposées pour améliorer la fairness dans le domaine du Machine Learning. Il propose également un *benchmark* qui peut s'avérer utile en tant que support pour d'autres études qui ont besoin de comparer leurs résultats exprérimentaux.

En ce qui concerne le NLP, nous mentionnons "It's All in the Name : Mitigating Gender Bias with Name-Based Counterfactual Data Substitution" [39] pour la mitigation des biais de genre en supplément à ce que nous proposons à la section 4.2.4 ainsi que "Lipstick on a Pig : Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them"[21] pour certaines difficultés rencontrées lors de tentatives de correction de biais. "Mitigating Gender Bias in Natural Language Processing : Literature Review"[54] propose également une revue de littérature propre au NLP qui vient supplémenter plus généralement la section 4.2.4.

Le bayésiannisme, que nous considérons comme philosophie prometteuse quant aux avancées qu'il reste à faire dans le domaine de la fairness et ses approches face aux biais, a vu de nombreux articles se reposant sur les approches qu'il offre et addressant des problématiques plus ou moins précises. Par exemple, pour des sujets plus spécifiques, [67], [1] ou encore [40]. L'ouvrage "La Formule du Savoir"[25] est une bonne entrée à sa philosophie ainsi qu'au développement de l'approche bayésienne.

La fairness étant dans ce papier le concept principal traité et discuté ainsi qu'une finalité évidente en matière d'éthique d'IA, les aspects de *transparency* ("transparence") et *accoun-*

tability (”responsabilité”) supportent aussi la fairness et sont eux-mêmes dignes d’un intérêt certain. L’article ”50 Years of Test (Un)Fairness : Lessons for Machine Learning”[27] discute la fairness sur une plus grande période que les autres publications mentionnées, ainsi que les aspects importants que sont l’accountability et la transparency.

Dans une perspective élargie, ”Le fabuleux chantier, rendre l’Intelligence Artificielle robustement bénéfique”[37] délivre d’intéressantes discussions et approfondissements de l’éthique des IAs et du débiaisement en mettant en évidence l’aspect du ”robustement bénéfique”. Il apparaît clair que l’éthique en matière d’Intelligence Artificielle est un chantier pour lequel il reste encore du travail.

Finalement, un ouvrage qui peut être jugé particulièrement utile est le travail ”Fairness and Machine Learning” [2] auquel il est pratique de se référer pour une description et analyse de nombreuses interventions propres à la fairness en IA.

4

DÉVELOPPEMENT

4.1 DES PROBLÈMES DE TAILLE

Autour des biais, nous retrouvons des fondements particulièrement importants mais toutefois encore fragiles. L'étude des biais et son avancée rencontrent des problèmes à résoudre absolument afin de pouvoir progresser sainement et correctement établir les enjeux qui leur sont liés. En plus de l'évolution croissante du nombre de publications scientifiques sur l'IA, le Machine Learning, etc, il existe déjà de nombreuses études sur la fairness et les biais [1].

Un problème que tout programmeur peut facilement ressentir est celui de l'immense complexité que peuvent atteindre les proccessus algorithmiques. Plus les puissances de calcul et technologies s'améliorent, plus il devient compliqué pour l'humain de maintenir une compréhension des opérations et interactions ayant lieu dans un système. Ainsi, le phénomène de boîte noire (redevenu populaire après le développement du *deep learning*) devient plus présent tout en ne semblant pas gêner grand nombre d'acteurs principaux de par le bénéfice de performances accrues. À cela peut aussi venir s'ajouter les contributions imprédictibles des utilisateurs des algorithmes pour lesquels ces utilisateurs fournissent des données d'entrée particulières et nouvelles (e.g., un site internet enregistrant les choix des utilisateurs ou le temps passé sur une page) [2] [3].

Quand bien même les efforts portés à la complexité des algorithmes aboutiraient à une quelconque amélioration, le problème de disponibilité et collection de données - à fournir aux algorithmes - reste un poids pour la progression souhaitée. En effet, il existe diverses raisons (e.g., manque absolu de données existantes pour le problème en question, minorité d'un groupe augmentant ainsi la difficulté de collection des données, lois et régulations compliquant voire interdisant la collection de données) pour lesquelles obtenir des données utiles au problème en question est délicat ou même impossible. Ce problème est par la suite aussi exacerbé par le besoin d'une grande quantité de données au préalable pour des algorithmes de Machine Learning ou Deep Learning ayant pour but une précision et robustesse justifiées. Par ailleurs, c'est ici un point intéressant qui est soulevé, en particulier par rapport à la discussion qui suivra plus tard sur l'approche bayésienne.

Pour appuyer le problème du manque général de données utilisables, il est à noter la partie légale à laquelle toute entité au sens large est sujette. Que ce soit en terme de *privacy*, atteinte

1. Voir, par exemple, les références utilisées pour cet article
2. Algorithmic bias detection and mitigation : Best practices and policies to reduce consumer harms
3. Controlling Machine Learning algorithms and their biases

à la liberté, etc, les lois et régulations varient fortement d'un endroit (i.e., continent, pays, etc) à l'autre. La RGPD [4] ("GDPR" en anglais) est un exemple bien connu d'opposition légale à l'obtention et manipulation de données personnelles à protéger et à l'application de certains algorithmes de décision (e.g., pour un système juridique au but d'évaluer la probabilité d'un coupable). Ainsi, bien que nous ne portons pas là de jugement sur le bénéfice ou déficit de tels systèmes - la RGDP pourrait en fait réduire les biais en contrôlant davantage les algorithmes par régulation - mis en place, la collection de données souhaitée se trouve encore plus contrainte. Un organisme de régulation tel que la RGPD ne résout certainement pas tous les problèmes constatés, néanmoins elle permet d'instaurer un cadre aidant à la structuration d'un certain ordre dont les enjeux liés au contrôle des biais ont besoin. Voir [22] pour des articles plus ciblés sur la RGPD qui ne peut pas être négligée par les chercheurs sujets au lois de l'Union Européenne, entre autres.

Précédemment, l'aspect de *fairness* dans les algorithmes a été introduit. Cependant, sa définition n'est pas (encore ?) bien établie. Un dilemme conséquent est celui de l'apparente dualité performance-fairness, notamment dans les algorithmes de Machine Learning et Deep Learning. De plus, un algorithme satisfisant une égalité de traitement des groupes concernés ne garantit pas une équité parmi ces mêmes groupes. Ainsi, bien que compliqué et souvent délicat (e.g., définition et distinction de ces différences, conflits d'intérêt), prendre en compte les différences entre ces groupes est essentiel pour atteindre la souhaitée *fairness*.

Une autre propriété importante qui a été mentionnée est celle de *transparency* des algorithmes. Le manque de transparence (i.e., celer l'algorithme lui-même et son design) est un des obstacles majeurs à la surveillance ou détection des algorithmes qui manqueraient de fairness, en plus des autres aspects mentionnés. Notamment, une entreprise notable (e.g., Google, YouTube) avec une grande influence et pouvoir considérable ne souhaiterait pas forcément rendre public les algorithmes utilisés dans ces systèmes. Ainsi, il est bien plus difficile de savoir si les résultats de leurs algorithmes sont manipulés et/ou si des biais sont introduits volontairement ou non. Malheureusement, partager son travail et ses avancées naturellement n'est pas l'attitude adoptée dans un contexte où les enjeux économiques et parfois sociaux sont si marqués.

4.2 UNE PARTIE DE L'ÉTAT DE L'ART

À ce jour, de nombreuses méthodes et approches visant à atteindre une fairness dans les algorithmes et la contrôler ont été créées, améliorées, discutées. Cependant, l'aboutissement final de la fairness semble encore bien loin.

Dans la prochaine sous-section, nous présentons brièvement plusieurs méthodes habituelles dites observationnelles qui sont, et surtout ont été, utilisées. Par la suite, nous nous penchons sur une approche et discussion de la fairness pour les algorithmes de NLP. Finalement, une

4. RGPD : Régulation Générale de la Protection des Données. Elle a été adoptée par l'Union Européenne en 2016 et est composé d'un grand nombre d'articles définissant les terminologies, limitations et aspects à considérer quant à la protection des données. Voir aussi : The impact of the GDRP on AI, publié par le Parlement européen

approche bayésienne, encore nouvelle mais prometteuse, est étudiée à travers la publication "Bayesian Fairness"[17].

4.2.1 *Critères et approches de fairness trop imparfaits*

Comme discuté plus tôt dans cet article, il y a en général plusieurs facteurs (e.g., personnes, méthodes, données) pouvant être responsables de l'ajout de biais, de discrimination, de manque de fairness. Nous présentons à présent plusieurs méthodes de base faisant principalement usage de caractéristiques observables/mesurées (e.g., les attributs à disposition, les labels/données réelles, les prédictions du modèle), à l'opposée de caractéristiques propres au modèle, par exemple. L'idée étant que les méthodes qui vont suivre peuvent se baser sur de telles caractéristiques afin de mesurer la fairness ou, similairement, détecter un niveau particulier de discrimination.

Pour des raisons de simplification, nous nous plaçons dans un contexte de classification (là où les discriminations sont par essence marquées) par un modèle et définissons les représentations mathématiques suivantes [5] [6]. L'ensemble des attributs est noté $\chi \subseteq \mathbb{R}^k$ et χ peut contenir ou non un ensemble d'attributs protégées A (e.g., origine éthnique, genre, croyance religieuse, orientation sexuelle). Nous utilisons ainsi le vecteur d'attributs $\mathbf{x} \in \chi$ pour prédire le label $Y \in \{0, ..., C\}$ où $C \in \mathbb{N}$, avec $\mathbf{x}$ étant ainsi séparable en les valeurs des attributs protégées, nous notons $\mathbf{x}^{prot}$, et celles des attributs acceptées que nous notons $\mathbf{x}^{acc}$. Ainsi, nous pouvons avoir $\mathbf{x} = [\mathbf{x}^{prot}, \mathbf{x}^{acc}]$. Nous définissons alors la fonction de modèle $\phi_\theta : \chi \longrightarrow \{0, ..., C\}$, et $\mathbf{x} \mapsto \phi_\theta(\mathbf{x}) = \hat{Y}$. Une prédiction effectuée par le modèle est dénotée $\hat{Y} = \hat{y}$. Remarquons que dans le cas d'un problème de classification binaire, nous avons les labels possibles $Y = 0$ ou $Y = 1$.

