UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

THÈSE DE DOCTORAT

L'équité dans les machines sociales

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN SCIENCES ET TECHNOLOGIES DE L'INFORMATION

PAR:

Mir Saeed Damadi

décembre 2025

Résumé

L'équité dans les machines sociales est une préoccupation essentielle alors que les systèmes automatisés et les interactions numériques deviennent de plus en plus omniprésents dans notre vie quotidienne. Cette recherche doctorale examine les enjeux d'équité dans les machines sociales et propose des critères pour mesurer l'équité. L'objectif de cette recherche est de créer un cadre global pour comprendre et aborder l'équité dans les machines sociales.

En passant en revue systématiquement la littérature existante et en réalisant des études de cas détaillées, cette recherche catégorise les phénomènes de biais dans différents types de machines sociales, identifie les préjudices causés par ces biais, et explore leurs relations causales.

À notre connaissance, cette recherche est la première à intégrer des métriques d'équité issues de l'apprentissage automatique dans les machines sociales, en introduisant des critères pour mesurer les biais dans ces systèmes. Nous cherchons à déterminer si la représentation dans les machines sociales est une réflexion fidèle de notre société biaisée ou si ces systèmes sociotechniques génèrent leurs propres biais, similaires aux biais algorithmiques. Enfin, nous présentons une solution pour réduire le biais dans l'étude de cas de Wikipédia. Ces résultats contribuent au domaine plus large de l'équité algorithmique et fournissent des perspectives et des stratégies pour créer des technologies et des interactions plus justes dans les machines sociales.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Motivation

L'équité est devenue un sujet important dans le cadre de FAccT (Fairness, Accountability, and Transparency) et de l'IA responsable, en mettant l'accent sur les résultats potentiellement injustes des décisions automatisées alimentées par l'apprentissage automatique (Machine Learning; ML) [25, 26]. Des recherches ont été menées sur l'équité en apprentissage automatique dans divers domaines, notamment la justice [5], les soins de santé [31, 68] et l'éducation [89, 90]. De nombreuses métriques d'équité [13, 78, 25] et mesures correctives [18, 11, 135] ont été proposées, avec la majorité des analyses causales et des mesures correctives spécifiquement adaptées à l'IA, et le plus souvent aux tâches de classification en apprentissage automatique.

Par exemple, le système COMPAS (Correctional Offender Management

Profiling for Alternative Sanctions), un outil de gestion de cas et de support à la décision utilisé dans le système de justice pénale aux États-Unis, a suscité des préoccupations importantes en matière d'équité. Selon ProPublica, les taux d'erreur étaient différents pour les personnes blanches et les personnes noires [6]. Dans le cas de COMPAS, différents critères d'équité tels que les Equalized Odds [59], qui exigent que les prédictions du système soient également précises pour tous les groupes démographiques, et la Predictive Parity [49], qui exige que les individus classés à haut risque aient la même probabilité de récidive, quel que soit leur groupe d'appartenance, ont été appliqués pour évaluer COMPAS. Ces critères visent à garantir que le système ne nuit pas de manière disproportionnée à certains groupes et que ses prédictions sont équitables et justes pour des populations diverses [26, 67].

Il y a également eu des *audits algorithmiques* de systèmes qui ne sont pas clairement basés sur l'apprentissage automatique. Par exemple, TaskRabbit et Fiverr, qui sont des places de marché en ligne de la soi-disant "gig economy", ont été auditées pour des questions d'équité [57]. Les chercheurs ont découvert que le genre et la race perçus étaient significativement corrélés avec les évaluations des travailleurs et les classements dans les recherches. Plus précisément, les travailleurs perçus comme étant noirs recevaient de moins bonnes évaluations que les travailleurs perçus comme étant blancs, bien qu'ils soient tout aussi qualifiés. Dans cette étude, contrairement au cas COMPAS, le biais a été mesuré avec des métriques ad hoc, sans analyse approfondie des causes ou des mesures correctives possibles. Nous supposons que cette ab-

sence d'analyse plus poussée pourrait être due au fait que les causes des biais sont présumées être des causes sociales, telles que le racisme, qui doivent être traitées au niveau social. D'autres systèmes sociotechniques (SST) similaires audités sont recensés dans la revue systématique de la littérature de Bandy [8], qui souligne l'importance croissante et la banalisation des audits algorithmiques. Dans la plupart des travaux examinés, comme dans les audits de TaskRabbit et Fiverr, il n'y a aucune référence explicite à des métriques d'équité spécifiques telles que les Equalized Odds ou la Predictive Parity.

Contrairement à COMPAS, où un algorithme spécifique est examiné, les systèmes sociotechniques audités impliquent un mélange de processus algorithmiques et d'interventions humaines, ce qui rend difficile de définir clairement la frontière entre l'algorithme et la composante sociale. Par exemple, l'algorithme de classement sur des plateformes telles que TaskRabbit et Fiverr agrège les votes des utilisateurs (la question de savoir si le vote des utilisateurs doit être un critère de classement fait encore débat), intégrant à la fois l'aspect social (les votes des utilisateurs) et l'aspect computationnel (les calculs algorithmiques). Cette intégration brouille la ligne entre les composantes sociales et algorithmiques.

Il y a également eu des analyses liées à l'équité des systèmes sociotechniques (SST) qui ne pourraient pas être décrites comme des audits *algorithmiques* car aucun algorithme identifiable n'est en cause. Par exemple, sur la version anglaise de Wikipédia, le nombre de biographies d'hommes est environ quatre fois supérieur à celui des femmes [121]. Dans le processus

d'acceptation des biographies sur Wikipédia, bien qu'il existe des règles de notabilité et que certains bots puissent être utilisés pour vérifier la crédibilité des biographies, il n'y a pas d'algorithme informatique spécifique régissant l'acceptation des biographies. Au lieu de cela, le contenu de Wikipédia est produit par une *machine sociale* (MS) [105], c'est-à-dire un réseau de personnes interagissant sur une infrastructure numérique, comprenant divers composants automatisés ou semi-automatisés. Ces systèmes forment une sous-catégorie distincte des systèmes sociotechniques, qui ont attiré une attention renouvelée ces dernières années [92, 106]. Ici encore, les biais sont évalués à l'aide de métriques ad hoc, et il est difficile de savoir comment ces biais peuvent être atténués : certaines solutions d'ordre social ont été proposées, comme l'organisation de "edit-a-thons" orientés vers les femmes pour aborder les biais de genre de Wikipédia [136, 37]. Cependant, il est nécessaire de mener des analyses causales plus systématiques et de développer des solutions applicables à ce type de SST.

1.2 Définition des questions de recherche

Dans cette recherche, nous étudions comment les concepts d'équité peuvent être examinés dans les machines sociales, y compris (mais sans s'y limiter) les composants algorithmiques. Les machines sociales sont des systèmes complexes; en raison de leur potentiel à grande échelle, impliquant parfois des centaines de millions de personnes interagissant, et en raison des com-

posants algorithmiques qui médiatisent certaines de ces interactions. En d'autres termes, pour citer Simon et al. [111], nous visons à élargir l'étude de l'équité algorithmique au "système socio-technique plus large dans lequel les technologies sont situées". La question se pose alors : les propriétés d'équité de ces systèmes sont-elles une préoccupation pertinente dans le domaine de l'informatique ? Dans ces situations où aucun système de ML ne produit directement de biais, les analyses causales spécifiques au ML et les solutions spécifiques au ML ne sont pas applicables ; il y a alors un manque de compréhension des causes possibles et des solutions. Cela a été un angle mort de la recherche sur l'équité algorithmique (délibérément laissé de côté parce qu'il n'y a pas d'"algorithm" à blâmer et à corriger).

Nous émettons l'hypothèse que la conception de l'infrastructure technique d'une machine sociale joue un rôle dans ces biais (même s'il est difficile de clairement identifier un "algorithme" impliqué) et peut être modifiée pour atténuer les biais.

Pour valider cette hypothèse, nous devons :

- mieux comprendre les questions liées à l'équité dans les machines sociales, qu'il y ait ou non des causes algorithmiques clairement identifiables [30, 29].

- être capables de comprendre le rôle que joue la machine sociale elle-même dans la formation des biais.

- comprendre comment la conception de la machine sociale peut être

modifiée pour atténuer les biais (cela inclut les processus ainsi que les composants pleinement “techniques”).

À la lumière des objectifs décrits ci-dessus, notre recherche vise à approfondir notre compréhension des questions liées à l'équité au sein des machines sociales, tant dans les cas où les causes algorithmiques sont clairement identifiables que dans ceux où elles ne le sont pas. De plus, nous cherchons à découvrir le rôle que la machine sociale elle-même—comprenant à la fois des composants humains et numériques—joue dans la formation et la perpétuation des biais. Enfin, notre recherche vise à explorer comment la conception et la structure de ces machines sociales peuvent être modifiées pour atténuer les biais, en englobant à la fois des changements procéduraux et des modifications des composants techniques.

Pour examiner le concept d'équité en apprentissage automatique dans le contexte plus large des machines sociales, nous devons d'abord définir puis analyser les composants impliqués dans l'évaluation de l'équité en apprentissage automatique. Dans l'apprentissage automatique, le processus de vérification de l'équité implique généralement plusieurs aspects :

1. Détection des biais : Tout d'abord, les chercheurs vérifient l'existence de biais dans le système en analysant les données d'entrée et de sortie. Ils identifient s'il y a des différences systématiques dans le traitement ou les résultats pour différents groupes.

2. Identification des groupes affectés : Une fois un biais détecté, ils déterminent

quels groupes spécifiques sont affectés par ce biais (par exemple, racial et de genre).

3. Évaluation de l'impact : Ensuite, ils évaluent l'impact des biais identifiés sur ces groupes, en comprenant l'étendue du préjudice causé. Cela implique de quantifier les effets négatifs et de comprendre comment ces biais perpétuent les inégalités.

4. Mesure du biais : Le biais identifié est mesuré à l'aide de critères spécifiques d'équité. Les critères courants incluent l'Equalized Odds, qui veille à ce que les taux d'erreur soient égaux entre les groupes, et la Predictive Parity, qui garantit que les prédictions positives sont également précises pour tous les groupes démographiques.

5. Identification de la source du biais : Les chercheurs identifient ensuite la source du biais, qui pourrait se trouver dans les données elles-mêmes ou dans le modèle d'apprentissage automatique.

6. Atténuation des biais : Enfin, ils mettent en œuvre des processus pour réduire ou éliminer le biais. Cela peut impliquer de réentraîner le modèle avec des données plus équilibrées, de modifier la conception de l'algorithme pour le rendre plus équitable, ou d'appliquer des techniques de post-traitement pour ajuster la sortie du système afin d'assurer l'équité.

Étant donné que les machines sociales englobent à la fois des personnes et

des infrastructures numériques dans l'analyse de l'équité, contrairement aux systèmes purement algorithmiques, il est crucial d'identifier le rôle spécifique que ces systèmes jouent dans la contribution aux biais. Il est important d'éviter l'attribution simpliste des biais uniquement aux facteurs humains ; au contraire, l'interaction entre les éléments sociaux et technologiques au sein de la machine sociale doit être examinée en profondeur. Pour définir et examiner les composants mentionnés ci-dessus dans les machines sociales et atteindre les objectifs mentionnés dans la section 1.1, nous visons à aborder les questions de recherche suivantes dans le cadre des objectifs.

Notre premier objectif est de comprendre les questions liées à l'équité dans les machines sociales. Cet objectif sera atteint par une étude approfondie de la littérature existante sur l'équité à travers les machines sociales, décomposée en les questions de recherche suivantes :

- **(QR1)** Lorsque des concepts liés à l'équité (tels que le "biais") sont discutés dans le contexte des machines sociales, quels phénomènes spécifiques sont évoqués ?
- **(QR2)** En ce qui concerne les problèmes d'équité entre groupes, quels groupes démographiques sont susceptibles de biais, et quels attributs démographiques protégés peuvent affecter la formation des biais ?
- **(QR3)** En quoi ces biais sont-ils considérés comme nuisibles ?
- **(QR4)** Pouvons-nous relier les biais aux attentes normatives exprimées sous forme de critères techniques d'équité ?

Nous pourrons alors satisfaire notre deuxième objectif, qui se concentre sur le rôle de la machine sociale dans la création des biais, à travers une étude de cas sur le biais de genre dans Wikipedia, ce qui répondra aux questions de recherche suivantes :

- **(QR5)** De plus, en définissant des métriques spécifiques d'équité, pouvons-nous identifier les sources de ces biais ?
- **(QR6)** Existe-t-il des relations causales claires entre ces phénomènes de biais ?