Ainsi, les méthodes se servant simplement de caractéristiques observables peuvent comparer les différences de résultats en interchangeant des attributs sensibles, protégées ou qui leur sont corrélées. Simultanément, une opération similaire est effectuée avec les groupes protégés et non protégés. Il est donc possible de conclure sur un éventuel problème de fairness.

Nous notons que retirer simplement les attributs protégées pourrait sembler être une bonne idée - car nous savons que ce sont des attributs à fort potentiel discriminatoire -, mais nous trouvons facilement des exemples où une telle pratique pourrait en fait aggraver le problème [2]. Cette méthode naïve est appelée *unawareness* ("ignorance" en français). Le classifieur n'utilisant que les attributs acceptées (i.e., ignore les attributs protégées), nous avons : $\phi_\theta(\mathbf{x}) = \phi_\theta(\mathbf{x}^{acc})$. Cette méthode a l'avantage de concorder avec la notion de *disparate treatment* retrouvée à l'origine de nombreux textes de loi. Malheureusement, retirer des attributs protégées n'empêche pas l'ensemble restant de variables d'être corrélées à l'attribut protégée retirée et ne résout ainsi pas le problème des proxies non plus.

Voyons à présent les approches et critères suivants[3].

---

5. Slides "AI Fairness" du cours Intelligence Artificielle, UniGe, par le Professeur Marchand-Maillet

6. fairmlbook

4.2.1.1 *Demographic parity*

Demographic parity (aussi appelé *group fairness criterion* ou parité de groupe en français) se concentre sur l'indépendance - au sens statistique - entre les prédictions $\hat{Y}$ et l'appartenance à un groupe protégé $A = a$. Ceci signifie que le pourcentage d'individus dans un groupe protégé qui sont classifiés comme positif doit être le même que le pourcentage global de tous les individus car il est indépendant du groupe. Elle fait partie d'une catégorie plus large appelée *group fairness*. Nous notons ici que nous pourrions à nouveau définir une fonction ϕ_θ afin d'utiliser $\phi_\theta(\mathbf{x})$ et ne pas considérer uniquement le cas des positifs $\hat{Y} = 1$. Pour des raisons de simplicité et cohérence, son utilisation n'a pas été jugée nécessaire.

Soit a appartenant à un groupe protégé et b un différent, le critère de *Demographic parity* est satisfait si nous avons :

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b) \tag{1}$$

Ainsi, la probabilité que la prédiction de la classification soit positive (i.e., $\hat{Y} = 1$) est indépendante de l'appartenance à un groupe protégé.

Nous remarquons aussi que ce critère de fairness de dépend pas des valeurs groundtruth Y. Ceci est un avantage et point important de cette approche car en pratique il n'est souvent pas possible d'obtenir les valeurs groundtruth ou de les évaluer. Pire, une tentative d'évaluation de ces valeurs groundtruth pourraient précipiter une insértion volontaire ou non de biais et ainsi aggravé le problème de fairness à résoudre. De plus, la contrainte de non-corrélation entre les valeurs groundtruth et les variables protégées réduit considérablement les possibilités pratiques de son utilisation concrète. Finalement, tenter de satisfaire (1) pourrait donner lieu à des compensations forcées entre les groupes notamment en favorisant excessivement le groupe protégé. Ceci est appelé "discrimination positive".

4.2.1.2 *Equal opportunity*

Cette approche de critère de fairness (appelée en français "égalité des chances") se concentre cette fois sur les observations d'individus similaires obtenant des résultats similaires. Elle fait partie de la catégorie de *group fairness*. *Equal opportunity*[2] prend en compte les classifications positives uniquement (voir ci-dessous *Equalized odds* pour une version plus stricte). Ainsi, la contrainte est sur les vrai positifs et pour a appartenant à un groupe protégé et b un différent, nous avons :

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b) \tag{2}$$

4.2.1.3 *Equalized odds*

Equalized odds[2] étend les contraintes de *Equal opportunity* aux classifications négatives (en plus des positives) et aux mauvaises classifications, où chacune d'entre elles doit être égale lorsque le groupe protégé sélectionné est comparé au reste de la population. Ainsi, il faut pouvoir obtenir le même nombre de vrai positifs et de faux positifs parmi tous les groupes protégés. Les probabilités considérées sont de type "probabilité jointe de la prédiction $\hat{Y}$ conditionné par la valeur ground truth Y et du groupe protégé A". Soit a appartenant à un

groupe protégé et b un différent.

Pour la contrainte de vrai positifs, nous avons :

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b) \tag{3}$$

Pour la contrainte de faux positifs, nous avons :

$$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = b) \tag{4}$$

Nous rappelons que dans le cas d'un classifieur continu, définir un seuil (*threshold*) est nécessaire.

4.2.1.4 *Calibration*

Calibration est une autre approche de critère de fairness qui inclut potentiellement les variables sensibles mais qui ne se base pas sur les valeurs ground truth. C'est une différence notable avec les précédentes. Elle repose aussi sur une contrainte d'indépendance conditionnelle. Soit a appartenant à un groupe protégé et b un différent. Soit $\hat{y} \in \{0, 1\}$ l'évaluation de la variable aléatoire $\hat{Y}$. Nous avons :

$$P(Y = 1|\hat{Y} = \hat{y}, A = a) = P(Y = 1|\hat{Y} = \hat{y}, A = b) \tag{5}$$

Certaines publications[2] appellent cette approche *sufficiency* - la calibration par groupe implique la sufficience - car, si la condition est satisfaite, le classifieur, soit $\hat{Y}$, est suffisant - les variables Y et A sont claires de par le contexte - pour la prédiction et les variables protégées sont incluses. La différence étant dans la façon de penser à cette notion : en terme de valeurs obtenues (positives ou négatives) ou en terme de calibration. Il arrive en pratique qu'il soit utile d'interpréter les valeurs de fonctions de coût (i.e., un score au sens large, pas restreint à un pur coût) en probabilités. Ainsi, pour un classifieur dit "correctement calibré", la contrainte est satisfaite et devient :

$$P(Y = 1|\hat{Y} = \hat{y}) = \hat{y}) \tag{6}$$

C'est-à-dire, en considérant les valeurs ground truth positives, l'ensemble des résultats (et non pas le résultat d'un individu uniquement) avec un score de $\hat{y}$ contient une proportion $\hat{y}$ de résultats positifs.

4.2.2 *Limites des méthodes observationnelles*

Certaines des tentatives d'amélioration (e.g., [66]) des approches observationnelles mentionnées ont été de s'attaquer à la phase de pré-traitement en tentant de rendre les données d'entraînement moins biaisées, plus équitables. Combiné à cela, l'algorithme est traîné de sorte à rassembler et maintenir un maximum d'information tout en célant les variables protégées. Malgré des performances satisfaisantes, un constat critique est effectué sur la trop grande apparence de faux positifs pour certains groupes dans le but de balancer les résultats d'autres afin d'essayer d'atteindre une fairness. En plus d'avoir développé et affiné certaines méthodes observationnelles (e.g., usage de *threshold* sur les probabilités produites par le classifieur), [24] ont conduit une analyse approfondie sur les limites de ce type de méthodes.

Plus généralement, *Demographic parity* souffre de plusieurs défauts dont ceux mentionnés ci-dessus et est difficilement acceptable comme une bonne méthode de rétablissement d'un déséquilibre entre groupes. Ainsi, [24] introduisent et discutent *Equal opportunity* et *Equal odds* en réponse aux limitations de *Demographic parity*. Cependant, ces deux critères de fairness ont aussi des défauts qui ne sont pas négligeables. Les critiques étant qu'ils ne résolvent pas l'éventuelle présence de biais dans les données originelles ou encore la nécessité de disposer des informations sur toutes les variables protégées. [65] définit la terminologie *disparate mistreatment* [7] et modifie le critère de base en passant à un problème de minimisation convexe. Additionnellement, un problème important et qui persiste pour ces deux approches est celui du besoin des valeurs groundtruth (ce qui n'est pas le cas pour *Demographic parity*) qui rend l'utilisation de ces critères délicats en pratique. Finalement, l'incompatibilité des critères de fairness (i.e., utiliser plusieurs critères en même temps afin d'optimiser la fairness) reste un problème commun et considérable (e.g., *Demographic parity* ne doit pas utiliser les attributs protégées mais *Equal odds* et *Equal opportunity* reposent sur cela).

Pour conclure cette section, nous relevons que, par exemple, la calibration peut naturellement être utilisée dans un contexte bayésien (voir [60]) et qu'ainsi ces approches ne sont pas à ignorer. Par ailleurs, [45] définit et propose des approches par *causal inference* (avec graphes utilisés comme support) qui pallient à certains problèmes des approches observationnelles et, par exemple, diminue le problème de ne pas savoir si des attributs protégées ont été utilisées ou non en étudiant les relations de causalité dans les données et entre les variables. D'autres publications plus nouvelles se sont par la suite fortement appuyées sur les travaux de [45] afin de proposer des approches modifiées et prometteuses ; voir [12] pour une étude de la fairness avec approche causale bayésienne, [57] et [64] pour des travaux intéressants par inférence causale. Bien entendu, de nombreux autres critères de fairness existent, souvent similaires par certains aspects, mais n'ont pas été vus dans cet article (e.g., *counterfactual fairness*[33], *treatment equality*[61], *equalizing disincentives*[29], etc).