Enfin, la QR6 répond également à notre troisième objectif, qui vise à comprendre comment le biais peut être atténué en modifiant la conception d'une social machine. Pour mettre cela en pratique, nous introduisons :

- **(QR7)** Comment le biais peut-il être atténué grâce à la refonte du système de catégorisation de Wikipedia ?

1.3 Aperçu des contributions

Approche pour QR1: Afin d'explorer de manière exhaustive les machines sociales dans cette recherche, nous sélectionnons des études de cas issues de cinq grandes catégories de réseaux de collaboration de masse : les places de marché en ligne, les réseaux sociaux, les jeux en ligne multijoueurs et mondes virtuels, et les systèmes de crowdsourcing, en fonction de la densité

des interactions. Les études de cas sont ensuite choisies dans chaque groupe, en tenant compte de la richesse des données qu'elles offrent pour la recherche.

Wikipedia, en tant que réseau de collaboration de masse, est un point focal principal en raison de l'étendue des données de recherche qu'il propose. Pour capturer la diversité des interactions et mettre en évidence l'impact potentiel des comportements problématiques dans les places de marché en ligne, nous élargissons notre sélection d'études de cas pour inclure quatre plateformes différentes : Uber, Airbnb, Fiverr et TaskRabbit. Ces plateformes représentent une large gamme d'interactions dans les places de marché en ligne.

Pour les catégories des réseaux sociaux et du crowdsourcing, nos études de cas sont respectivement Reddit, YouTube, et les systèmes d'évaluation par les pairs académiques. Comme la plupart des études sur les jeux en ligne s'appuient sur des interviews et des enquêtes auprès de joueurs et joueuses concernant leurs expériences sans se concentrer sur un jeu spécifique, nous investiguons les biais dans cette catégorie de manière générale, sans nous focaliser sur un jeu particulier.

Nous procédons donc d'abord à une revue systématique des études sur les "biais" dans les machines sociales. Notre revue systématique couvre 196 articles étudiant les "biais" dans les machines sociales, incluant 86 articles sur Wikipedia, 26 articles sur les places de marché en ligne, 31 articles sur les réseaux sociaux, 12 articles sur les jeux en ligne et 41 articles sur l'évaluation par les pairs.

Approche pour QR2: Nous catégorisons les phénomènes de biais en deux types principaux : ceux formés en fonction d'attributs sensibles (tels que le genre, la culture, la race et l'orientation sexuelle) et ceux spécifiques à certaines machines sociales (comme le statut d'inscription sur Wikipedia ou l'affiliation académique). L'analyse révèle des biais communs à de nombreuses machines sociales, en particulier lorsque les attributs sensibles sont visibles, conduisant à des comportements biaisés et à des résultats injustes. Il est suggéré que concevoir des machines sociales pour masquer les attributs sensibles, lorsque cela n'est pas essentiel, peut aider à atténuer ces biais. Des exemples incluent l'utilisation de l'évaluation en double aveugle dans les contextes académiques et l'utilisation d'avatars non humains sur des plateformes comme Fiverr pour dissimuler la race et le genre, réduisant ainsi la probabilité de comportements discriminatoires. Nous notons également que certains biais, tels que ceux fondés sur la religion ou l'orientation sexuelle, sont plus difficiles à détecter en raison de leur nature moins visible.

Approche pour QR3: Pour répondre à cette question de recherche, nous avons mené une analyse détaillée des biais identifiés dans les machines sociales, en nous concentrant sur la nature et l'étendue des notions établies de préjudice que ces biais causent, guidés par la taxonomie des préjudices de Shelby et al. Nous avons lié chaque phénomène de biais à des catégories de préjudice connexes, en les regroupant en un ensemble plus restreint de phénomènes abstraits basés sur les préjudices spécifiques qu'ils produisent.

Notre analyse a révélé que ces biais entraînent souvent des conséquences importantes, telles que la sous-participation et la sous-représentation des groupes minoritaires, une représentation biaisée, et un impact disparate.

Approche pour QR4: Dans cette approche, nous proposons que l'équité dans les machines sociales (MSs) puisse être efficacement analysée en utilisant des concepts et des techniques issus des systèmes algorithmiques. L'analyse se concentre sur la question de savoir si les biais identifiés correspondent aux critères techniques d'équité établis et utilisés dans les contextes algorithmiques. Le modèle principal appliqué est celui des systèmes de décision binaire automatisés, où les décisions sont basées sur des données associées à des individus, conduisant à des résultats souhaitables ou non souhaitables. Les phénomènes de biais, tels que la sous-représentation et l'impact disparate, peuvent souvent être modélisés dans ce cadre. Par exemple, des problèmes comme la sur-police et les biais de notation peuvent être compris comme des décisions binaires qui acceptent ou rejettent la contribution ou l'attribut d'une personne. Pour évaluer l'équité, divers critères d'équité de groupe, y compris la parité démographique, l'égalité des chances et l'égalité des chances en résultat, sont utilisés.

Approche pour QR5: Nous fournissons une analyse approfondie des biais de genre au sein de Wikipédia, en examinant spécifiquement la sous-représentation (biais de couverture) et la représentation biaisée (biais lexical) des femmes. Notre objectif est de déterminer si ces biais sur Wikipédia

reflètent les biais sociétaux ou si la machine sociale de Wikipédia génère ses propres biais, semblables aux biais algorithmiques. Étant donné la prévalence des biais de genre dans la société, nous explorons si Wikipédia reflète fidèlement ces biais ou les déforme à travers ses processus socio-techniques.

Wikipédia, une machine sociale où les actions humaines et automatisées interagissent, peut être évaluée en termes d'équité en utilisant les principes d'équité algorithmique. Nous évaluons si Wikipédia est juste ou biaisé en fonction de critères tels que la parité démographique et l'égalité des chances. Par exemple, l'équité pourrait signifier une représentation égale des hommes et des femmes dans les biographies, en tenant compte de leur notoriété dans la société.

Nous examinons également la représentation biaisée en analysant le contenu des biographies pour voir si les rôles traditionnels de genre sont trop mis en avant. En utilisant un ensemble de données de biographies de politiciens américains, nous contrôlons les facteurs de confusion comme la profession. Nos résultats suggèrent que le biais de couverture est moins prononcé chez les politiciens par rapport à d'autres groupes, et les techniques modernes de traitement du langage naturel révèlent que les différences de contenu sont en grande partie factuelles, conformes aux rôles de genre traditionnels plutôt que de manifester un biais préjudiciable.

Approche pour QR6: Nous explorons les relations causales entre les divers phénomènes biaisés observés sur Wikipédia, en utilisant la plateforme

comme étude de cas pour répondre à la question de recherche : existe-t-il des relations causales claires entre ces phénomènes problématiques ? Nous suggérons que ces phénomènes sont interconnectés, avec des facteurs sociaux et techniques jouant tous deux un rôle dans la perpétuation des biais sur la plateforme. De plus, la complexité des relations causales entre les différents phénomènes de biais justifie une vision holistique de la machine sociale (SM), car les interventions sur un phénomène biaisé sont susceptibles d'avoir des conséquences sur plusieurs autres.

Approche pour la QR7 : Afin d'atteindre notre troisième objectif — comprendre comment les biais peuvent être atténués en modifiant la conception d'une social machine — nous procédons à une refonte de l'interface de catégorisation de Wikipedia. En nous appuyant sur les résultats de la QR6, cette refonte se concentre sur les problèmes structurels et représentationnels identifiés dans la catégorisation des biographies sur Wikipedia. L'un des principaux problèmes est l'isolement des catégories fondées sur l'identité — un schéma où les individus sont classés uniquement selon des attributs tels que le genre ou l'ethnicité (par exemple, "Women novelists"), sans être également placés dans des catégories plus larges, fondées sur la profession (par exemple, "Women novelists"). Cette pratique peut renforcer la marginalisation en présentant l'identité comme le trait définissant, plutôt qu'en situant les individus dans leurs domaines d'expertise. En développant et en mettant en œuvre un cadre de catégorisation fondé sur les attributs, nous montrons com-

ment des conceptions alternatives d'interface et de schéma peuvent réduire ce type d'isolement catégoriel et promouvoir des représentations plus équitables entre les groupes démographiques. Cette intervention fondée sur la conception complète nos résultats empiriques et met en évidence des pistes pratiques pour atténuer les biais au sein des social machines.

1.4 Structure de la thèse

Le chapitre 2 explore les principaux critères d'équité pour les systèmes d'apprentissage automatique, en examinant les raisons sous-jacentes à l'émergence des biais dans ces systèmes. Ce chapitre vise à fournir une compréhension complète de la manière dont les biais apparaissent et des métriques utilisées pour évaluer l'équité. Le chapitre 3 se concentre sur les systèmes socio-techniques et les machines sociales, offrant un aperçu détaillé de leurs différents types et caractéristiques. Il comprend également une revue de la littérature existante sur l'équité au sein des machines sociales, en mettant en lumière les principales études et découvertes dans ce domaine de recherche émergent.

Chapitre 4 décrit la méthodologie utilisée pour mener une revue systématique des problèmes d'équité dans les machines sociales. Il inclut la sélection des études de cas, l'identification et la catégorisation des phénomènes de biais, ainsi que l'analyse des groupes sociaux affectés par les biais.

Chapitre 5 approfondit les phénomènes de biais, les dommages résultants, et les critères d'équité au sein des machines sociales. Les phénomènes rap-

portés sont regroupés en un petit nombre de phénomènes plus abstraits. Ce chapitre explore également les métriques d'équité décisionnelle comme la Parité Démographique et les Chances Égalisées, fournissant un cadre complet pour évaluer l'équité dans les systèmes socio-techniques.

Dans le chapitre 6 nous nous penchons sur le cas spécifique du biais de genre dans les biographies de Wikipédia, en nous concentrant particulièrement sur le biais de couverture et les représentations biaisées des femmes à travers les biographies des politiciennes américaines. Nous déterminons si la représentation des femmes politiques sur Wikipédia est un reflet honnête de notre société biaisée ou si cette machine sociale génère ses propres biais, similaires aux biais algorithmiques.

Chapitre 7 présente une analyse détaillée des relations de causalité au sein de Wikipédia, en se concentrant sur des problèmes tels que la sous-participation, la sous-représentation des membres de certains groupes, et la représentation biaisée. Il fournit des perspectives sur les interactions complexes au sein des machines sociales et les problèmes d'équité qui en résultent.

Le chapitre 8 aborde notre troisième objectif en présentant une proposition de refonte du système de catégorisation de Wikipedia, en mettant l'accent sur la manière dont des modifications structurelles peuvent contribuer à réduire les biais. Il détaille le développement et la mise en œuvre d'un cadre de catégorisation fondé sur les attributs, conçu pour assurer une représentation plus cohérente et plus équitable entre les groupes démographiques.

Enfin, le chapitre 9 propose un résumé des principaux résultats et conclusions de cette recherche. Le chapitre inclut également une liste des publications dérivées issues de ce travail.

Chapitre 2

Contexte : L'équité algorithmique

Dans ce chapitre, nous explorerons le rôle crucial des systèmes de prise de décision automatisée, en particulier ceux alimentés par l'apprentissage automatique (Machine Learning; ML), et leur influence croissante sur la société. Nous examinerons les questions d'équité qui surgissent au sein de ces systèmes, en discutant de la manière dont les biais peuvent se manifester et se propager à travers les modèles de ML. De plus, nous introduirons divers critères d'équité qui peuvent être appliqués à ces systèmes pour atténuer les biais. Enfin, nous approfondirons les raisons sous-jacentes pour lesquelles les biais se produisent dans le ML, y compris les échantillons biaisés et inadéquats, ainsi que le biais indirect, afin de poser les bases de la compréhension des implications plus larges de l'équité dans les machines

sociales.

2.1 Décision automatisée avec le ML

L'apprentissage automatique (ML) a connu une croissance exponentielle et joue aujourd'hui un rôle significatif dans les décisions cruciales prises au sein des sociétés humaines, telles que le recrutement [17, 61], les soins de santé [31, 68], et l'éducation [89, 90, 91].

Les décisions qui étaient auparavant prises par des humains affectaient une communauté cible plus restreinte. Avec la croissance continue de la technologie et la prolifération des systèmes de prise de décision automatisée, l'apprentissage automatique peut affecter un large éventail d'individus dans la société. La plupart de ces décisions automatiques sont prises à l'aide du ML. Chaque décision représente en réalité un choix parmi plusieurs options qui peuvent être mises en œuvre par le ML. Si nous considérons chacune des différentes options dans la prise de décision automatique comme une classe, nous pouvons effectuer une prise de décision automatique avec un classificateur ML.