4.2.3 *Fairness avec approche Bayésienne*

Cette section se base sur et relève principalement le travail produit dans l'article [17] "*Bayesian Fairness*" par Christos Dimitrakakis, Yang Liu, David C. Parkes, Goran Radanovic. Plusieurs références citées ont aussi été étudiées. Le choix - applicabilité, généralité, résultats expérimentaux satisfaisants - a été fait après avoir étudié plusieurs articles présentant une approche bayésienne.

Rappelons qu'une motivation à l'approche bayésienne est l'incertitude trop commune provenant des modèles probabilistiques utilisés et de leurs paramètres. Elle est renforcée par l'insuffisance de données à disposition ainsi que des systèmes où les actions influencent les informations à disposition.

Nous comprenons jusque là que les modèles probabilistiques représentatifs de notre monde sont souvent délicats à utiliser avec une certitude assurée. Ainsi, une approche bayésienne

7. Lorsque les taux de faux positifs et faux négatifs produits par un classifieur ne sont pas égaux parmi les groupes protégés

pour la fairness et traitement des biais peut être adaptée et mérite d'être méticuleusement investiguée. Dans cette section, nous étudions le développement des recherches de l'article "Bayesian Fairness"[17]. Plutôt que de proposer une nouvelle définition de fairness, il introduit une perspective bayésienne sur la fairness et montre comment une approche bayésienne peut amener à de meilleures décisions - plus équitables - malgré les modèles encore imparfaits utilisés.

L'approche bayésienne permettrait d'obtenir des informations supplémentaires - potentiellement nécessaires - ainsi que de faire meilleur usage de l'information disponible et des modèles utilisés. [17] propose alors un framework bien défini se basant sur une approche bayésienne. L'introduction au bayésianisme donnée à la section Bayésianisme pose les bases utiles à la compréhension - technique mais aussi philosophique - de l'approche étudiée.

Le contexte suivant est donné : un preneur de décision (DM) effectue une suite de choix selon une stratégie (ou règle) π afin de maximiser un certain résultat, un certain profit ou une certaine performance u. Ainsi, il doit compromettre entre la performance voulue et une contrainte de fairness f. Nous admettons l'existence d'une distribution de probabilité P qui permet d'écrire le problème de décision comme :

$$\max_{\pi}(1-\lambda)\mathbb{E}_P^{\pi}u - \lambda\mathbb{E}_P^{\pi}f \tag{7}$$

, avec λ le *trade-off* ("compromis") entre la fairness voulue et la performance.

Un premier aspect bayésien réside au paramètre de croyance $\beta \in \mathcal{B}$ du DM par rapport à une famille de distributions $\mathscr{P} \triangleq \{P_\theta \mid \theta \in \Theta\}$ pour laquelle il est possible que $P_{\theta^*} = P$ pour un certain paramètre θ^* (i.e., la distribution réelle P fait partie de cette famille).

Par rapport aux informations disponibles, la stratégie π du DM définit quelles actions $a_t \in \mathcal{A}$ prendre à chaque intervalle de temps t. À un temps t, le DM observe des données $x_t \in \mathcal{X}$ puis, en fonction de β_t, prend une décision a_t. Le but du DM étant de maximiser la performance u, elle est définie par la fonction $u : \mathcal{A} \times \mathcal{Y} \longrightarrow \mathbb{R}$, où $\mathcal{Y}$ est un ensemble de résultats.

L'approche ou méthodologie bayésienne de [17] se base sur les notions d'indépendance conditionnelle (voir aussi la section Critères et approches de fairness trop imparfaits) et est construite sur un critère de fairness nommé *balance* introduit par [32]. *Balance* est en fait assez proche de la définition de critère de fairness appelée *calibration* vue à la section Calibration dans le sens où le but général est que la probabilité estimée ne dépende pas du groupe auquel appartient un individu. Ceci laisserait même penser qu'il serait possible de les conjuguer dans leur utilisation ([32] montre que ce n'est pas possible sous certaines conditions). Dans son ensemble, le DM prend des décisions par rapport à la fairness mesurée dans les différents modèles, pondérée par leur probabilité.

Afin de permettre à la fairness d'être une caractéristique intrinsèque de la stratégie de décision, [17] proposent de séparer clairement les paramètres du modèle de la stratégie de décision et des informations à disposition du DM. Ils relèvent aussi qu'en considérant l'incertitude parmi les modèles de façon bayésienne (opposée à une façon non-bayésienne), la stratégie de décision est drastiquement modifiée. De plus, les algorithmes bayésiens de descente de

gradient développés par [17] considèrent ainsi l'incertitude du modèle et sa fairness, par rapport aux informations du DM.

Une action a est dite *test-fair* par rapport au résultat y et la variable sensible z si y est indépendante de z sachant a et θ. Nous introduisons à présent deux définitions de [17] basées sur les travaux de [13] et [32].

Definition 4.2.1 (Règle de décision calibrée)**.** Une règle de décision $\pi(a|x)$ est dite *calibrée* par rapport à une distribution P_θ si y et z sont indépendantes pour n'importe quelle action a effectuée. Avec P_θ^π la distribution induite par P_θ et la règle de décision π, nous notons la condition comme :

$$P_\theta^\pi(y, z \mid a) = P_\theta^\pi(y \mid a) P_\theta^\pi(z \mid a) \tag{8}$$

Definition 4.2.2 (Règle de décision balancée)**.** Une règle de décision $\pi(a|x)$ est dite *balancée* par rapport à une distribution P_θ si a et z sont indépendantes pour tout y. Avec P_θ^π la distribution induite par P_θ et la règle de décision π, nous notons la condition comme :

$$P_\theta^\pi(a, z \mid y) = P_\theta^\pi(a \mid y) P_\theta^\pi(z \mid y) \tag{9}$$

Dans la majorité des cas (voir [32] et "Théorème 3" de [17]), les deux conditions mentionnées ci-dessus ne peuvent être achevées simultanément. Ainsi, il faut choisir quelle règle de décision utiliser en pratique. [17] choisissent de travailler avec la définition de décision balancée (Règle de décision balancée) qui se conjugue mieux face aux incertitudes d'un modèle.

Toujours selon [17], nous nous plaçons à présent dans le contexte d'un problème de décision statistique concret afin d'introduire une formulation bayésienne de règle de décision. Considérons le schéma de la figure 1 avec explications. Le problème de décision est le suivant.

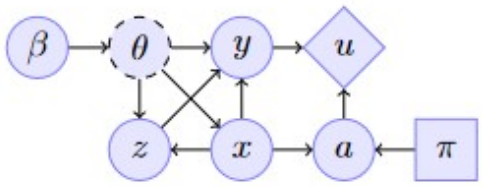

FIGURE 1 – Schéma du problème de décision. a est une action, u la performance, x une observation, y un résultat (pas observable et généré en fonction d'une distribution dépendant de x et pas du DM), z une variable sensible, θ un paramètre (inconnu) de distribution, π la stratégie de décision et β la croyance du DM sur la distribution (l'approche bayésienne relève cette croyance qui est un posterior résultant de l'a priori du DM ainsi que des données disponibles).

Le DM observe un événement $x \in \mathcal{X}$ et effectue une action (décision dépendant de la stratégie π) $a \in \mathcal{A}$. Il en résulte une performance $u(y, a)$ en fonction de l'action a et du résultat vrai $y \in \mathcal{Y}$ généré par une distribution $P_\theta(y|x)$. La croyance β introduite au début de cette

section est essentielle dans les stratégies de décision bayésienne de [17] - d'abord sans prendre en compte l'aspect de fairness, puis en l'incluant - qui vont être présentées ci-dessous. En particulier, il faut comprendre que les décisions prises par le DM sont basées sur la croyance posterior β obtenue.

Definition 4.2.3 (Règle de décision bayésienne optimale)**.** La règle de décision bayésienne optimale est définie par et notée :

$$\pi^*(\beta, x) \in \arg\max_{a \in \mathcal{A}} u_\beta(a \mid x) \tag{10}$$

, avec $u_\beta(a \mid x) \triangleq \sum_y u(y,a)\mathbb{P}_\beta(y \mid x)$ où $\mathbb{P}_\beta(y \mid x) \triangleq \int_\Theta P_\theta(y \mid x) d\beta(\theta)$ est la distribution marginale (i.e., la distribution de probabilité des variables sans être conditionné par les variables a évidemment et z) des résultats conditionnée par les observations x selon la croyance β du DM. $\pi^* : \mathcal{B} \times \mathcal{X} \to \mathcal{A}$ est donc une stratégie déterministe qui maximise la performance en espérance. Nous notons que cette règle de décision ne dépend pas directement de la variable sensible z.

Dans le cadre - introduit au début de la section - du DM effectuant des observations et choix à plusieurs temps t et ayant pour but de maximiser en espérance la performance globale définie par $U \triangleq \sum_{t=1}^{T} u(y_t, a_t)$ en trouvant la stratégie de décision adéquate, la règle de décision Règle de décision bayésienne optimale est en effet optimale. Cependant, elle n'inclut pas encore d'aspect de fairness pour lequel l'intérêt versera vers une optimisation raisonnable du *trade-off* performance-fairness. Pour cela, [17] définit une règle de décision bayésienne similaire à Règle de décision bayésienne optimale, mais incorporant la notion de *balance* vue dans Règle de décision balancée pour la fairness de la stratégie de décision. Une première étape est de redéfinir le problème de maximisation (7) de façon bayésienne. [17] le définit comme :

$$\max_{\pi} \int_\Theta [(1-\lambda)\mathbb{E}_\theta^\pi u - \lambda \mathbb{E}_\theta^\pi f] \, \mathrm{d}\beta(\theta) \tag{11}$$

Nous pouvons alors introduire ce que [17] définit et nomme "*Bayesian balance*". Comme expliqué dans la section Bayésianisme, l'approche bayésienne s'efforce à considérer tous les modèles possibles, et dans le cas de *bayesian balance*, l'impact des décisions prises qui vont à leur tour dépendre de l'évolution de l'état probabiliste des modèles. La définition de *bayesian balance* mesure la déviation de la stratégie π choisie par rapport à la balance idéale, puis la pondère par la probabilité du modèle. La mesure prend en compte la déviation (pour laquelle la p-norme est utilisée) de la fairness dans le cadre du problème de décision avec un DM bayésien.