Par exemple, le système d'admission universitaire décide si un individu peut être admis à l'université en fonction de son profil. Un classificateur ML est utilisé pour mettre en œuvre cette décision automatique. Le système est formé en utilisant les caractéristiques des individus qui ont postulé à l'université dans le passé et dont la candidature a été "acceptée" or "re-

jetée". Les données sont fournies à la machine, et la machine apprend que chaque individu appartient à la classe "acceptée" ou "rejetée" en fonction de ses caractéristiques. Après l'apprentissage, la machine peut décider si un nouvel individu qui postule actuellement pour une admission appartient à la classe "acceptée" ou à la classe "rejetée". La technologie de reconnaissance faciale est l'un des outils qui utilise l'apprentissage automatique pour "confirmer" or "nier" l'identité des individus et est largement utilisée dans divers centres publics et privés. Dans le système COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), conçu comme un logiciel d'assistance et un outil de soutien utilisé pour prédire le risque de récidive, le système décide si un individu appartient à une classe de "ceux qui doivent être maintenus en prison" ou à une classe de "ceux qui peuvent être libérés temporairement". Dans le système de recrutement, la machine décide si un individu doit être embauché ou non. Ce sont des exemples que nous référons dans diverses sections de cette recherche.

2.2 Problèmes d'équité

Dans les sociétés humaines, malgré les lois qui ont été édictées pour prévenir les violations des droits de l'homme et la non-discrimination entre les individus et les différents groupes de la société, selon notre contexte mental et historique en tant qu'êtres humains, la discrimination est encore visible dans certaines décisions. Par exemple, dans les admissions universitaires,

les recherches montrent que le taux d'erreur pour les femmes à la peau plus foncée est d'environ 35 %, contre moins de 1 % pour les hommes à la peau plus claire.

Ainsi, disposer d'un système de prise de décision équitable capable de décider au nom des êtres humains et de réduire la discrimination dans les sociétés humaines est plus remarquable que jamais. Si le système de prise de décision est basé sur des données discriminatoires, même si le biais est minime, ce biais peut s'aggraver à long terme. De plus, si le système de décision automatique n'est pas équitable, il est de loin plus destructeur qu'un décideur humain injuste, car le décideur automatique peut affecter une société cible plus large. En revanche, si le décideur automatique est équitable, non seulement il peut prendre des décisions justes actuellement, mais il peut également réduire le biais historique existant à long terme. Par exemple, un système équitable d'admissions universitaires peut accroître la participation des femmes dans la société à long terme et réduire l'ancienne vision patriarcale qui persiste encore dans certaines sociétés.

Nous croyons que l'apprentissage automatique et les machines sociales sont similaires en termes de prise de décision et de leur impact dans les sociétés. Il est donc probablement possible de définir certains concepts d'équité de manière à ce qu'ils puissent être utilisés dans les machines sociales. Aujourd'hui, les machines sociales sont devenues une partie intégrante de la vie humaine. En raison de la croissance croissante des machines sociales et du grand nombre de leurs utilisateurs, leurs effets, qu'ils soient positifs ou

négatifs, peuvent toucher un large éventail d'utilisateurs. Considérons, par exemple, une personne qui reçoit des centaines ou des milliers de commentaires négatifs ou de menaces en réponse à un commentaire qu'elle a fait, ce qui aurait été impossible sans les machines sociales. Par conséquent, nous croyons que le concept d'équité dans les machines sociales est aussi important que l'apprentissage automatique, et nous devrions avoir des critères pour mesurer les comportements sociaux équitables et injustes des utilisateurs, ainsi que la réponse équitable des machines sociales aux utilisateurs. Les sous-sections suivantes expliquent les raisons pour lesquelles des biais peuvent se produire dans l'apprentissage automatique [18, 11], et les critères d'équité pour les systèmes ML [34, 125, 135].

2.3 Critères d'équité pour les systèmes ML

Afin de pouvoir fournir une définition mathématique pour chaque groupe de critères d'équité, en considérant un modèle d'apprentissage automatique comme un classificateur binaire, nous utilisons les notations suivantes :

- P est un prédicteur binaire
- A est une caractéristique sensible telle que le sexe, le genre, la race, ...
- X est un ensemble d'attributs excluant A.
- $Y \in \{0, 1\}$ est la valeur cible du prédicteur P .

- $\hat{Y} \in \{0, 1\}$ est la valeur prédite ou la sortie du classificateur.

Le prédicteur P souhaite prédire Y à partir d'un ensemble d'attributs X et produire une valeur prédite Y. Pour Y, l'une de ces valeurs (1) est considérée comme la valeur positive et est supposée être plus désirable que la valeur négative (0). Étant donné que les critères de groupe mentionnés dans cette recherche sont basés sur la matrice de confusion, rappeler les éléments de la matrice de confusion aide le lecteur à avoir une meilleure vue :

- Vrai positif (VP): le nombre total de prédictions correctes lorsque la classe réelle était positive.
- Faux positif (FP): le nombre total de prédictions erronées lorsque la classe réelle était positive.
- Faux négatif (FN): le nombre total de prédictions erronées lorsque la classe réelle était négative.
- Vrai négatif (VN): le nombre total de prédictions correctes lorsque la classe réelle était négative.

La figure 2.1 montre le résultat d'un classificateur hypothétique. Les échantillons rouges indiquent les positifs réels et les échantillons bleus indiquent les négatifs réels. Les échantillons au-dessus de la ligne en pointillés ont été prédits positivement par le classificateur. Les échantillons en dessous de la ligne en pointillés ont été prédits négativement par le classificateur.

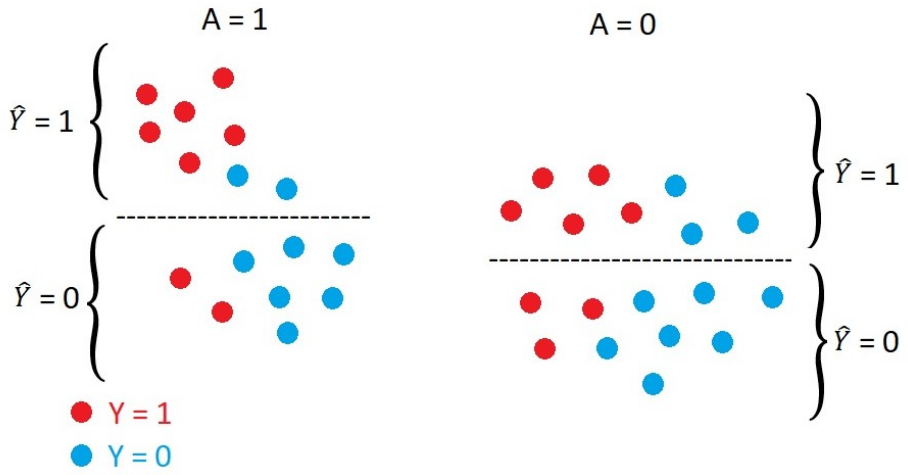


Figure 2.1. *Les points rouges indiquent les positifs réels et les points bleus indiquent les négatifs réels, qui sont séparés à gauche et à droite en fonction de la valeur de l'attribut sensible A.*

2.3.1 Inconscience

L'inconscience est compatible avec le concept de traitement disparate, qui exige de ne pas utiliser un attribut sensible (protégé). Pour satisfaire à l'équité par inconscience, les données d'entraînement utilisées ne doivent pas contenir d'attributs sensibles. C'est la manière la plus simple et la plus basique de satisfaire à l'équité en apprentissage automatique, mais ce n'est pas suffisant [46, 125]. Il peut exister des caractéristiques fortement corrélées avec des attributs sensibles, et cette méthode n'est pas en mesure de les détecter.

2.3.2 Équité de groupe

Il existe plusieurs critères d'équité que nous avons placés dans la catégorie de l'équité. Ici, le critère d'équité est le respect de l'égalité entre les groupes.

2.3.2.1 Parité démographique

La parité démographique (Demographic Parity) est l'un des critères les plus connus, également appelée indépendance ainsi que parité statistique [135, 137]. La parité démographique exige qu'une décision, telle que l'acceptation ou le rejet d'un prêt, soit prise indépendamment d'un attribut sensible. La parité démographique exprime que la proportion de chaque groupe d'une classe sensible (par exemple, le genre ou la race) devrait recevoir une sortie positive à des taux égaux. Une sortie positive réelle peut conduire, par exemple, à "entrer à l'université" ou à "obtenir un prêt" [135, 137]. La formulation est montrée ci-dessous.

$$P[\hat{Y} = y|A = 0] = P[\hat{Y} = y|A = 1] \tag{2.1}$$

Ce concept présente deux inconvénients majeurs [135, 59, 137]. Tout d'abord, il ne garantit pas l'équité. Avec cette méthode, nous pouvons accepter des personnes qualifiées dans un groupe, mais dans l'autre groupe, en raison du petit nombre d'échantillons, un certain nombre de personnes non qualifiées seront également acceptées. Cela se produit lorsque les

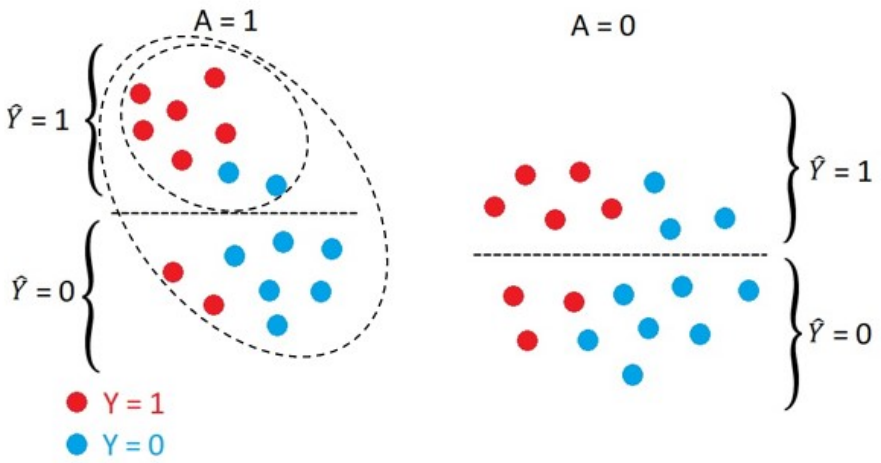


Figure 2.2. *Les points rouges indiquent les positifs réels et les points bleus indiquent les négatifs réels. Les cercles en pointillés montrent* $\frac{Positifs}{Positifs\ +\ Négatifs}$

données d'entraînement d'un groupe sont plus nombreuses que celles de l'autre groupe. Le deuxième inconvénient est que cette méthode peut grandement affecter la précision du système de décision et la réduire. Cela signifie qu'en équilibrant les personnes sélectionnées dans différents groupes, un certain nombre de personnes non qualifiées peuvent être sélectionnées dans un groupe et réduire la précision du système.

2.3.2.2 Equalized Odds

Equalized odds, également appelées séparation et parité du taux positif, examinent l'établissement de l'équité sous deux aspects[59] :

1. L'égalité de la détection correcte des sorties positives entre différents groupes (la valeur de $\frac{VP}{(VP+FN)}$ ou taux de vrais positifs (TVP) pour le

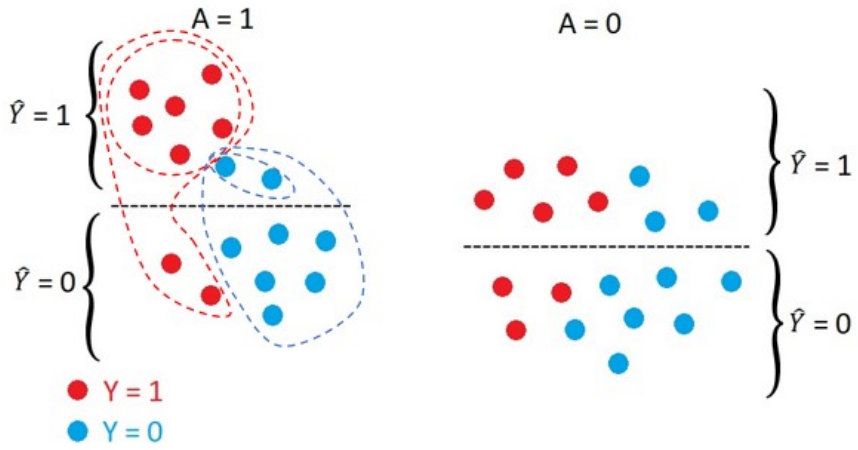


Figure 2.3. *Les points rouges indiquent les positifs réels et les points bleus indiquent les négatifs réels. Les cercles en pointillés rouges et bleus montrent respectivement* $\frac{VP}{VP+FN}$ *et* $\frac{FP}{FP+VN}$.

groupe $A = 0$ et le groupe $A = 1$ devrait être égale).

2. L'égalité de la détection correcte des sorties négatives entre différents groupes (la valeur de $\frac{FP}{(FP+VN)}$ ou taux de faux positifs (TFP) pour le groupe $A = 0$ et le groupe $A = 1$ devrait être égale).

Equalized odds est satisfaite si la prédiction est conditionnellement indépendante de l'attribut sensible A étant donné la valeur vraie y:

$$P[\hat{Y} = 1|Y = y, A = 0] = P[\hat{Y} = 1|Y = y, A = 1], \hat{Y} \in \{0, 1\} \qquad (2.2)$$

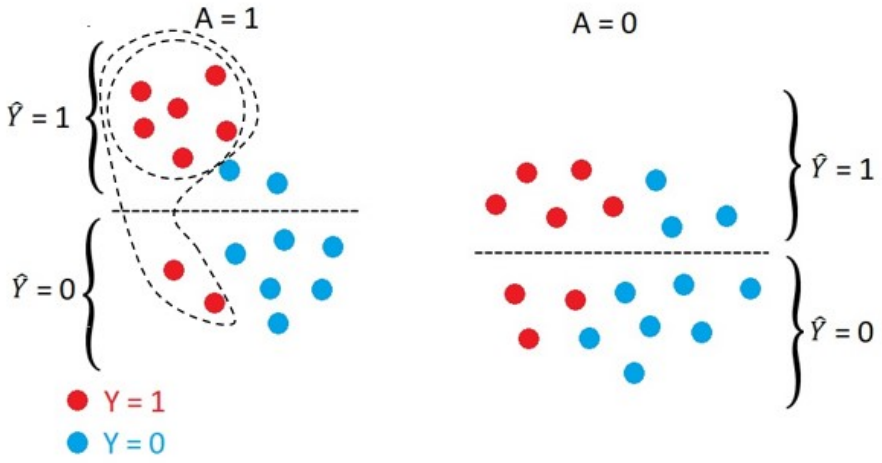


Figure 2.4. *Les points rouges indiquent les positifs réels et les points bleus indiquent les négatifs réels. Les cercles en pointillés montrent* $\frac{VP}{VP+FN}$.

2.3.2.3 Equal opportunity

Equal opportunity est une notion plus faible que celle Equalized odd [59]. Dans de nombreuses applications (par exemple, un système d'alerte aux dangers), les gens se préoccupent davantage du taux de positifs réels ($\frac{VP}{(VP+FN)}$) que du taux de négatifs réels. Ainsi, dans de nombreuses applications, la formule suivante, qui est une version assouplie Equalized odds, peut être utilisée [59].

$$P[\hat{Y} = 1|Y = 1, A = 0] = P[\hat{Y} = 1|Y = 1, A = 1] \tag{2.3}$$

Equal opportunity et Equalized Odds qui sont une forme générale de ce critère, peuvent ne pas aider à atténuer la discrimination dans la société

[59]. Considérons une société où les femmes ont des difficultés à accéder à l'éducation de base en raison de la discrimination. Par conséquent, le nombre de femmes qui postulent pour entrer à l'université est beaucoup moins élevé que celui des hommes. Afin de pouvoir atténuer ce biais dans la société, la fraction du nombre de femmes admises par rapport au nombre total de femmes qualifiées doit être supérieure à cette fraction pour les hommes. Par exemple, si 200 hommes et 20 femmes sont qualifiés pour entrer à l'université et que la capacité d'admission est de 55, le choix de 5 femmes et de 50 hommes satisfera à cette méthode. Toutefois, cette méthode de sélection non seulement ne réduit pas la discrimination mais perpétue également ce biais dans la société. Si davantage de femmes étaient admises, cette discrimination serait réduite à l'avenir.

2.3.2.4 Predictive Rate Parity

Predictive Rate Parity, , également appelée Sufficiency [49]. Dans ce critère, pour qu'un système soit reconnu comme équitable, les deux aspects suivants doivent être satisfaits :

- Parité Prédictive Positive ($\frac{VP}{VP+FP}$)

$$P[Y = 1|\hat{Y} = 1, A = 0] = P[Y = 1|\hat{Y} = 1, A = 1] \tag{2.4}$$

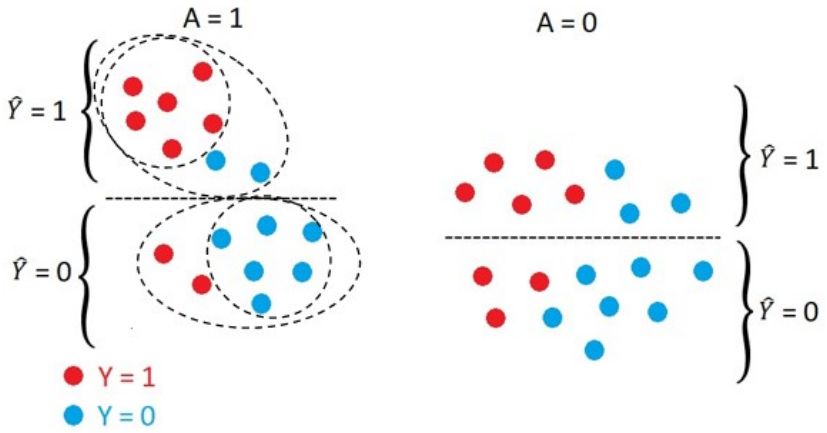


Figure 2.5. *Les points rouges indiquent les positifs réels et les points bleus indiquent les négatifs réels. Les cercles en pointillés au-dessus de la ligne pointillée indiquent* $\frac{VP}{VP+FP}$ *et en dessous de la ligne pointillée indiquent* $\frac{VN}{VN+FN}$.

- Parité Prédictive Négative ($\frac{VN}{VN+FN}$)

$$P[Y = 0|\hat{Y} = 0, A = 0] = P[Y = 0|\hat{Y} = 0, A = 1] \tag{2.5}$$

Si nous considérons l'exemple de l'admission des hommes et des femmes à l'université mentionné ci-dessus pour la Parité Prédictive, cette méthode, tout comme les Odds Égalisés, ne serait pas capable d'atténuer la discrimination dans la société.

2.3.2.5 Autres mesures d'équité

Bien que l'équité individuelle et l'équité contrefactuelle soient des critères importants pour évaluer l'équité dans les systèmes d'apprentissage automa-

tique, ils ne seront pas le principal objectif de cette thèse. Cette thèse se concentrera plutôt sur l'équité de groupe, telle que définie par des métriques plus courantes et largement utilisées, qui correspondent mieux aux objectifs de notre recherche.

Équité Individuelle Cette méthode est complètement différente des concepts d'équité de groupe. Dans l'équité de groupe, le critère est basé sur la comparaison de deux groupes, mais dans cette méthode, le critère d'équité est les individus. Pour atteindre l'équité, des décisions similaires doivent être prises concernant les individus qui sont similaires [34].

Pour réaliser l'équité basée sur les individus, une métrique de distance est considérée pour mesurer la similarité entre les individus en fonction de leurs caractéristiques. Le principe de Lipschitz est utilisé pour calculer la distance. Si deux individus x_1 et x_2 sont à une distance $d(x_1, x_2)$ l'un de l'autre, leur résultat prédit $(\hat{Y}(x_1), \hat{Y}(x_2))$ doit avoir une distance statistique d'au plus $d(x_1, x_2)$ [34].

Il est plus difficile que prévu de sélectionner des caractéristiques des individus dont la distance et la similarité peuvent être mesurées. Il peut y avoir des caractéristiques qui ne sont pas considérées dans le système de décision, mais dont nous avons besoin pour mesurer la similarité des personnes [43].

Pour de nombreux partisans de l'équité individuelle, cette méthode semble idéale, mais avec les inconvénients de cette méthode en pratique, il est préférable de ne pas la considérer comme une méthode complète pour

définir l'équité. Si elle peut être appliquée au système, il est préférable de la considérer comme un critère en complément d'autres critères pour mesurer l'équité.

Équité Contrefactuelle L'équité contrefactuelle est basée sur la causalité. L'équité est mesurée par le modèle causal qui est construit pour un système d'apprentissage automatique en utilisant les caractéristiques d'entrée. L'équité contrefactuelle considère qu'un système est équitable si, dans le cas où les caractéristiques sensibles avaient une valeur différente de la valeur actuelle, la valeur de sortie prédite du système aurait été égale à la valeur obtenue en fonction de la valeur des caractéristiques sensibles dans la réalité. La formule suivante exprime cela mathématiquement.

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \quad (2.6)$$

2.4 Raisons pour lesquelles le biais peut se produire dans l'apprentissage automatique

Comme mentionné précédemment, dans l'apprentissage automatique, l'apprentissage se fait à travers les données. Les données collectées dans les sociétés humaines peuvent être biaisées, consciemment ou inconsciemment, en raison de la discrimination dans les sociétés humaines. Par conséquent, l'entraînement d'une machine avec des données biaisées conduit à la forma-

tion d'un système biaisé. Dans cette thèse, basé sur des caractéristiques sensibles, un groupe minoritaire est défini comme un groupe culturellement, ethnically ou racialement distinct qui est subordonné à un groupe plus dominant (majorité). Voici les principales raisons qui peuvent causer des biais.

2.4.1 Échantillons biaisés

- *Échantillons discriminatoires :* Si l'environnement réel à partir duquel les données sont collectées est biaisé par les humains, les données collectées seront biaisées. Par exemple, considérez un système de décision d'embauche qui se base sur les mérites des individus. Si le système est formé avec des données reflétant les résultats d'embauche de gestionnaires biaisés, le système conçu aura le même biais que ces gestionnaires [18, 11].

- *Échantillon avec une distribution biaisée :* Si les échantillons représentant le paysage du problème ne couvrent qu'une partie de l'espace problème plutôt que l'ensemble de l'espace, une dispersion incorrecte des échantillons peut provoquer un biais. Ce biais peut devenir plus sévère avec le temps [18, 11]. Prenons l'exemple du système de dossier de la police. Le taux de criminalité dans chaque zone est déterminé en fonction des rapports de police. Le département de police tend à envoyer plus d'agents dans les endroits où le taux de criminalité est plus élevé. Par conséquent, les crimes sont plus susceptibles d'être enregistrés dans ces zones en raison

du plus grand nombre de policiers. Même si les gens dans d'autres zones commettent plus de crimes par la suite, le département de police peut continuer à enregistrer des taux de criminalité plus bas en raison d'une attention moindre et du manque de policiers dans ces zones. Il peut y avoir des crimes dans ces zones, mais les données d'entraînement du système n'incluent pas ces échantillons, et avec le temps, le nombre d'échantillons provenant de cette zone diminue par rapport aux zones avec plus d'agents, et la dispersion des échantillons devient progressivement biaisée. Cela signifie que le système conçu sur cette base a une vue négative des zones avec le taux de criminalité initial le plus élevé, et ce biais augmente progressivement [18, 11].

2.4.2 Échantillons inadéquats

- *Caractéristiques limitées :* Les caractéristiques peuvent ne pas être informatives ; cela signifie que les caractéristiques fournissent moins d'informations pour un groupe minoritaire que pour un groupe majoritaire. Un système conçu sur la base de ce type de données peut ne pas être suffisamment précis pour étiqueter un groupe minoritaire [18, 11].

- *Disparité de taille d'échantillon :* Si les échantillons d'entraînement ne sont pas équilibrés. Cela signifie que le nombre d'échantillons d'entraînement pour le groupe minoritaire est inférieur à celui du groupe majoritaire. Le système est entraîné de manière injuste et ne sera pas suffisam-

ment précis pour le groupe minoritaire [18, 11]. Par exemple, si dans le système de reconnaissance faciale, le nombre d'images d'hommes à peau claire dans les données d'entraînement est supérieur à celui des femmes à peau foncée, cela entraînera un biais dans le système.

2.4.3 Biais indirect

En supposant que les caractéristiques sensibles ne sont pas utilisées pour l'entraînement du système, il peut y avoir d'autres caractéristiques qui pourraient être affectées par les caractéristiques sensibles et porter indirectement le poids des caractéristiques sensibles. Dans ce cas, utiliser ces caractéristiques pour l'entraînement du système peut provoquer un biais dans le système [18, 11]. Par exemple, Amazon utilise un système qui propose un service de livraison gratuite le jour même dans certains quartiers des États-Unis en fonction du nombre d'acheteurs et du montant qu'ils achètent. Selon une étude de 2016, dans de nombreuses villes américaines, le nombre de quartiers sélectionnés pour le service de livraison gratuite le jour même pour les quartiers blancs est plus de deux fois supérieur à celui des quartiers noirs. Leur objectif n'était pas de discriminer en fonction de la race, mais le système a pu utiliser des caractéristiques corrélées avec des caractéristiques sensibles et créer indirectement un système biaisé [58].