Definition 4.2.4 (Bayesian balance)**.** Une règle de décision $\pi(\cdot)$ est dite (α, p)-Bayes-balancée pour une croyance β si elle satisfait la condition :

$$f(\pi) \triangleq \int_\Theta \sum_{a,y,z} \mid \sum_x \pi(a \mid x) \left[P_\theta(x,z \mid y) - P_\theta(x \mid y) P_\theta(z \mid y)\right] \mid^p \, \mathrm{d}\beta(\theta) \leq \alpha^p \tag{12}$$

Il est important de noter que, dans cette mesure, chaque paramètre θ possible est considéré. Ceci est alors différent d'une définition de *marginal balance* ("balance marginale") pour laquelle la mesure est effectuée en un seul point (différentes valeurs des paramètres θ ne sont pas prises en compte). Une critique faite à l'usage d'une telle définition serait que

le modèle marginal serait alors jugé comme l'unique modèle correct, ce qui amplifierait l'inéquité par rapport aux autres modèles malgré quils aient une forte probabilité. Or, le but est d'optimiser la fairness pour les modèles à haute probabilité et la maintenir de façon raisonnable pour les autres. Nous rappelons que la *bayesian balance* obtenue n'est pas nécessairement celle du vrai modèle. Toutefois, une règle de décision équitable selon la *bayesian balance* pour une croyance β suffisamment proche du vrai modèle implique une fairness par rapport au vrai modèle.

Par la suite, [17] étend l'approche étudiée aux problèmes possédant une caractéristique séquentielle (i.e., les actions du DM influencent les informations utilisables) et propose une analyse expérimentale de tels cas. Ils se basent sur un problème d'optimisation - qui sera adapté de façon à ce que le changement de croyance soit pris en compte - présenté par [48] pour lequel aussi les futures croyances (à un certain temps t) sont directement utilisées par la fonction de stratégie π (qui donnera une probabilité par rapport à une action). Une implication est que le DM doit considérer une stratégie de décision qu'il devra adapter (puisque sa croyance change en fonction des données à disposition qui dépendent elles-mêmes de ses croyances et donc de la stratégie de décision).

Les algorithmes utilisés par [17] sont des algorithmes de descente de gradient stochastiques qui se concentrent sur les optimisations discutées plus tôt où les objets mathématiques principaux sont les différents modèles (définis par les paramètres θ) et la croyance $\beta(\theta)$ sur la distribution de probabilité, avec lesquels le but est de trouver une stratégie de décision adéquate au *trade-off* performance-balance et donc par extension à celui de performance-fairness. Pour les résultats d'expériences (sur un ensemble de données synthétiques pour lequel l'exacte distribution selon laquelle il a été généré sera utilisée pour la mesure, ainsi qu'un ensemble de données de COMPAS [8] pour lequel ce sera cette fois une distribution empirique). Lors des expériences, le framework bayésien est comparé à l'approche à performance optimale sans compromis pour la fairness (le modèle marginal est considéré comme le vrai modèle au contraire de l'approche bayésienne qui prend en compte plusieurs modèles pour optimiser la stratégie de décision). Ceci signifie que bien que la méthodologie soit identique, des problèmes différents sont optimisés (voir (11) et (7)) et des paramètres différents utilisés (pour l'approche bayésienne, les paramètres viennent du posterior de la croyance sur la distribution alors que pour l'autre ils viennent directement du modèle marginal). Nous relevons que la descente de gradient stochastique pour la stratégie marginale est directement effectuée selon la pente la plus forte.

Les résultats des expériences sont que l'approche bayésienne s'avère en effet adaptée pour prendre en compte raisonnablement bien la fairness ainsi que l'incertitude des modèles. Ce second point est mis en valeur dans l'expérience sur l'ensemble de données synthétiques où il est observé que, dans plusieurs cas, la valeur de la performance de l'approche marginale diminiue au début (i.e., durant les première observations de données générées) puis atteint la même valeur que l'approche bayésienne après qu'une quantité suffisante de données aient été obtenues. Dans les expériences sur les deux ensembles de données mentionnés, l'approche bayésienne n'a pas été dominé par celle marginale. Quant au premier point, à nouveau dans l'expérience avec l'ensemble de données synthétiques, il a été relevé que l'approche bayésienne est de façon consistente meilleure que l'approche marginale vis-à-vis de la mesure

8. Discuté brièvement dans le chapitre 2, voir aussi Propublica analysis of COMPAS

de fairness, tout en ayant une performance à nouveau équivalente. Plus le trade-off λ est augmenté (e.g., de 0 à 1), plus l'approche bayésienne confirme sa superiorité rapidement quant au score de fairness ; cela sans une baisse de performance.

Une première remarque que nous pouvons émettre est à quel point, bien que [17] affirme ainsi permettre une comparaison équitable, le choix de commencer avec le même prior pour les deux approches (bayésienne et simple marginale) influence les résultats sachant que les paramètres sont samplés du posterior (mis à jour) dans la méthode bayésienne. Éventuellement, un ajustement du prior initial pourrait conduire à des différences de résultats intéressantes, bien que les comparaisons seraient alors plus délicates. De plus, connaissant un des problèmes fondamentaux des approches bayésiennes qu'est le manque de puissance de calculs pour des problèmes complexes, nous pouvons questionner l'efficacité d'une telle méthode pour un scénario avec un bien plus grand nombre de modèles que, par exemple, les 8 modèles possibles dans l'expérience de [17] sur l'ensemble de données synthétique ; la convergence du posterior au vrai modèle se fait dans ce cas bien plus rapidement et facilement. Finalement, bien que les résultats semblent prometteurs, nous pensons qu'il est encore nécessaire de conduire de nouvelles expériences similaires avec des scénarios moins simplifiés. [17] suggère à juste titre d'étendre ces recherches aux problèmes séquentiels dans un contexte bayésien, là où la prise en compte d'une future croyance par la stratégie de décision semble encore plus prône à la réussite d'une approche bayésienne.

4.2.4 *Fairness dans le NLP avec Word Embeddings*

Afin d'élargir notre vision d'application de la fairness ainsi que de cibler un domaine plus spécifique, nous présentons le contexte des biais dans le NLP ainsi en s'appuyant constamment sur et relevant principalement les éléments de recherche jugés pertinents de l'article [7] "*Man is to Programmer as Woman is to Homemaker ? Debiasing Word Embeddings*" qui vise à réduire les biais d'approches par Word Embeddings. Ceci permet ainsi de mieux comprendre le processus général de *debiasing* ("enlever les biais") ou mitigation de biais dans un cas plus concret.

[7] critiquent le manque de publications dénonçant l'existence prononcée des biais de genre, puis relèvent que des biais marqués ont été observés suite à l'utilisation de word embeddings sur des textes de *Google news*. L'inquiétude quant aux capacités d'amplification des frameworks tels que le word embeddings est alors justifiée au vu de leur forte présence dans les systèmes de NLP. [7] ont pour objectif d'obtenir des embeddings qui réduiront la présence ou l'impact des biais et, surtout, qui ne les amplifieront pas.

Un *word embedding* est une représentation d'un mot w en un vecteur $\mathbf{w} \in \mathbb{R}^d$, où $d \in \mathbb{N}$ est la dimension du vecteur. Deux mots w_1, w_2 similaires (e.g., même sens sémantique) devraient donner lieu à une équation de la forme $\mathbf{w_1} - \mathbf{w_2} \approx 0$. Plus généralement, la différence entre deux vecteurs issus de word embeddings représente certaines relations entre les mots[51]. Ces relations peuvent être quelconques et complexes à percevoir par humain, ainsi, un programme typique a pour but de retourner le ou les mots avec les relations les plus fortes. Le titre de l'article ici principale, "Man is to computer programmer as woman is to homemaker ?"[7], interroge en pointant un exemple de biais trouvé dans des algorithmes avec word embed-

dings. Étant donné l'analogie "$man : computerprogrammer :: woman : X$", le résultat premier retourné est "$X = homemaker$" qui est une conséquence de la relation trouvée : "$w_{man} - w_{woman} \approx w_{computerprogrammer} - w_{homemaker}$.

À titre d'exemples, nous mentionnons quelques résultats de [7] exhibant des biais, suivi de résultats considérés appropriés ou raisonnables (sans biais).

Pour les analogies de genre ("femme vs. homme" ou "elle vs. il") biaisées, en anglais [9] :

— nurse vs. surgeon
— interior designer vs. architect
— giggle vs. chuckle
— charming vs. affable
— vocalist vs. guitarist

Pour les analogies de genre jugées appropriées, en anglais [10] :

— queen vs. king
— sister vs. brother
— ovarian cancer vs. prostate cancer
— convent vs. monastery

La première attention est portée à l'aspect géométrique des word embeddings qui exhibent déjà des caractéristiques biaisées. Atteindre une fairness dans les relations entre mots pourrait sembler plus facile qu'il en est le cas. Par exemple, en considérant l'approche naïve d'observer les apparitions par paire (e.g., "female nurse" par rapport à "male nurse") et les classer numériquement (i.e., par leur nombre d'apparitions) ne permet pas d'obtenir des résultats corrects. Ceci peut être expliqué par une investigation plus profonde, au-delà de la fonction de l'algorithme. La présence supérieure du groupe de mots "male nurse" ou "female quarterback" peut être due à la perception des auteurs - dans le sens de "personne ayant écrit le texte d'où il est tiré" - quant à la supposée nécessité - de par le caractère inhabituel d'une telle association - de préciser que c'est un "nurse" male ou que c'est un "quarterback" female. En d'autres termes, ce qui est commun ou habituel ne sera probablement pas précisé ou détaillé, et le contraire pour ce qui semble plus rare ou inhabituel pour l'auteur.