2.5 Résumé

Bien que les causes de biais dans les systèmes d'apprentissage automatique découlent en grande partie de l'entraînement basé sur les données, ces mêmes causes peuvent ne pas s'appliquer dans le contexte des machines sociales, où il n'y a pas de données d'entraînement. Cette distinction soulève des questions importantes sur les sources de biais dans les machines sociales. Bien que les critères d'équité développés pour les systèmes d'apprentissage automatique puissent être pertinents si nous modélisons une machine sociale comme un système de classification, les causes sous-jacentes du biais peuvent différer considérablement en raison de l'absence de données d'entraînement. Cela nécessite une exploration plus approfondie de la façon dont les biais émergent et se propagent au sein des machines sociales, et de la manière dont l'équité peut être assurée dans ces machines sociales complexes.

Chapitre 3

État de l'art: Équité dans les machines sociales

Ce chapitre explore le concept d'équité dans le cadre plus large des machines sociales. Il commence par définir les machines sociales, en les distinguant des systèmes socio-techniques traditionnels en mettant l'accent sur leur dépendance à l'infrastructure informatique, qui permet des réalisations collaboratives que ni les humains ni les machines ne pourraient accomplir indépendamment. Après cette vue d'ensemble conceptuelle, nous abordons la question critique de l'équité au sein des machines sociales. Nous passons en revue les études clés pertinentes à ce sujet, y compris les recherches sur l'équité dans l'apprentissage automatique, et nous cherchons à étendre les limites traditionnelles de l'équité algorithmique pour englober les contextes dans lesquels les machines sociales opèrent. Nous abordons également

des travaux sur l'équité liée au traitement automatique des langues (Natural language processing ; NLP), en mettant particulièrement en évidence les complexités et les défis de l'identification et de l'atténuation des biais, comme discuté dans l'enquête de Blodgett. De plus, nous examinons les audits algorithmiques des plateformes sociales, qui révèlent l'interaction entre les hiérarchies sociales et les biais perçus au sein de ces écosystèmes numériques. En synthétisant ces divers courants de recherche, le chapitre vise à offrir une perspective complète sur le discours évolutif autour de l'équité dans les machines sociales et démontre la nécessité d'une revue systématique plus large qui va au-delà des systèmes algorithmiques pour comprendre pleinement et aborder l'équité au sein des machines sociales complexes.

3.1 Systèmes socio-techniques et machines sociales

Les machines sociales sont une classe de systèmes conçus pour organiser un travail complexe où humains et machines interagissent de manière à ce que l'infrastructure informatique soutienne la créativité humaine. Les machines sociales font référence à l'interaction entre les infrastructures complexes de la société et le comportement humain. Dans cette section, nous expliquons le concept de machines sociales, ainsi que la théorie des systèmes socio-techniques qui consiste fondamentalement à étudier la technologie et ses utilisateurs ensemble plutôt que séparément [123, 60, 101]. Cette section

aborde les machines sociales dans différentes catégories basées sur les relations homme-machine et les interactions entre eux à travers divers exemples de machines sociales [85, 123, 88, 27, 45].

La théorie des systèmes socio-techniques fournit un aperçu du double façonnage de la technologie et du contexte social, et reconnaît les organisations comme des systèmes complexes d'humains et de technologies visant à atteindre des objectifs donnés dans le cadre d'un environnement organisationnel donné [123]. Les machines sociales sont un concept relativement nouveau, plus spécifique que les SST car elles impliquent que la technologie dans les SST est principalement un réseau d'ordinateurs (y compris les smartphones, etc.). Les machines sociales ne sont également pas limitées au "travail" au sens du travail salarié, alors que les STS ont été conceptualisés à une époque où les machines ou les ordinateurs étaient principalement utilisés pour le travail.

Dans les machines sociales, le collectif homme-machine réalise des choses plus grandes que ce que les 'parties' individuelles pourraient accomplir seules. Le terme "machines sociales" introduit par Hendler et al.[60] désigne une classe de systèmes où les humains et les machines interagissent de manière à ce que l'infrastructure informatique soutienne la créativité humaine. Par la suite, Shadbolt et al. ont exploré ce concept à travers plusieurs travaux en 2013, 2016 et 2019 [105, 106, 107]. Malgré ces recherches, une définition complète du terme reste insaisissable.

Ce sont des systèmes "dans lesquels les gens font le travail créatif et

la machine fait l'administration" [60]. La machine sociale peut atteindre des objectifs que ni la machine seule ni l'humain ne peuvent accomplir. Les machines sociales prennent de nombreuses formes. Dans leur livre, Shadbolt et al. [105] fournissent de nombreux exemples de machines sociales sans les catégoriser clairement. À l'inverse, Tsvetkova et al. [123] dans leurs recherches sur les réseaux homme-machine, proposent une catégorisation structurée. En comparant les exemples cités par Shadbolt et al. [105] avec ceux de la catégorisation de Tsvetkova, nous avons découvert un chevauchement significatif. Plus précisément, nous avons trouvé des exemples identiques. En nous basant sur ce terrain commun, nous proposons d'adopter cinq des huit catégories suivantes : 1- Calcul des ressources publiques, 2- Crowdsourcing, 3- Moteurs de recherche sur le web, 4- Crowdsensing, 5- Marchés en ligne, 6- Médias sociaux, 7- Jeux en ligne multijoueurs et mondes virtuels, 8- Collaboration de masse, délimitées dans le cadre de Tsvetkova pour les réseaux homme-machine, comme moyen de catégoriser les machines sociales. Cette catégorisation nous permet de définir clairement le périmètre de notre recherche, en spécifiant les types de systèmes auxquels notre étude s'applique.

3.1.1 Catégories de machines sociales

- **Les réseaux de collaboration de masse** sont une forme d'actions collectives qui se produisent lorsque de grands nombres de personnes travaillent indépendamment sur un même projet qui se déroule sur

Internet en utilisant des logiciels sociaux [123]. Wikipédia et les projets de logiciels open-source tels que Linux présentent deux des exemples les plus marquants de réseaux de collaboration de masse.

- **Les marchés en ligne** sont des sites de commerce électronique qui connectent les prestataires de services aux clients, tels qu'Uber, Airbnb, Fiverr et TaskRabbit. Dans ces systèmes, les utilisateurs peuvent s'évaluer et se noter mutuellement [85].

- **Les médias sociaux** sont des applications technologiques basées sur l'informatique qui facilitent le partage d'idées, de pensées, de médias et d'informations à travers la création de réseaux sociaux et de communautés virtuels [123]. Reddit, YouTube, LinkedIn et Facebook sont des exemples réussis de ces systèmes.

- **Les jeux en ligne multijoueurs et les mondes virtuels** sont des environnements simulés par ordinateur qui sont peuplés par de nombreux utilisateurs qui peuvent créer un avatar personnel, et explorer simultanément et indépendamment le monde virtuel, participer à ses activités et communiquer avec les autres. Les actions et contributions d'un utilisateur affectent directement et indirectement les perceptions et actions des autres utilisateurs [88].

- **Les systèmes de crowdsourcing** sont basés sur des appels ouverts pour l'accomplissement volontaire de tâches. Dans les systèmes de

crowdsourcing, les humains choisissent activement et contribuent aux tâches, mais en général, les agents humains n'interagissent pas directement entre eux. La revue par les pairs académique, Amazon Mechanical Turk, Kaggle et le défi réseau DARPA de 2009 [99] sont des exemples de systèmes de crowdsourcing.

Les types de machines sociales mentionnés diffèrent par la structure et l'intensité des interactions entre les humains et les machines. Cependant, il n'existe pas de frontière précise entre ces groupes, et certains produits peuvent appartenir à plusieurs catégories. Par exemple, Facebook peut être classé dans deux catégories : marché en ligne et médias sociaux.

3.2 Équité dans les machines sociales

Avec l'émergence des machines sociales, l'étude de l'équité doit être élargie pour englober un cadre plus large. Les machines sociales, qui incluent des réseaux de personnes interagissant via des infrastructures numériques, intègrent souvent divers composants automatisés ou semi-automatisés. Ces systèmes ne sont pas uniquement algorithmiques ; ils peuvent également comprendre un mélange de processus algorithmiques et d'interventions humaines, voire fonctionner principalement par le biais d'interactions sociales sans algorithmes informatiques spécifiques. Cependant, les interactions sociales doivent être principalement médiées par l'infrastructure informatique ; sinon, elles ne relèvent pas de notre recherche.

Les plateformes sociales telles que TaskRabbit et Fiverr ont fait l'objet d'audits algorithmiques [57], révélant une association notable entre la perception du genre et de la race, et les évaluations et classements des travailleurs. Plus précisément, les travailleurs noirs reçoivent souvent des évaluations inférieures par rapport à leurs homologues blancs également qualifiés. Cependant, ces investigations n'approfondissent pas explicitement les mesures d'équité spécifiques. La revue systématique de Bandy et al. [8] discute d'audits similaires à travers les plateformes sociales, soulignant la prévalence croissante et l'importance des audits dans les machines sociales. Malgré leur exhaustivité, ces revues ne se réfèrent pas spécifiquement à des métriques telles que les Equalized Odds ou la Predictive Parity.

Blodgett et al. [16] dans leur enquête critique, examinent la nature multifacette du biais dans les systèmes de Traitement automatique des langues (NLP). Ils soulignent les motivations souvent ambiguës et incohérentes derrière l'identification et la mitigation de ces biais, ainsi que l'absence de mesures d'équité. Ce travail est pertinent pour notre enquête sur l'équité dans les machines sociales, car il met en lumière des lacunes critiques dans la compréhension actuelle et l'atténuation du biais, des préjudices connexes et des groupes affectés. Il fournit une base pour étendre les considérations d'équité au-delà des systèmes purement algorithmiques vers ceux impliquant une interaction humaine significative.

Selbst et al. [103] ont exploré les défis et les méthodologies dans la conception des systèmes socio-techniques qui équilibrent efficacement les proces-

sus algorithmiques et l'intervention humaine. Leur travail souligne l'importance de contextualiser l'équité dans ces systèmes, reconnaissant que les biais et l'équité ne peuvent être pleinement compris ou abordés sans tenir compte des contextes sociaux complexes dans lesquels ces technologies opèrent.

L'analyse de Blodgett et al. [16] souligne la nécessité d'une articulation plus claire de ce qui constitue un biais et du raisonnement normatif qui sous-tend cette compréhension. Cela s'aligne sur notre objectif d'intégrer les mesures d'équité issues de l'apprentissage automatique dans des cadres plus larges de machines sociales, où les biais peuvent provenir non seulement des résultats algorithmiques mais aussi des processus influencés par les humains. En suivant leur recommandation d'ancrer l'analyse des biais dans les hiérarchies sociales, notre cadre proposé vise à évaluer systématiquement et à atténuer ces biais, assurant des résultats plus équitables à travers diverses machines sociales.

De plus, Blodgett et al. [16] appellent à un engagement plus profond avec les communautés affectées et les dynamiques de pouvoir entre les technologues et ces communautés. Cette perspective est cruciale pour notre approche, qui vise à développer des stratégies qui promeuvent l'équité dans les machines sociales en prenant en compte à la fois les modèles computationnels et les éléments humains qui interagissent avec ces systèmes. Notre cadre intégrera ces perspectives pour mieux aborder les nuances de l'équité dans le contexte des machines sociales, contribuant ainsi à la création de technologies plus équitables qui reconnaissent et atténuent les sources complexes de biais

identifiées dans leur enquête.

3.3 Résumé

Dans ce chapitre, nous avons examiné différentes catégories de machines sociales et abordé la question cruciale de l'équité, en soulignant la nécessité d'étendre les cadres traditionnels d'équité pour tenir compte des complexités des interactions socialement médiatisées. À travers des études clés sur les audits algorithmiques et les biais dans le traitement automatique des langues, nous avons identifié des lacunes dans les approches actuelles de l'équité, notamment la nécessité de prendre en compte les biais dans des contextes sociaux plus larges. Cette enquête sur l'état de l'art en matière d'équité pour les machines sociales prépare le terrain pour la revue systématique présentée dans le prochain chapitre. Cette revue élargira davantage le discours actuel en offrant une analyse complète de l'équité à travers divers cadres de machines sociales, poursuivant notre examen des méthodologies de pointe et proposant de nouvelles orientations pour la recherche future.

Chapitre 4

Revue systématique

4.1 Méthodologie

Afin d'avoir une vue d'ensemble des problèmes d'équité et de biais dans les machines sociales, nous avons d'abord sélectionné un échantillon représentatif de machines sociales, puis nous avons collecté des articles décrivant les problèmes de biais affectant ces systèmes.