Les mots en anglais ont différents genres : masculin (spécifique au male), féminin (spécifique à la femelle), neutre. Obtenir que "king" est plus similaire ou proche de "man" que de "woman" n'est ainsi pas étonnant et devrait être pris en compte lors de la confection d'algorithme servant à éviter ou corriger les biais. Par conséquent, [7] précise que leur algorithme de correction de biais prend en compte la différence entre les mots spécifiques à un genre (e.g., "businesswoman" et "businessman") et ceux neutres en ne corrigeant les biais que pour ces derniers. Ils notent aussi le besoin d'une attention particulière aux biais dits "indirects" qui sont une conséquence d'une relation entre un autre mot spécifique à un genre et le mot neutre. Bien que les genres sont fortement présents, il est important de garder à l'esprit que d'autres relations/associations jouent un rôle non-négligeable dans les résultats de similarité entre deux mots affichant des biais de genre (e.g., "businessman" proche de "finance"). C'est pourquoi

9. Bien qu'il soit probable que le même type de relations soient trouvées dans d'autres langues, nous ne traduisons pas les résultats puisqu'ils ont été obtenus suite à des analyses de textes anglais.
10. *Ibid.*

analyser dans le détail les résultats géométriques peut donner une idée de l'impact de l'association au genre par rapport aux autres relations existantes.

Dans l'approche de correction de biais, les objectifs que [7] se sont fixés sont les suivants : s'assurer que les mots neutres (e.g., "nurse") sont à une même distance entre les paires de genre (e.g., "she" et "he"), réduire les biais de genre associés indirectement aux mots neutres, préserver la performance du word embeddings ainsi que les relations appropriées [11] qui ne devraient pas être débiaisées). Pour cela, un aspect fondamental est la création d'un *gender subspace* qui va permettre d'effectuer les distinctions nécesssaires.

[7] pointent que, contrairement à l'étude de fairness et biais dans le cadre de problèmes de classification (voir sections 4.2.1 et 4.2.3), les word embeddings ne distinguent pas des individus particuliers ni un ensemble d'attributs sensibles. Ceci impose d'envisager la résolution et étude de fairness différemment.

[7] introduisent un embedding - avec $d = 300$ - particulier appelé *w2vNEWS* et qui est utilisé principalement dans l'article. Nous considérons alors un embedding unitaire (i.e., $\|\mathbf{w}\| = 1$) pour chaque mot $w \in W$. De plus, nous considérons l'ensemble $N \subset W$ de mots neutres ainsi que l'ensemble $P \subset W \times W$ de paires femme-homme (e.g., she-he, sister-brother) qui sont des ensembles possibles à obtenir par les embeddings. [7] discutent une méthodologie pour le faire à la section 7 de leur publication. La mesure de similarité utilisée entre deux mots w_1 et w_2 normalisés (i.e., de norme unitaire) est un simple cosinus ou produit scalaire [12] de façon à ce que nous ayons :

$$\mathbf{w_1} \cdot \mathbf{w_2} = cos(\mathbf{w_1}, \mathbf{w_2}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \tag{13}$$

Une fois que la mesure de similarité est obtenue, il est possible de l'évaluer par rapport à ce qui est considéré biaisé ou non. Par exemple, il est possible de demander à un groupe d'individus volontaires d'attribuer une valeur de 0-10 où 10 signifie "fort stéréotype" d'association entre "she" ou "he" et les mots en question. [7] ont conduit les expériences avec intervention humaine en demandant, à travers une plateforme de sondage, de partager des mots (e.g., compléter une analogie) ou d'évaluer le niveau de stéréotype de certaines paires de mots ou analogies. Une remarque très pertinente de [7] est que les stéréotypes discutés ne sont pas propre à Word2Vec et se retrouvent aussi dans d'autres frameworks d'embeddings tels que GloVe.

Plutôt que d'avoir à compléter une analogie comme présenté plus tôt (e.g., "$a : a' :: b : b'$" où a,a' et b sont donnés et b' est à trouver), il est possible de demander la tâche suivante. Compléter une analogie de la forme "$a : a' :: b : b'$" où cette fois la paire (a, b) est donnée et la paire (a', b') est à trouver. Ainsi, l'algorithme a pour but de générer des paires de mots pour lesquelles il trouve une relation forte avec la paire de mots données en entrée. Pour une

11. Un moyen de s'accorder sur ce qui est approprié dans les cas délicts est la collection de réponses d'un grand nombre d'individus variés

12. Bien qu'il existe différentes mesures ayant été proposées. Voir [49] pour un exemple de mesure plus poussée.

paire (a,b), $(\mathbf{a}-\mathbf{b})$ est appelée la *seed direction* (car (a,b) est la *seed*, input de l'algorithme). Ensuite, toutes les paires de mots (x,y) se voient attribuées un score selon la formule :

$$S_{(a,b)}(x,y) = \begin{cases} \cos(\mathbf{a}-\mathbf{b},\mathbf{x}-\mathbf{y}) & \text{, si } \|\mathbf{x}-\mathbf{y}\| \leq \delta \\ 0 & \text{, sinon} \end{cases} \tag{14}$$

, où δ est le seuil de similarité à définir (e.g., $\delta = 1$).

Il faut comprendre cette formule comme maximisant le score pour des paires de mots parallèles en plus de contraindre les mots de la paire (x,y) à être suffisamment proches. En pratique, [7] mentionnent qu'un seuil $\delta = 1$ et les embeddings normalisés revient à avoir les deux mots complétant l'analogie plus proches l'un de l'autre que le seraient deux embeddings aléatoires. L'algorithme utilisé par [7] retourne les paires avec le plus haut score et élimine les analogies supplémentaires partageant le même mot x pour des questions de redondance.

Pour les biais de genre indirects, [7] se sont aussi servis d'individus par sondage en plus d'analyses des résultats géométriques. Le type de mots choisi pour l'expérience est celui des "occupations" ou "activités". En prenant une paire de mots neutres (e.g., "softball" et "football"), les mots d'occupations ou activités (e.g., waitress, businessman, receptionist, etc) sont projetés selon la direction de la différence des mots de la paire choisie (e.g., selon $\mathbf{w}_{softball} - \mathbf{w}_{football}$) et se sont intéressés aux mots les plus proches d'un des mots de la paire choisie (i.e., les mots les plus bas sur une projection négative ou les plus hauts sur une projection positive). Il a été observé, par exemple, que "bookkeeper" et "receptionist" sont beaucoup plus proches de "softball" que de "football", ce qui indique potentiellement un biais de genre - c'est-à-dire que "bookkeeper", "receptionist" et "softball" seraient alors interprétés de façon biaisée vers le genre féminin ou "she").

Dans la section 5 de leur publication, [7] définissent un protocole d'expérience avec une mesure rigoureuse des effets de biais directs et indirects en définissant et identifiant d'abord un sous-espace de genre ainsi qu'une direction de genre $g \in \mathbb{R}^d$ (obtenue par *PCA* [13]) permettant de capturer le genre dans un embedding. Ils identifient ainsi les biais géométriquement et les quantifient selon une formule définie, toujours basée sur le produit scalaire.

Jusque là, les algorithmes utilisés forment une base pour la détection et correction de biais mais n'ont pas pour but direct leur correction. [7] introduisent alors des algorithmes de correction de biais. En premier lieu, il faut identifier le sous-espace ainsi que la direction de l'embedding qui capture les biais - cette étape est appelée "*Identify gender subspace*". La prochaine étape, finale, donne lieu à deux possibilités qui sont respectivement un *hard-debiasing* et *soft-debiasing* : "*Neutralize and equalize*" ou "*Soften*". Le première assure que les mots neutres sont sans importance (i.e., sont nuls) dans le sous-espace ("Neutralize") et que tous les mots neutres sont à une même distance de tous les mots de chaque ensemble (appelé "*equality set*") de mots en dehors du sous-espace ("Equalize"). Par exemple, en considérant les equality sets grandmother, grandfather et guy, gal, nous obtenons après Equalize que le mot "babysit" (qui est un mot neutre) est à une même distance de "grandmother" et "grandfather" ainsi que de "gal" et "guy", mais d'une distance probablement plus grande (car il est plus commun d'associer le mot babysit avec les grands-parents). Cette première possibilité est à envisager dans le cas

13. PCA : Principal Components Analysis

où nous souhaitons éviter les biais entre ces ensembles et les mots neutres. Cependant, un problème à considérer est que les mots de l'ensemble deviennent égaux et perdent donc de leurs caractéristiques. Un exemple concret est les textes contenant "to grandfather something" où le mot "grandfather" est alors utilisé différemment de ce pour quoi le mot "grandmother" pourrait être utilisé (nous n'écrivons pas "to grandmother something"). Suivant les cas et le modèles souhaité, cela peut être plus ou moins négligeable. La seconde possibilité, "Soften", a pour but de réduire les différences entre les ensembles de façon moins radicales, le but étant de raisonnablement préserver l'embedding original. Un paramètre de *trade-off* est à ajuster. La transformation pour Soften est une matrice (transformation linéaire) qui va minimiser la projection des mots neutres dans le sous-espace tout en préservant les produits scalaires entre les embeddings. Les formules à utiliser et l'équation à optimiser peuvent être trouvés dans la section 6 de [7].

Comme remarqué précédemment, afin de pouvoir utiliser les algorithmes vus jusqu'ici, il est nécessaire de pouvoir identifier les mots neutres parmi le vocabulaire disponible (issu de traitements de corpora de texte). Puisque les mots spécifiques à un genre sont bien moins nombreux, c'est en premier lieu à eux qu'il faut s'intéresser, puis simplement prendre le complémentaire de l'ensemble de ces mots. Mathématiquement, nous écrivons $N = W \setminus S$. L'ensemble S peut être obtenu à l'aide d'un classifieur linéaire (e.g., SVM [14]) à partir duquel il est possible d'obtenir un ensemble de mots spécifiques à un genre qui va être ajouté à une liste de base de mots (voir Appendice C de [7]) sous-ensemble de l'ensemble de données initialement considéré.