4.2 Revue systématique

Compte tenu de notre focus sur les comportements problématiques et l'exploration de l'équité dans les machines sociales, nous avons établi trois critères clés pour la sélection de nos études de cas. Premièrement, nous avons choisi des études de cas couvrant toutes les catégories de machines

sociales. Deuxièmement, au sein de chaque catégorie, nous avons sélectionné des machines sociales offrant une richesse de données de recherche. De plus, nous avons observé que des interactions d'utilisateurs notables mettent souvent en évidence la présence de comportements problématiques. Ainsi, dans des catégories telles que les marchés en ligne, caractérisées par des modèles d'interaction variés, nous avons élargi notre sélection pour inclure plusieurs exemples. Cette approche garantit une couverture complète et nous permet de capturer une gamme diversifiée d'interactions, mettant en lumière l'impact potentiellement destructeur des comportements problématiques. La liste des études de cas dans chaque catégorie est fournie dans le Tableau 4.1.

Nos études de cas dans le marché en ligne représentent une large gamme de ce type. TaskRabbit sert de plateforme pour la réalisation de tâches à court terme, pouvant potentiellement impliquer des interactions réelles en face à face entre le travailleur et le client. En revanche, Fiverr facilite la prestation de services qui ne nécessitent pas de présence physique ou de contact direct et ne sont pas contraints par des limitations géographiques. Pour approfondir et comprendre les dynamiques des machines sociales, notre investigation s'étend à deux exemples de premier plan : Uber, une plateforme de covoiturage leader, et Airbnb, un marché d'hébergement renommé. En scrutant diverses plateformes de machines sociales, nous obtenons des perspectives sur les comportements problématiques dans différents environnements. Les résultats de notre recherche ont le potentiel d'être largement généralisés à travers un spectre de machines sociales.

Le Tableau 4.1 ne spécifie pas d'études de cas individuelles pour les jeux en ligne, car la plupart des études dans cette catégorie reposent sur des interviews et des enquêtes avec des joueurs masculins et féminins sur leurs expériences. De plus, d'autres recherches dans ce domaine couvrent un large éventail de jeux. Par exemple, les études de Rennick et al. [97] se basent sur un grand corpus de dialogues recueillis à partir de divers jeux, et Williams et al. [128] ont examiné 150 jeux dans son travail. Il est à noter que le jeu World of Warcraft a été l'objet de deux études [24, 19], et Gao et al. [47] ont spécifiquement examiné le jeu League of Legends.

Afin de répondre à **(QR1)** and **(QR2)**, nos investigations englobent cinq types distincts de machines sociales. Nous avons mené une revue systématique des articles rapportant des *biais* (et d'autres concepts liés à l'équité) dans les machines sociales sélectionnées. Pour identifier les articles pertinents, nous avons recherché dans les bases de données Scopus et Google Scholar en utilisant les mots-clés "bias", "fairness", "discrimination", "fair" et "unfair" en conjonction avec le nom de la machine sociale cible (par exemple, Wikipedia) ou la catégorie cible (par exemple, jeux en ligne), et avons collecté des articles publiés jusqu'en décembre 2023. Dans Scopus, nous avons limité la portée de la recherche aux titres, résumés et introductions. De plus, nous avons également collecté des revues plus générales de la recherche académique pour les études de cas – en utilisant les mots-clés "review" ou "survey" et "le nom des études de cas" dans Google Scholar – et recherché dans ces revues d'autres articles et termes pertinents pour décrire les problèmes liés à l'équité. Les

termes “disparity” et “inequality” ont été identifiés de cette manière. Nous avons obtenu un total de 622 articles, avec le nombre d’articles dans chaque catégorie détaillé dans le Tableau 4.1.

Tableau 4.1. *Études de cas dans chaque catégorie de machines sociales et répartition des articles par catégorie*

Machine sociale	Études de cas	Répartition des articles collectés	articles liés à l’équité de groupe
Réseaux de collaboration massive	Wikipédia	213	86
Marché en ligne	TaskRabbit, Fiverr, Uber et Airbnb	71	26
Médias sociaux	Reddit, YouTube	183	31
Jeux en ligne multijoueurs et mondes virtuels	jeux en ligne	29	12
Systèmes de crowdsourcing	Révision par les pairs académique	126	41

Le mot-clé lié à l’équité le plus courant dans les articles d’enquête étudiés est “biais”, Nous avons éliminé les articles où ce terme ne faisait pas référence à un manque d’équité envers des groupes spécifiques. En particulier, nous avons écarté les articles qui se référaient à des biais statistiques, cognitifs, ou à une violation d’une politique au sein de la machine sociale concernée, telle que la NPOV (Neutral Point of View) politique de Wikipédia. Parmi ces articles, nous avons retenu ceux qui décrivaient un biais envers certaines personnes ou groupes sociaux, ou ceux où les points de vue de ces groupes étaient délibérément ignorés. Nous avons éliminé les articles qui discutaient de “texte biaisé” sans aucune référence aux personnes ou groupes

spécifiques affectés par ces biais, par exemple, les articles présentant des techniques pour identifier ce type de "texte biaisé".

Nous avons retenu 196 articles, avec le nombre d'articles dans chaque catégorie détaillé dans le tableau 4.1. La liste des articles étudiés est fournie en annexe pour référence.

Pour chaque article, nous avons identifié les phénomènes décrits comme problématiques et les groupes sociaux qu'ils affectent. Nous avons ensuite catégorisé ces phénomènes et les avons reliés à différents types de préjudices, en nous référant à une taxonomie établie des préjudices sociotechniques [108].

4.3 Problèmes d'équité

4.3.1 Problèmes d'équité dans les réseaux de collaboration massive

Dans cette section, nous commençons à répondre à **QR1** en étudiant Wikipédia comme un réseau emblématique de collaboration massive. Wikipédia a des processus d'interaction complexes, régis par de nombreuses règles et normes [66], avec de nombreux composants automatisés impliqués (y compris des milliers de bots [138, 53] responsables de 20 % des modifications [71]), mais sans décisions automatisées majeures qui affectent directement les personnes comme dans le modèle habituel de l'équité dans les systèmes

algorithmiques. Wikipédia étant une plateforme bien étudiée de réseaux de collaboration massive, de nombreux types de “biais” pertinents à notre problème ont déjà été étudiés sur Wikipédia. Nous listons et catégorisons les phénomènes décrits comme des biais dans la littérature sur Wikipédia.

4.3.1.1 Inégalité de participation

En examinant la dynamique de la participation sur Wikipédia, des différences importantes apparaissent selon le *sexe* et la *localisation géographique*.

Dans de nombreux articles étudiés, les termes *gender bias / gender gap* se réfèrent à la disparité entre le nombre d'hommes et de femmes participant à Wikipédia, en tant que lecteurs, rédacteurs et éditeurs [94, 132, 54, 87, 83, 134]. Selon des enquêtes de la Wikimedia Foundation et d'autres analyses, en 2013, la proportion de femmes participant à Wikipédia se situait entre 13 % et 22 % [136]. Nous notons qu'il y a rarement de discussion sur d'autres identités de genre (par exemple, transgenre ou non-binaire) dans ces analyses.

En plus du biais de genre, le *biais géographique* est un autre phénomène significatif observé dans la participation à Wikipédia. Alors que le biais de genre concerne la sous-représentation des femmes, le biais géographique met en évidence les disparités dans la répartition des contributeurs selon les pays. Un rapport de la Wikimedia Foundation [55] note que 45% des contributeurs de Wikipédia proviennent de seulement cinq pays (Italie, France, Grande-Bretagne, Allemagne et États-Unis). Si une seule édition linguistique de Wikipédia est considérée, une situation similaire se produira : les contribu-

teurs de cette édition seront principalement originaires du pays ou des pays où la langue est parlée. De plus, la composition démographique des contributeurs de Wikipédia dans un pays donné peut ne pas refléter la distribution démographique globale du pays, ce qui signifie qu'il existe potentiellement d'autres manières de définir des groupes sous-représentés (par exemple, selon les lignes raciales).

4.3.1.2 Nombre de biographies

L'un des principaux phénomènes de biais rapportés sur Wikipédia est la distribution des biographies, toujours en ce qui concerne le genre et la localisation géographique. De nombreux articles comparent le nombre de biographies d'hommes et de femmes [126, 121, 136, 134, 129, 133] (les autres identités de genre ne sont pas étudiées ici non plus) et constatent que les femmes sont nettement sous-représentées : en juillet 2019, seulement 17,9 % des biographies sur Wikipédia étaient des biographies de femmes [121]. Cependant, à moins de considérer que les hommes et les femmes sont également susceptibles d'être *notables*, ce chiffre est difficile à interpréter. Wagner et al. [126] étudient le phénomène en utilisant des bases de données externes de personnes "notable" comme références pour savoir qui *devrait* avoir une biographie sur Wikipédia : dans ce sens, ils constatent que les femmes ne sont pas sous-représentées, dans le sens où un homme notable et une femme notable ont des chances similaires d'avoir une biographie sur Wikipédia. Dans une autre étude, Wagner et al. [127] trouvent que les femmes sur Wikipédia

sont généralement plus notables que leurs homologues masculins.

D'autre part, Adams et al. [1] étudient les biographies de sociologues américains sur Wikipédia, et en utilisant des définitions bibliométriques de la notabilité, ils trouvent que pour des niveaux de notabilité similaires, les hommes sont au moins deux fois plus susceptibles d'avoir une biographie sur Wikipédia. En d'autres termes, il y a une claire sous-représentation des femmes par rapport au ratio approximatif de femmes par rapport aux hommes dans la population mondiale en général, mais les résultats sont moins concluants lorsque nous tentons de contrôler la notion (mal définie) de "notabilité". Cela suggère que le biais n'est pas dans le processus d'attribution des pages Wikipédia aux personnes notables, mais plutôt dans le processus de devenir "notable" en premier lieu.

En plus du biais de genre, un phénomène parallèle émerge sous la forme du biais géographique. [12] souligne que 62 % des biographies sur Wikipédia concernent également des personnes provenant des cinq mêmes pays, mentionnés dans la section 4.3.1.1. En considérant les groupes sociaux définis par la race aux États-Unis, [1] trouve que les chercheurs non blancs sont sous-représentés sur Wikipédia, par rapport à leurs homologues blancs de notabilité similaire (mesurée par des indicateurs bibliométriques).

Suppression de biographies Un phénomène connexe mais distinct est le taux de *suppression* disparate des biographies d'hommes et de femmes : pour l'un des processus de suppression de pages de Wikipédia (Article for

deletion, AFD), une étude a révélé que 25 % des biographies supprimées étaient celles de femmes [121]. Bien que 25 % soit bien inférieur à 50 %, la comparaison pertinente ici est avec la proportion des biographies de femmes existantes et susceptibles d'être supprimées (c'est-à-dire 17,9 %). Selon cette analyse, les femmes sont donc sur-représentées dans le processus de suppression de biographies, par rapport au nombre total de biographies de femmes sur Wikipédia.

4.3.1.3 Liens dans les biographies

Dans quelques articles, une disparité est signalée concernant le nombre de liens entre les biographies de personnes de différents genres [126, 136] : selon cette analyse, il y a statistiquement plus de liens des biographies de femmes vers celles des hommes que de liens des biographies d'hommes vers celles des femmes, même en tenant compte des tailles des groupes. Ce type de biais est appelé biais structurel dans certains articles.

4.3.1.4 Biais lexical dans les biographies

Un autre type de biais rapporté sur Wikipédia se trouve dans le contenu lexical des articles, en particulier dans le texte des biographies [126, 94, 132, 136, 95, 81, 79, 64, 129]. Ce type de biais est appelé biais lexical dans certains articles [126]. e biais lexical est visible lorsque le contenu des biographies des hommes et des femmes diffère selon des schémas stéréotypés : les

biographies des femmes contiennent plus d'informations liées à leur famille, leur conjoint et leurs enfants, tandis que les biographies des hommes contiennent plus d'informations sur eux-mêmes, leurs intérêts, leurs activités et leurs réalisations [126, 136, 114].

4.3.1.5 Catégorisation des pages pour les biographies

Un biais dans le système de catégorisation des pages de Wikipédia est décrit par Yanisky et al. [132]: dans certaines catégories de biographies, un sous-groupe particulier (défini par une valeur d'attribut démographique) est traité comme une exception au sein du groupe "par défaut", constitué par le groupe majoritaire. Dans le cas décrit par Yanisky et al. [132], la catégorie des "American novelists" contient des romanciers masculins, et les romancières sont placées dans la sous-catégorie de "American female novelists". Bien que ce cas particulier ait ensuite été résolu, nous avons observé que des catégorisations similaires se produisent pour d'autres professions et attributs démographiques, dans plusieurs éditions linguistiques de Wikipédia. Nous notons que Young et al. [134] ne considèrent pas ce *biais de catégorisation* comme problématique, car il pourrait augmenter l'attention sur les groupes sous-représentés.