Avant de conclure sur l'article de recherches, nous présentons les résultats des expériences de [7] pour la correction de biais. Concernant les biais directs, la génération (ou complétion) d'analogie a été utilisée pour la paire "she-he" et les réponses des sondages (où il a été demandé aux participants d'évaluer le niveau de stéréotype des analogies créées par l'algorithme) adressés à un groupe d'humains ont été prises en compte, comme expliqué précédemment. Le soft-debiasing n'a pas offert de résultats probants pour les biais directs. Toutefois, [7] donnent l'exemple suivant de résultat de hard-debiasing. L'analogie à compléter "he to doctor is as she to X" a donné lieu au résultat $X = nurse$ par l'algorithme *avant* debiasing. Or, après hard-debiasing, la réponse a été $X = physician$ qui perd ainsi le fort biais de genre tout en restant cohérent. De plus, il est noté que l'algorithme a bien préservé les relations appropriées telles que "she to ovarian cancer is as he to prostate cancer", puisque le nombre d'analogies jugées appropriées avant et après hard-debiasing est resté le même sur les 150 analogies, bien que légèrement variable. Ceci témoigne d'une réussite de l'objectif de préservation de la qualité des embeddings. [7] pousse encore les tests expérimentaux (voir leur Appendice) en comparant la qualité des embeddings débiaisés obtenus à plusieurs benchmarks. La figure 2 de [7] affiche les résultats finaux de leur algorithme de correction de biais dans le cadre de l'expérience décrite. Pour ce qui est des biais indirects, [7] ont relevé la difficulté d'évaluer la performance de leurs embeddings débiaisés sans posséder de valeurs ground truth sur les effets indirects des biais de genre. Malgré cela, ils ont tout de même pu obtenir des résultats significatifs de débiaisement réussi tout en semblant préserver la qualité des embeddings. Pour l'exemple des mots selon la direction $\mathbf{w}_{softball} - \mathbf{w}_{football}$ qui avait donné lieu, après "pitcher", à "bookkeeper" ou "receptionist" pour "softball", les résultats "infielder" et "major leaguer" (toujours après "pitcher") ont été obtenus après correction des biais de genre sans

14. SVM : Support Vector Machine

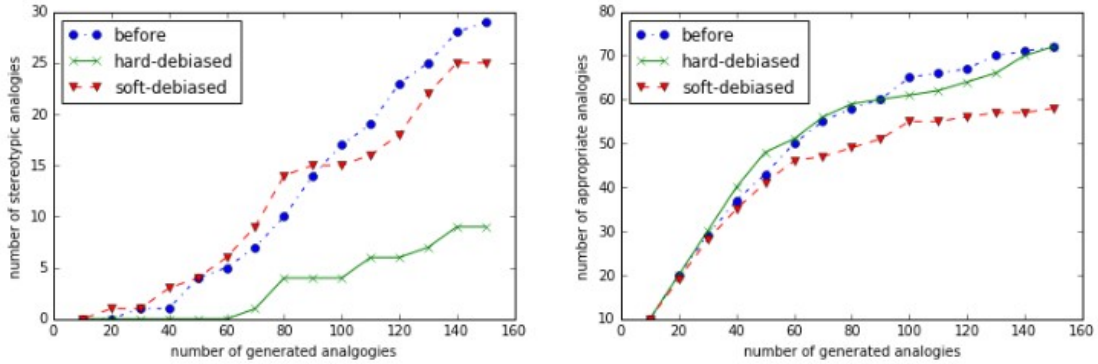

FIGURE 2 – À gauche, le nombre d'analogies considérées stéréotypées par rapport au nombre d'analogies générées. À droite, le nombre d'analogies jugées appropriées par rapport au nombre d'analogies générées. Les graphes affichent les résultats avant la correction de biais (en bleu), après un hard-debiasing (en vert) et après un soft-debiasing (en rouge). Note : les décisions ont été prises selon une majorité sur 10 participants par analogie. [7]

entraver ceux de "football". Ceci est un constat positif du fonctionnement des algorithmes de débiaisement des embeddings. La figure 3 de [7] montre quelques exemples de résultats des associations avant et après correction de biais indirects.

softball **extreme**	**gender portion**	**after debiasing**
1. pitcher	-1%	1. pitcher
2. bookkeeper	20%	2. infielder
3. receptionist	67%	3. major leaguer
4. registered nurse	29%	4. bookkeeper
5. waitress	35%	5. investigator
football **extreme**	**gender portion**	**after debiasing**
1. footballer	2%	1. footballer
2. businessman	31%	2. cleric
3. pundit	10%	3. vice chancellor
4. maestro	42%	4. lecturer
5. cleric	2%	5. midfielder

FIGURE 3 – Résultats des associations de l'algorithme concernant les occupations/activités avant et après correction des biais indirects de genre. Les pourcentages indiquent avec quelle proportion une association représente un biais de genre. [7]

Pour terminer, nous résumons ce qui a été vu dans cette publication et ce que nous pouvons en tirer. Bien que l'approche de NLP par word embeddings s'avère dans certains cas ambiguë dans ses implications, elle dispose d'atouts pratiques, efficients, voire même élégants. En plus de la capacité des word embeddings à capturer l'information attendue quant aux problèmes à résoudre, [18] remarquent que les word embeddings contiennent également de l'information utile pour la correction de biais. Nous avons vu à travers cet article une

approche, fidèle au word embeddings conventionnel, mais qui a su incorporer des éléments apportant une amélioration conséquente dans la correction de biais tout en maintenant la qualité des embeddings initiaux. La définition d'un sous-espace ainsi que d'une direction, issue d'un ensemble de paires de mots, capturant le genre, le long de laquelle il est possible de projeter les mots neutres afin d'évaluer les biais indirects, a été un des pilliers importants du développement de la méthode de correction de biais de genre. Le hard-debiasing a montré des résultats convaincants tandis que le soft-debiasing devrait, et mérite, encore d'être testé dans des scénarios plus spécifiques. Les algorithmes étudiés ont pu préserver leur faculté à résoudre de façon adéquare la plupart des analogies ainsi que les capacités des embeddings à expliciter les groupements de mots aux caractéristiques similaires tout en accomplissant une partie de leur tâche de correction des biais.

Parmi les aspects qui mériteraient d'être étudiés plus profondément afin d'observer les conséquences sur les méthodes proposées jusque là, nous notons la mesure de similarité ainsi que le contrôle général de l'impact des modifications des embeddings qui semble manquer de preuves d'assurance. En particulier pour les biais indirects qui sont plus complexes à contrôler que ceux directs. Cet article ne pointe là qu'un type de biais - celui de genre - qui, bien que fortement présent, ne permet de corriger tous les autres types de biais présents. Néanmoins, cela pave le chemin à diverses possibilités inspirées des méthodes vues qui englobent plusieurs aspects essentiels dans l'approche de correction de biais. Nous faisons aussi remarquer que les recherches discutées ont été menées dans le contexte particulier de l'Anglais. Au vu des approches prometteuses, ou en tout cas considérées, il serait intéressant de conduire une étude similaire sur des langages dont la grammaire est fondamentalement différente (e.g., le français par rapport à la présence marquée de genres, ou encore le russe).

Une autre publication qui est une revue de littérature et rejoint les idées de [7] est "Mitigating Gender Bias in Natural Language Processing : Literature Review"[55]. De plus, il existe aussi plusieurs frameworks pour le NLP auxquels il pourrait être pertinent de s'intéresser, par exemple [53].

Finalement, le Word Embeddings n'est là qu'un outil qui souffre lourdement des biais déjà inclus dans la société et transcrits sans contrôle particulier. Pour cette raison, sans qu'un travail ne soit effectué au-delà du Word Embeddings, sa tâche risquerait - et c'est déjà quelque chose - de ne rester qu'à l'évitement d'amplification des biais présents. Ceci est à noter tout en gardant en tête que lorsque certains cas semblent être biaisés, il se pourrait que ce ne soit là qu'une réalité difficile à se représenter ou accepter.

5

RÉSULTATS

5.1 FAIRNESS RÉSOLUE ? QUELLES SOLUTIONS À ENVISAGER ?

Il est évident que la réponse au titre de cette section, en un sens provocateur, est claire. Toutefois, elle appelle au développement et à la mise en valeur des progrès et avancées qui ont été observés jusque là. Un premier constat est la hausse croissante des publications sur le sujet ainsi que de l'intérêt suscité auprès des chercheurs. La prise de conscience et de de responsabilité des chercheurs et sociétés directement concernées est un premier pas raisonnable vers une attention commune et générale des enjeux considérés.

Nous avons constaté que le framework relativement simple décomposant l'étude de biais en trois catégories "pré-traitement ou *pre-processing*, traitement ou *in-processing*, post-traitement ou *post-processing*" est celui généralement utilisé. Pour chacune de ces catégories, les phases de détection puis correction ou mitigation des biais peuvent être entreprises. Conscients du besoin de marge de manoeuvre dans l'utilisation de framework, nous maintenons qu'il s'avérerait utile d'en définir pour des problématiques plus précises. En tout cas, maintenir une fairness adéquate requiert à ce jour un sacrifice de performance qu'il est difficile à balancer, comme nous le retrouvons par exemple dans [23], mais loin d'être impossible (voir le développement de la section 4.2.4). Un but serait alors de requérir le ou les critères de fairness de faire partie intégrante de l'évaluation de la performance, plutôt que de toujours faire face au *trade-off* performance-fairness.

Nous notons que parmi les méthodes d'amélioration de fairness, aucune usant en partie de la création de variables supplémentaires afin de "diluer" l'importance des variables sensibles - plutôt que de les retirer comme la méthode peu efficace de *unawareness* - n'a été mentionnée. Les définitions de critères de fairness ont toutes présentées certaines qualités et certains défauts, avec une forte dépendance au type de données utilisé par l'algorithme pour les mesures numériques de fairness. Par exemple, l'apparente incompatibilité entre les critères basés sur les groupes et ceux basés sur les individus a brièvement été mentionnée. Un malheureux constat est donc celui de la difficulté à "mixer" ou conjuguer les diverses définitions de fairness, souvent contradictoires. Ceci a pour but de pallier au manque d'une définition absolue admissible. Toutefois, nous avons aussi conclu sur les difficultés mathématiques qui peuvent survenir, comme le relèvent [11] et [14]. Il s'est avéré, parmi les sources étudiées, que des critères plus sophistiqués ont donné lieu à de meilleurs résultats. Parmi les critères de fairness discutés et les méthodologies proposées, l'approche bayésienne, en particulier pour les problèmes de classification où l'aspect de discrimination apparaît naturellement, a montré de probants résultats (voir 4.2.3). La capacité à considérer de façon appropriée

plusieurs modèles pouvant être mis à jour contribue de façon sûre à la création d'une meilleure structure permettant de prendre en compte une fairness. Ce qui a été relevé de l'approche bayésienne laisse penser que c'est le type d'approche principal méritant d'être investigué. Aussi, suite à la discussion des tests dans la section 2.3, il serait envisageable de les faire passer aux modèles de ML en les adaptant à l'approche bayésienne. Bien sûr, nous n'avons pas du tout discuté ici des inférences causales et nous rappelons les critiques émises dans 4.2.3 quant à la puissance de calcul requise pour des problèmes hautement complexes. Nous insistons cependant sur la spécificité des conditions des expériences et la nécessité de les conduire sur d'autres ensembles de données de façon plus poussée, comme mentionné dans 4.2.3. Il est naturellement délicat d'apporter un jugement sur des expériences de ce type, pour lesquelles des itérations sur d'autres données avec diverses contraintes seraient requises avant de témoigner d'une efficacité définitive. Cependant, nous sommes conscients qu'il faut plutôt voir là des pistes et résultats prometteurs.

Le défi des problèmes de word embeddings en NLP qui consistent à corriger les biais, ou du moins, éviter leur amplification tout en maintenant des embeddings de qualité a été observé et discuté dans 4.2.4. En général, les tentatives de correction de biais sont effectuées durant la phase de pré-traitement. Un autre constat a été celui de la précaution avec laquelle il faut analyser les biais à travers les analogies. L'intérêt dans 4.2.4 a été porté vers les biais de genre qui sont importants et très communs, mais potentiellement plus faciles à détecter et corriger que les autres types de biais (e.g., lieu d'habitation, âge, éthnicité). Les résultats présentés sont eux aussi annonceurs d'un chemin à progrès. L'utilisation de sous-espaces et de formules définies autour d'une direction bien choisie issue de paires de mots a été un choix essentiel à la réussite de la méthode présentée. Il serait alors intéressant d'essayer de l'appliquer pour différents types de biais. Bien entendu, des modifications devront être apportées puisque la séparation en mots spécifiques à un genre et en mots neutres n'aura plus d'intérêt.

L'existence et importance des variables sensibles et de leurs proxies (i.e., variables liées/corrélées aux variables sensibles et donc devant également être surveillées) a été soulignée. Le danger des ajustements en phase de post-traitement (i.e., des résultats) a aussi été noté étant donné que cela peut conduire à aggraver les biais dans d'autres configurations pour des modifications complexes à estimer. De plus, cela reviendrait à éviter de comprendre la source d'inéquité. Toutefois, ce procédé possède les avantages de ne pas avoir besoin d'accéder au modèle ayant produit les résultats - car en effet, il n'est pas toujours possible d'étudier les détails du modèle - ni d'accéder à des informations privées (i.e., données sensibles). L'importance de travailler avec des données réalistes, sûres et suffisantes a été marquée par les différences de résultats entre un modèle entraîné sur un certain type de données (i.e., certaines attributs avec une certaine distribution) et le déploiement de ce même modèle faisant face à de nouvelles données, potentiellement très différentes. Ceci met en évidence le besoin d'une supervision continue d'un modèle déployé et des résultats produits. Bien que délicates dans le cadre de conflits d'intérêt, compétitivité et *privacy*, une collaboration et des discussions entre les personnes directement concernées (e.g., chercheurs, responsable de production, etc) semblent être très bénéfiques à un meilleur contrôle de la fairness.

Bien que nous ne nions pas le fait que les efforts effectués visant à une amélioration de la fairness manquent de structure et d'entente communautaire, il existe des initiatives fortement positives que [10] listent et qui sont des projets participatifs d'une ampleur notable

tels que ”AIF360”[4] ou encore ”ML Fairness Gym” [1] qui fait partie du projet Google Open AI.

L'analyse et régulation de la fairness requièrent un ensemble de compétences et un savoir allant au-delà des Sciences Informatiques. Ceci limite fortement l'accessibilité à l'amélioration de l'état de la fairness sans consensus inter-disciplinaire et frameworks établis. Nous avons compris que des régulations légales telles que celles de la RGPD ne seront jamais suffisantes pour résoudre le problème de fairness, si ce n'est fournir un cadre général et restreindre l'utilisation de certains types de données. Ce cadre est d'autant plus justifié qu'un trop grand nombre de systèmes d'IA sont encore mal sécurisés. En somme, une partie du contrôle des biais et de la fairness dans les algorithmes quitte le domaine apparent des Sciences Informatiques et relève d'un travail en commun nécessitant une excellente communication et des efforts constants, là où l'objectif pourrait encore sembler hors d'atteinte. Une attention particulière bien que brève a aussi été portée sur les dangers de vouloir forcer une correction de biais là où il n'y en aurait en fait pas. Par exemple, appliquer naïvement et de manière forcée le critère de demographic parity (voir section 4.2.1.1) dans le cadre d'un système juridique de prédiction de récidive où les hommes ont un taux de récidive plus élevé que celui des femmes pourrait résulter en des temps d'emprisonnement plus longs pour les femmes sans que ce ne soit justifié (car selon les statistiques à disposition les femmes auraient en effet moins de chance de récidiver) ; une méthode par le critère de demographic parity aurait alors augmenté les biais au lieu d'améliorer la fairness.

Un autre constat est que l'idée d'atteindre une fairness absolue semble disparaître, bien que c'est ce à quoi aspire grand nombre de chercheurs. Il devrait en fait être considéré d'office ”une fairness” plutôt que ”la fairness” ou ”l'équité”, signifiant qu'il faudrait alors en général discuter ”des” fairness, au pluriel. Il va ainsi s'agir de proposer des solutions adaptées en fonction de la situation. En effet, en plus des problèmes de transcription des idées de mitigation de biais de la société à la machine, [47] ont relevé que les différences démographiques ou de genre influencent de façon conséquente l'évaluation de la fairness par un humain. Par exemple, [47] concluent que la conception de fairness d'un individu pour une problématique donnée peut changer. Ceci pose alors un obstacle de plus quant à l'évaluation de la fairness par un programme. Ainsi, mettre en place des frameworks adaptifs dans lesquels des mesures appropriées de fairness pourraient être incorporées en fonction de la problématique est un but vers lequel tendre. Ceci implique la mise à disposition d'un document de référence à la complétion duquel pourraient contribuer les chercheurs. Ceci devrait naturellement résulter du constat de l'état actuel - et direction - de la fairness en IA. Dans un monde où les domaines tels que le Machine Learning se démocratisent de plus en plus et prennent de l'importance dans de nombreux secteurs, même a priori éloignés, nous avons besoin d'un consensus général évolutif, de documents officiels raisonnablement aboutis, accessibles publiquement, et d'outils facilitant le travail d'évaluation et correction de la, ou plutôt ”des”, fairness dans les algorithmes ou machines.

Remarque. Le chapitre 4, produit par un choix méticuleux et appuyé des informations à relever et détailler, peut être considéré comme fournissant des observations ou résultats appartenant à ce chapitre 5.

1. ML Fairness Gym

6

CONCLUSION

Après avoir défini le contexte de recherche et motivé l'importance des enjeux liés au sujet d'études dans le chapitre 1, une partie préliminaire a servi de fondation aux éléments par la suite développés ou discutés. En particulier, le développement du bayésiannisme - auquel le fréquentisme, approche populaire mais avec d'importants défauts, a été opposé -, de sa philosophie et de ses fondements dans le cadre de l'évaluation par p-value et des tests randomisés et contrôlés en double aveugle est une partie intéressante qui instaure une nouvelle idée de perspective dans l'étude de la fairness. La section Fairness avec approche Bayésienne du Chapitre 4 a relevé les recherches d'une publication proposant un framework se basant sur une approche bayésienne et produisant des résultats intéressants qui poussent à investiguer ce chemin avec attention. Le procédé de développement de la méthode ainsi que les limitations rencontrées témoignent aussi des aspects sur lesquels il serait utile de se pencher. La section Fairness dans le NLP avec Word Embeddings du Chapitre 4 a montré à travers une publication les difficultés intrinsèques auxquelles font face les chercheurs dans le NLP avec embeddings. En particulier, la complexité que peut représenter la correction de biais sans entraver la qualité des embeddings initiaux, l'attention particulière qu'il faut porter au potentiel d'amplification involontaire des biais, ainsi que le choix d'approches pour la détection et correction des biais (e.g., analyser à travers des analogies pourraient céler ou, au contraire, sembler faire apparaître des biais qui n'en sont pas). La difficulté d'appliquer une méthode à un autre type de biais que ceux de genre - qui sont ceux qui ont été étudiés dans la section mentionnée - est aussi un problème auxquels les méthodes se basant sur des caractéristiques (e.g., la distinction entre les mots spécifiques à un genre et les mots neutres) propres au type de biais font face. Les premières section du Chapitre 4 décrivent les difficultés rencontrées avec les approches et critères de fairness qui sont ou ont été utilisés. Cela prépare le terrain au deux sections principales du chapitre de développement mentionnées plus haut, en particulier en motivant le besoin de meilleures approches en matière de détection et correction des biais. En somme, le Chapitre 4 conduit une étude de l'état des recherches récentes et discute l'état de l'art à travers des articles jugés pertinents. Le Chapitre 3 propose un élargissement de la vue d'ensemble de l'état de la fairness et les concepts qui lui sont reliés (e.g., transparency, accountability) en proposant divers articles et ouvrages méritant d'être consultés. La forte abondance de publications nouvelles dans la littérature d'IA témoigne de l'importance de progresser dans la bonne direction quant aux biais et à l'état de la fairness. Les résultats partagés au Chapitre 5 représentent les conclusions des recherches menées quant à "l'état de la fairness dans les algorithmes d'IA et des biais introduits".