4.3.1.6 Sous-représentation des intérêts des femmes

Un petit nombre d'études ont exploré si les sujets d'intérêt pour les femmes sont sous-représentés sur Wikipédia [94, 132, 129]. La question

est de savoir si Wikipédia contient plus d'articles sur des sujets d'intérêt pour les hommes que d'articles sur des sujets d'intérêt pour les femmes. Différentes expériences ont attribué des sujets aux hommes et aux femmes soit en utilisant des associations stéréotypées "bien connues" soit en extrayant ces sujets à partir de corpus de magazines destinés à un public masculin/féminin. L'analyse de Worku et al. [129] montre également que les articles présumés d'intérêt pour les femmes étaient nominés pour suppression à des taux légèrement plus élevés que les articles présumés d'intérêt pour les hommes.

4.3.1.7 Biais culturel

De manière similaire au phénomène précédent de sous-représentation des intérêts des femmes, il a été rapporté que les perspectives de certains groupes culturels sont également sous-représentées dans les articles de Wikipédia. Par exemple, [82] évoquent la *colonization of Wikipedia*, c'est-à-dire la domination des perspectives occidentales, même sur des sujets principalement associés à d'autres cultures (par exemple, la médecine traditionnelle chinoise). Ils citent l'exemple d'un article décrivant une plante utilisée dans la médecine traditionnelle chinoise, où toutes les mentions de ses utilisations médicinales traditionnelles ont été supprimées, ainsi que toutes les références à l'exception d'un article scientifique occidental.

Plusieurs auteurs ont également étudié les *linguistic points of view* [4, 21, 100, 75], c'est-à-dire les différences dans la manière dont un sujet est décrit

dans les différentes éditions linguistiques de Wikipédia. Un exemple frappant est la manière dont les conflits internationaux historiques sont décrits par les groupes culturels représentant les deux parties du conflit, ces deux groupes éditant principalement les différentes éditions linguistiques de Wikipédia [4].

4.3.1.8 Surveillance excessive des utilisateurs anonymes

En plus des groupes démographiques précédemment discutés, pour lesquels les biais sont couramment discutés dans le contexte de l'équité algorithmique, un petit nombre d'articles [119, 33, 32] ont discuté du risque de biais à l'encontre du groupe social des *anonymous editors*, c'est-à-dire les éditeurs de Wikipédia qui ne se sont pas inscrits pour un compte. Dans ce cas, la forme de biais qui affecte potentiellement les utilisateurs est ce que nous pourrions appeler une *surveillance excessive*. Wikipédia permet à la fois aux contributeurs inscrits et anonymes, mais les contributions des éditeurs anonymes sont plus étroitement surveillées que celles des utilisateurs enregistrés, et sont plus susceptibles d'être révoquées [119]. Cela est dû à un logiciel de "signalement" semi-automatisé, qui utilise l'apprentissage automatique pour détecter les modifications potentiellement problématiques et les signaler aux éditeurs chargés de la "surveillance". L'interface du système indique à la fois un score de risque (calculé par l'outil) et le statut d'inscription de l'éditeur. de Laat et al. [33] suggère que le logiciel de signalement peut ne pas être biaisé, mais le fait d'afficher le statut d'inscription amène les éditeurs

surveillants à se concentrer sur les utilisateurs anonymes, les “profilant” ainsi.

4.3.2 Problèmes d’équité dans les marchés en ligne

Poursuivant notre exploration de **(QR1)**, cette section examine les phénomènes de biais dans quatre marchés en ligne distincts : TaskRabbit, Fiverr, Airbnb et Uber. Nous collectons et catégorisons les biais spécifiques affectant différents groupes démographiques au sein de ces plateformes.

4.3.2.1 Nombre d’avis

Ce phénomène exprime comment le genre et la race perçus sont liés au nombre d’avis que les travailleurs reçoivent dans les marchés en ligne [116, 117, 120, 80]. Yanisky et al. [57] étudient le phénomène sur les données de TaskRabbit en utilisant un modèle de régression binomiale négative. Ils explorent divers facteurs influençant la quantité d’avis, y compris la date d’inscription, l’activité récente, et les tâches accomplies. Notamment, les résultats révèlent que la perception d’être une femme est liée à un nombre d’avis inférieur, les femmes blanches recevant 10 % d’avis en moins que leurs homologues masculins. Dans toutes les catégories raciales, les femmes, en moyenne, reçoivent moins d’avis que les hommes.

En ce qui concerne la race perçue, l’étude ne trouve pas de différences significatives entre les travailleurs noirs et blancs Hannák et al. [57]. Cependant, les travailleurs perçus comme des hommes asiatiques reçoivent 13 % d’avis en moins que ceux perçus comme des hommes blancs. Il est impor-

tant de noter que, bien qu'un grand nombre d'avis puisse suggérer un embauche fréquente et sembler être souhaitable, le contenu de ces avis n'est pas nécessairement positif. La section 4.3.2.4 passe en revue les biais exprimés dans le sentiment des avis.

TLa même étude révèle également des schémas différents dans le retour social sur Fiverr, comparativement à TaskRabbit. Une proportion substantielle des utilisateurs de Fiverr n'ont pas de photo de profil ou utilisent des images non humaines, souvent des publicités liées aux tâches. Ces choix peuvent affecter la perception des clients, étant donné la nature virtuelle des tâches sur Fiverr, où les interactions en personne n'ont pas lieu.

Hannák et al. [57] révèlent que l'absence de photo de profil a une forte corrélation négative avec le nombre d'avis, tandis que l'utilisation d'images non humaines montre une corrélation positive. En ce qui concerne le genre et la race, ils ont effectué deux tests différents et ont observé que les travailleurs perçus comme noirs reçoivent significativement moins d'avis que ceux perçus comme blancs : les travailleurs noirs ont été trouvés pour recevoir 38 % d'avis en moins que les travailleurs blancs dans un test, et 32 % d'avis en moins dans l'autre.

4.3.2.2 Biais de notation

En plus des avis, TaskRabbit dispose également d'un système d'évaluation, où d'autres différences peuvent être mesurées. Hannák et al. [57] trouvent une association significative entre la perception d'être noir et des scores

d'évaluation plus bas, mais pas de différences substantielles entre les genres en général, sauf dans le cas des travailleurs noirs, où les hommes noirs reçoivent des évaluations inférieures par rapport aux femmes noires.

Les articles examinés [110, 131, 57] montrent que le biais racial est évident dans les évaluations sur Fiverr, où être perçu comme noir ou asiatique est associé à des évaluations significativement plus basses par rapport à ceux perçus comme blancs. Étonnamment, les femmes blanches reçoivent de meilleures évaluations, tandis que les travailleurs non blancs et masculins reçoivent des évaluations plus basses. Selon Hannák et al. [57], les données suggèrent un schéma nuancé, où les femmes perçues dans des catégories spécifiques, telles que "Bases de données" and "Analyse Web," reçoivent des évaluations notablement plus élevées que les hommes.

4.3.2.3 Biais dans le classement de recherche

La race ou le genre perçu des travailleurs influence leurs classements dans les résultats de recherche sur TaskRabbit. Les clients s'appuient fortement sur les moteurs de recherche de ces plateformes pour trouver des travailleurs adaptés. Même si les variables démographiques ne sont pas explicitement prises en compte, des études suggèrent que les classements peuvent toujours être implicitement influencés par des variables telles que les avis et les évaluations, qui ont montré des corrélations avec les caractéristiques démographiques perçues.

Pour TaskRabbit, des facteurs tels que le nombre de tâches accomplies,

la durée d'adhésion, et l'activité récente montrent des corrélations positives avec des classements plus élevés [57]. En ce qui concerne la race, les travailleurs perçus comme noirs tendent à avoir des classements plus bas par rapport à ceux perçus comme blancs, tandis que les travailleurs perçus comme asiatiques ont tendance à avoir des classements significativement plus élevés.

4.3.2.4 Sentiment dans les avis

Hannák et al. [57] ont également mené une analyse du sentiment dans les avis sur Fiverr, déterminant les ratios de probabilité des adjectifs positifs et négatifs en fonction du genre et de la race perçus du travailleur évalué. Ils ont observé des biais de genre et de race significatifs sur Fiverr. Pour mieux interpréter les effets, ils ont regroupé les travailleurs en six combinaisons de genre et de race perçus, y compris homme blanc, homme noir, homme asiatique, femme blanche, femme noire, et femme asiatique. Un biais de contenu a été observé, en particulier pour les femmes noires, qui étaient moins susceptibles d'être décrites avec des adjectifs positifs. L'utilisation d'adjectifs comme mots négatifs était la plus prononcée et significative pour les travailleurs perçus comme noirs, quel que soit le genre.

4.3.2.5 Disparité dans la qualité du service

Ce phénomène se manifeste sous diverses formes sur différentes plateformes. Les membres de certains groupes connaissent des taux d'acceptation des invités plus faibles sur Airbnb, des temps d'attente plus longs pour les

passagers et un plus grand nombre d'annulations de trajets sur Uber.

Les faibles taux d'acceptation des invités sur les plateformes de location comme Airbnb révèlent des tendances préoccupantes, notamment en ce qui concerne les disparités raciales et de genre. La recherche menée par Edelman et al. [35] et Li et al. [70] indique que les invités afro-américains connaissent des taux d'acceptation significativement plus bas par rapport à leurs homologues blancs. Plus précisément, les invités afro-américains avaient un taux d'acceptation de 29 %, nettement inférieur au taux d'acceptation de 49 % pour les autres. Lorsque les invités afro-américains avaient des avis positifs, leur taux d'acceptation s'améliorait pour atteindre 56 % (comparé à 58 % pour les autres). De plus, les invités portant des noms à consonance afro-américaine ont reçu une réponse positive à 42 %, contre 50 % pour les autres.

Ces résultats sont en accord avec des incidents couverts par les médias, tels que le cas de Gregory Selden, qui a fait l'objet de discrimination raciale sur Airbnb et a poursuivi la société en justice pour cela [69].

Une autre dimension de ce phénomène est révélée dans une étude [2] explorant la discrimination à l'égard des personnes en couple homosexuel (same-sex relationships ; SSRs) sur Airbnb à Dublin. La recherche a montré que les invités en SSR masculin avaient 20 % à 30 % moins de chances d'être acceptés, notant que cette différence découle principalement du fait que les hôtes ne répondent pas plutôt que d'un rejet explicite.

Un autre phénomène de biais est révélé par les disparités de prix

sur Airbnb et d'autres plateformes de location entre particuliers. Plusieurs études [36, 35, 70, 72, 65] ont montré que les hôtes blancs fixent des prix plus élevés que les hôtes non blancs. Les hôtes afro-américains, en particulier, tendent à demander et à recevoir des prix plus bas que les hôtes blancs : une étude [124] a révélé que les hôtes non noirs parvenaient à obtenir un prix 12 % plus élevé que leurs homologues noirs.

Un autre phénomène observable est le temps d'attente plus long que connaissent les passagers Uber appartenant à certains groupes [98, 51], ce qui conduit à une expérience moins favorable. Des études aux États-Unis montrent que les passagers afro-américains rencontrent fréquemment des temps d'attente prolongés pour que leurs demandes de trajet soient acceptées, révélant une manifestation de discrimination raciale dans le contexte du covoiturage. Ge et al. [51] affirment que les temps d'attente sont 30 % plus longs pour les passagers afro-américains qui utilisent Uber.

Les défis de la discrimination chez Uber ne se limitent pas aux biais raciaux, mais incluent également les problèmes d'accessibilité rencontrés par les passagers handicapés ou accompagnés d'enfants. Bien qu'aucune statistique sur les temps d'attente ou la disponibilité des trajets n'ait été publiée, il existe des preuves anecdotiques que la qualité du service est inférieure pour ces passagers, y compris un cas juridique contre Uber intenté par un passager handicapé [124, 69].

En plus des temps d'attente, la recherche de Ge et al. [51, 52] montre également que les passagers afro-américains sont plus susceptibles de subir

des annulations de trajets par rapport aux autres groupes raciaux. Cette pratique discriminatoire ne cause pas seulement des désagréments pour les passagers concernés en perturbant leurs plans de voyage, mais soulève également des questions sur l'équité et l'inclusivité des services d'Uber.

4.3.3 Problèmes d'équité sur les réseaux sociaux

Les plateformes de médias sociaux comme Reddit ne sont pas seulement des miroirs du monde réel, mais elles sont aussi devenues des espaces importants où différents groupes et idées façonnent notre société moderne. Reconnaissant leur importance, les chercheurs ont mené des études pour examiner les biais potentiels dans divers aspects au sein de ces plateformes.

En explorant la représentation biaisée sur les plateformes de médias sociaux comme Reddit, un aspect important réside dans la représentation biaisée de certains groupes démographiques par rapport à d'autres. Des études ont mis en évidence une représentation biaisée des femmes et d'autres groupes protégés.