Nous concluons que la fairness est fortement liée au contexte et à la problématique dans laquelle elle est étudiée ou discutée. Afin de pouvoir faire valoir une fairness satisfaisante, il

ne faut pas être trop attaché à une optimisation de la performance face au trade-off inévitable de performance-fairness ; en tout cas pas dans un premier temps. Comme discuté au Chapitre 5, il faudrait incorporer une notion ou critère de fairness directement dans la mesure de performance afin de ne plus être confronté explicitement au *trade-off* performance-fairness qui ne devrait idéalement plus être considéré comme un problème, mais plutôt un but de performance - au sens large - naturel. En se concentrant sur l'amélioration de la qualité des données, en fournissant des moyens de collection de données appropriées, il faudrait se défaire de l'utilisation des attributs sensibles puis se concentrer sur l'impact des attributs corrélés aux attributs sensibles. Toutefois, nous nuançons en rappelant que pour certains domaines tels que le système de santé ou encore le système juridique et pénal, la performance ou précision est extrêmement importante et la balance performance-fairness se trouve être plus complexe à décider ; des efforts particuliers sont alors à fournir. Ainsi, dans ce type de cas, il est suggéré de travailler sur les données d'apprentissage (e.g., obtenir plus de données et de meilleures qualité) plutôt que d'essayer de corriger les biais du modèle lui-même. Nous notons néanmoins qu'il faut faire attention lors de cette phase de pré-traitement qui est sujette à l'introduction de biais humains (i.e., causée par les manipulations d'un humain).

En ce qui concerne les perspectives futures, il serait intéressant de travailler sur des problèmes de décisions séquentielles qui n'ont pas spécifiquement été adressés mais pour lesquels un framework suivant une approche bayésienne pourrait produire d'excellents résultats. Aussi, la littérature est majoritairement constituée de problèmes de classification simplifiés. Les publications les traitant mentionnent généralement qu'il est possible d'étendre leur approche à d'autres types de problèmes. En comprenant bien que les problèmes de classification, par exemple binaires, mettent facilement en valeur l'aspect de discrimination et donc éventuellement de biais, nous pensons tout de même qu'une recherche orientée dans le sens de problèmes plus complexes, ou en tout cas différents de par leurs applications, serait plus bénéfique. Une autre piste, plus générale, serait l'approfondissement des liens entre la fairness, et la transparence et interprétabilité afin de possiblement faciliter l'étude de la fairness et des biais.

En remarque finale, le présent travail n'indique pas qu'il est nécessaire d'exiger une résolution immédiate et absolue de la fairness. Premièrement car ce n'est pas un but forcément atteignable, et deuxièmement car les progrès et technologies vont constamment évoluer et maintenir une fairness - si ce n'est "la" - se présente comme un travail continu dont un des buts serait plutôt de faciliter son adaptation afin de toujours pouvoir proposer des solutions satisfaisantes dans le contexte actuel, pour une problématique particulière.

BIBLIOGRAPHIE

[1] M. T. Barendse, C. J. Albers, F. J. Oort, and M. E. Timmerman. Measurement bias detection through bayesian factor analysis. *Frontiers in Psychology*, 5 :1087, 2014.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[3] Elias Baumann and Josef Lorenz Rumberger. State of the art in fair ml : From moral philosophy and legislation to fair classifiers, 2018.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360 : An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[5] Jelke Bethlehem. Selection bias in web surveys. *International Statistical Review / Revue Internationale de Statistique*, 78(2) :161–188, 2010.

[6] D. Biau, B. Jolles, and R. Porcher. P value and the theory of hypothesis testing : An explanation for new researchers. *Clinical Orthopaedics and Related Research®*, 468 :885–892, 2010.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker ? debiasing word embeddings, 2016.

[8] William M. Briggs. It is time to stop teaching frequentism to non-statisticians, 2012.

[9] Jianping Cao, Ke Zeng, Hui Wang, Jiajun Cheng, Fengcai Qiao, Ding Wen, and Yanqing Gao. Web-based traffic sentiment analysis : Methods and applications. *IEEE Transactions on Intelligent Transportation Systems*, 15(2) :844–853, 2014.

[10] Simon Caton and Christian Haas. Fairness in machine learning : A survey, 2020.

[11] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. How to be fair and diverse ?, 2016.

[12] Silvia Chiappa and William S. Isaac. A causal bayesian networks viewpoint on fairness. *IFIP Advances in Information and Communication Technology*, page 3–20, 2019.

[13] Alexandra Chouldechova. Fair prediction with disparate impact : A study of bias in recidivism prediction instruments, 2016.

[14] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness : A critical review of fair machine learning, 2018.

[15] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA, 2016. Association for Computing Machinery.

[16] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210 :2–21, 2018. Randomized Controlled Trials and Evidence-based Policy : A Multidisciplinary Dialogue.

[17] Christos Dimitrakakis, Yang Liu, David Parkes, and Goran Radanovic. Bayesian fairness, 2018.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011.

[19] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning, 2018.

[20] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3) :330–347, July 1996.

[21] Hila Gonen and Yoav Goldberg. Lipstick on a pig : Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, 2019.

[22] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3) :50–57, Oct 2017.

[23] Christian Haas. The price of fairness - a framework to explore trade-offs in algorithmic fairness. 12 2019.

[24] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

[25] Lê Nguyên HOANG. *La formule du savoir : Une philosophie unifiée du savoir fondée sur le théorème de Bayes*. EDP Sciences, 2018.

[26] Simon Hug. Selection bias in comparative research : The case of incomplete data sets. *Political Analysis*, 11(3) :255–274, 2003.

[27] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan 2019.

[28] John Ioannidis. Why most published research findings are false. *PLoS medicine*, 2 :e124, 09 2005.

[29] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior, 2020.

[30] Akash Junnarkar, Siddhant Adhikari, Jainam Fagania, Priya Chimurkar, and Deepak Karia. E-mail spam classification via machine learning and natural language processing. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pages 693–699, 2021.

[31] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. Preference amplification in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 805–815, New York, NY, USA, 2021. Association for Computing Machinery.

[32] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016.

[33] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

[34] Michael Lavine. Frequentist, bayes, or other ? *The American Statistician*, 73(sup1) :312–318, 2019.

[35] Mark Ledwich and Anna Zaitsev. Algorithmic extremism : Examining youtube's rabbit hole of radicalization, 2019.

[36] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai : A review of machine learning interpretability methods. *Entropy*, 23(1), 2021.

[37] El Mahdi EL MHAMDI Lê Nguyên HOANG. *Le fabuleux chantier : rendre l'intelligence artificielle Robustement Bénéfique*. EDP Sciences, Place of publication not identified, 2019.

[38] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Feedback loop and bias amplification in recommender systems. *CoRR*, abs/2007.13019, 2020.

[39] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name : Mitigating gender bias with name-based counterfactual data substitution. *CoRR*, abs/1909.00871, 2019.

[40] Douglas S. McNair. Preventing disparities : Bayesian and frequentist methods for assessing fairness in machine-learning decision-support models, new insights into bayesian inference. 2018.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[42] I. J. Myung and M. Pitt. Applying occam's razor in modeling cognition : A bayesian approach. *Psychonomic Bulletin & Review*, 4 :79–95, 1997.

[43] Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 83–84, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[44] Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational : Man is to doctor as woman is to doctor. *CoRR*, abs/1905.09866, 2019.

[45] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[46] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[47] Emma Pierson. Demographics and discussion influence views on algorithmic fairness, 2018.

[48] Martin L Puterman. *Markov decision processes : discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[49] Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. Measuring semantic similarity of words using concept networks. pages 193–200, 08 2016.

[50] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. *CoRR*, abs/1908.08313, 2019.

[51] Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10) :627–633, October 1965.

[52] Terrence J. Sejnowski. *The Deep Learning Revolution*. The MIT Press, 2018.

[53] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models : A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics.

[54] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing : Literature review, 2019.

[55] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing : Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics.

[56] Zhixing Tan, Shuo Wang, Zonghan Yang, Gang Chen, Xuancheng Huang, Maosong Sun, and Yang Liu. Neural machine translation : A review of methods, resources, and tools, 2020.

[57] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training, 2020.

[58] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. The computational limits of deep learning, 2020.

[59] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai) : Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–21, 2020.

[60] Stephen Vardeman and Glen Meeden. Calibration, sufficiency, and domination considerations for bayesian probability assessors. *Journal of the American Statistical Association*, 78(384) :808–816, 1983.

[61] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[62] Ronald Wasserstein, Allen Schirm, and Nicole Lazar. Moving to a world beyond "p ¡ 0.05". *American Statistician*, 73 :1–19, 03 2019.

[63] Whittaker, J. & Looney, S. & Reed, and A. & Votta. Recommender systems and the amplification of extremist content. internet policy review, 10(2), 2021.

[64] Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 285–294, New York, NY, USA, 2019. Association for Computing Machinery.

[65] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017.

[66] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[67] Panagiotis Zervopoulos, Ali Emrouznejad, and Sokratis Sklavos. A bayesian approach for correcting bias of data envelopment analysis estimators, 02 2019.