4.3.3.1 Biais linguistique et discours toxique

Comme discuté dans la section 4.3.1.4, le biais lexical se rapporte au contenu du texte, qui, dans le contexte des plateformes de médias sociaux comme Reddit, est évident dans les commentaires. Ferrer et al. [40] explorent ce phénomène, le qualifiant de biais linguistique, et examinent les biais dans les commentaires sur Reddit concernant le genre, l'ethnicité et la religion. De

même, Banerjee et al. [9] étudient le biais de genre sur Reddit et Fitocracy, soulignant comment les commentaires sont biaisés en fonction du genre. Ils démontrent, à travers une analyse de sentiment, que les textes contenant un plus grand nombre de phrases négatives sont plus susceptibles de montrer un biais de genre. Dans une autre étude, Farrell et al. [38] analysent le langage misogyne dans 300,000 conversations sur Reddit. Ils étudient la propagation des idées misogynes au sein et entre sept communautés en ligne, montrant des tendances croissantes de misogynie, d'hostilité et d'attitudes violentes, corroborant ainsi les théories féministes sur la prévalence croissante de tels contenus. De plus, Marjanovic et al. [73] examinent le biais lexical dirigé contre les femmes politiques, révélant que "les femmes politiques sont souvent appelées par leur prénom et sont décrites en relation avec leur corps, leurs vêtements ou leur famille; ce traitement n'est pas étendu de manière similaire aux hommes."

En outre, Almerekhi et al. [3] et Xia et al. [130] abordent les problèmes de harcèlement, de discours haineux et de toxicité dans les commentaires sur Reddit, proposant des méthodes pour identifier les commentaires toxiques. Bien qu'ils n'examinent pas des groupes spécifiques, ils utilisent des techniques d'apprentissage automatique pour détecter et analyser les schémas de discours haineux.

4.3.3.2 Biais de genre dans le système de recommandation de YouTube

Le biais de genre dans le système de recommandation de YouTube se produit lorsque l'algorithme de la plateforme favorise le contenu créé par un genre plutôt qu'un autre, reflétant souvent les biais sociétaux. Cela peut conduire à une sous-représentation ou à une mauvaise représentation du contenu des créateurs d'un certain genre. Dans de nombreux cas, les vidéos créées par des femmes reçoivent moins de recommandations par rapport à celles des hommes, limitant ainsi leur visibilité et leur portée. Ce biais peut être perpétué par les interactions des utilisateurs, où le contenu de certains genres est plus aimé ou partagé, renforçant ainsi la tendance de l'algorithme à favoriser ce contenu. En conséquence, le système de recommandation peut finir par amplifier les disparités de genre existantes, rendant plus difficile pour le contenu des genres sous-représentés de gagner en visibilité [102].

4.3.4 Problèmes d'équité dans les jeux en ligne multi-joueurs

4.3.4.1 Inégalité de participation

La recherche menée par Williams et al. [128] indique une disparité dans les démographies des joueurs aux États-Unis, avec un nombre plus élevé de joueurs masculins par rapport aux joueuses, et une plus grande représentation

des joueurs blancs par rapport aux Hispaniques et aux Amérindiens. L'inégalité de participation dans les jeux en ligne est influencée par divers facteurs, y compris la conception du jeu et les normes sociales. Comme le souligne Norris et al. [86], certaines caractéristiques agressives inhérentes à certains jeux vidéo tendent à plaire davantage aux individus ayant des traits de personnalité similaires. Nous pensons que cette préférence peut contribuer à la sous-représentation des femmes dans les communautés de joueurs, car les normes sociales découragent souvent les femmes d'adopter des comportements agressifs.

4.3.4.2 Représentation des genres dans les personnages de jeux

La représentation des femmes en tant que personnages dans les jeux en ligne a attiré l'attention des chercheurs, en particulier en ce qui concerne le nombre de personnages, le type de personnage et les dialogues des personnages. La recherche menée par Ivory et al. [62] met en évidence une prévalence des personnages masculins surpassant en nombre les personnages féminins dans les contextes de jeux. En outre, Brehm et al. [19] souligne un biais dans la représentation des personnages dans des jeux comme World of Warcraft, où les rôles de personnages négatifs, tels que les sorciers, sont principalement attribués aux personnages féminins. De plus, Rennick et al. [97] suggèrent une disparité dans les dialogues, les personnages masculins ayant plus de répliques que leurs homologues féminins. Ces conclusions collectives soulignent les biais systémiques perpétuant une représentation inégale des

femmes dans les jeux en ligne.

4.3.5 Problèmes d'équité dans les systèmes de crowdsourcing

Dans cette section, étant donné la richesse des recherches disponibles, nous nous concentrons sur le processus de révision par les pairs académique. Ci-dessous, nous décrivons les phénomènes de biais mis en évidence dans les articles examinés.

4.3.5.1 Biais de genre dans la sélection des évaluateurs

Une manifestation du biais de genre émerge dans le choix des évaluateurs [14, 112]: les éditeurs associés (EAs) qu'ils soient hommes ou femmes, ont nommé plus d'évaluateurs masculins que féminins, mais les EAs féminins étaient significativement plus susceptibles de nommer des évaluatrices par rapport aux EAs masculins [14, 112].

4.3.5.2 Taux d'acceptation des articles

Dans le processus de révision par les pairs, les auteurs peuvent faire l'objet d'un biais en fonction des groupes auxquels ils appartiennent.

Biais de genre Concernant le genre des auteurs, Severin et al. [104] ont montré que les auteurs masculins recevaient plus d'attention de la part des évaluateurs, et Biolkova et al. [14] ont fourni des preuves suggérant que les

manuscrits écrits par des femmes et évalués par des pairs féminins ont tendance à recevoir des évaluations plus positives par rapport à ceux évalués par des pairs masculins. Cela suggère un biais dans les évaluations des examinateurs influencé par le genre de l'auteur.

Biais d'affiliation ou de prestige La recherche indique que le processus de révision par les pairs est susceptible d'être biaisé en fonction de la réputation et de l'affiliation institutionnelle des auteurs, que nous appelons biais d'affiliation ou de prestige [109, 115, 77, 84]: la qualité des articles provenant de chercheurs renommés ou d'institutions prestigieuses tend à être surévaluée par rapport aux articles d'autres chercheurs ou institutions.

En outre, un biais similaire s'applique aux articles associés aux institutions hôtes des revues, comme en témoigne le fait que les articles acceptés de ces institutions reçoivent ensuite en moyenne moins de citations que ceux rédigés par des chercheurs d'institutions non affiliées [109]. Idéalement, dans un processus de révision par les pairs équitable, les articles de différentes institutions devraient avoir la même probabilité de recevoir des citations ou des évaluations positives s'ils sont de qualité ou d'impact prédictif similaire.

Ces résultats sont corroborés par d'autres études qui explorent les effets du passage à la révision par les pairs en double aveugle dans une conférence de science informatique [115, 44]: leur analyse révèle une diminution notable des scores attribués aux auteurs prestigieux après la transition vers la révision en double aveugle, indiquant un biais de prestige dans le mode de révision

en simple aveugle, qui disparaît avec la révision en double aveugle.

4.4 Groupes sociaux affectés par les biais

Les problèmes d'équité pour les groupes concernent les préjudices systémiques dirigés vers un groupe défini par un attribut démographique protégé. Par conséquent, la première étape pour aborder les biais est d'identifier ces groupes. Pour répondre à **(QR2)**, nous explorons quels groupes démographiques sont susceptibles de subir des biais et examinons comment ces biais se manifestent.

Nous catégorisons les groupes sociaux affectés par les biais dans les machines sociales en deux catégories distinctes : les groupes formés sur la base d'attributs sensibles (et protégés par la loi dans de nombreuses juridictions), présents dans divers types de machines sociales, et les groupes uniques à certaines machines sociales, significatifs uniquement dans le contexte de ces machines sociales (attributs indiqués par * dans le Tableau 4.2)

4.4.1 Attributs sensibles

L'attribut démographique le plus fréquemment cité dans les études sur les biais est le *genre*: de nombreux articles décrivent le "gender gap", les "gender biases" ou les "gender disparities". Ces articles décrivent plusieurs phénomènes où les femmes sont désavantagées par rapport aux hommes. Il est à noter que nous n'avons trouvé aucune recherche discutant des biais affectant d'autres identités de genre (par exemple, les personnes non-binaires

Tableau 4.2. *Le nombre d'articles décrivant les différents groupes affectés par des biais dans chaque catégorie de machines sociales. Le total peut dépasser 181 car plusieurs articles discutent de plusieurs problèmes de biais. * Ces attributs ne sont pas intrinsèquement sensibles ; ils acquièrent leur signification à travers la machine sociale spécifique ou sont générés par elle.*

Groupe/attribut	Wikipédia	Marché en ligne	Révision par les pairs	Reddit, YouTube	Jeux en ligne
genre	81	13	24	19	12
race	2	13	-	6	1
culture/géographique/nationalité	44	2	12	4	-
orientation sexuelle	1	1	-	-	-
religion	-	-	-	2	-
statut de handicap	-	1	-	-	-
utilisateurs anonymes*	3	-	-	-	-
affiliation académique/prestige*	-	-	14	-	-
être avec des enfants*	-	1	-	-	-

ou transgenres).

Un deuxième attribut démographique pertinent est la *culture*, que nous pouvons associer à la *langue*, *pays*, et à la *nationalité*. Les articles que nous regroupons dans cette catégorie décrivent le biais comme un "cultural bias", "geographical bias", "language-specific bias", "local bias", "nationality bias" and "political bias". Dans ces articles, les groupes sociaux affectés par les biais sont directement ou indirectement décrits comme des groupes culturels –

résidents de certains pays ou locuteurs de certaines langues, avec l'implication que ces attributs définissent des groupes culturels relativement homogènes.

Le troisième attribut sensible est la *race* ou *l'ethnicité*, qui est fréquemment abordée dans les articles examinés. Ces articles soulignent souvent la manière dont les individus non-blancs, y compris les personnes noires, asiatiques et du Moyen-Orient, sont traités différemment par rapport aux individus blancs.

Enfin, un petit nombre d'articles ont rapporté des biais fondés sur *l'orientation sexuelle* (comment les couples de même sexe sont traités sur Airbnb [2]), la religion (comment certaines religions sont représentées sur Reddit [10]) et le statut de handicap (comment les personnes handicapées sont mal servies par Uber [69]).

4.4.2 Attributs définis par des machines sociales spécifiques

Sur Wikipédia, les utilisateurs peuvent être traités différemment selon leur *statut d'enregistrement*, ce qui distingue les utilisateurs qui contribuent anonymement de ceux qui possèdent un compte Wikipédia. Ces groupes peuvent avoir une corrélation avec des groupes démographiques définis par le genre ou la race, mais cela ne peut être confirmé car nous ne connaissons pas la composition démographique des utilisateurs anonymes.

L'affiliation/le prestige est un attribut non sensible inhérent au milieu académique et, contrairement au statut d'enregistrement sur Wikipédia, il n'est pas directement généré par la machine sociale. Cependant, dans le contexte de la révision par les pairs académiques, il est considéré comme

inacceptable de traiter les gens différemment en fonction de leur affiliation.

De même, être accompagné d'enfants est un attribut non sensible qui peut influencer la qualité du service dans la machine sociale d'Uber.

4.4.3 Discussion

Comme le montre le Tableau 4.2, certains biais sont répandus dans la plupart des machines sociales et ont été observés dans de nombreuses études. Par exemple, les biais de genre, de race et de culture sont courants, car l'appartenance à ces groupes est relativement facile à observer dans le contexte de ces machines sociales (par exemple, à travers les noms, les images ou les étiquettes spécifiques à la machine sociale) et incite les utilisateurs à agir de manière biaisée.

La visibilité de ces attributs permet également aux chercheurs d'enquêter et de mesurer efficacement ces biais. D'autres attributs, comme la religion ou l'orientation sexuelle d'une personne, ne sont pas aussi facilement observables. En conséquence, ils peuvent se manifester moins fréquemment dans les machines sociales, ou ils peuvent être présents mais plus difficiles à détecter. Cependant, nous avons également observé que de nombreux biais affectent plusieurs groupes de manière parallèle (par exemple, les femmes et les minorités raciales), ce qui suggère que l'étude présente des phénomènes de biais existants peut servir de guide pour les recherches futures, où l'on pourrait vérifier si d'autres groupes sont affectés par les mêmes types de biais.

Enfin, les biais liés au statut de handicap, à l'anonymat, à l'affiliation

académique ou au fait d'être accompagné d'enfants ont tendance à se manifester moins fréquemment car ils sont principalement pertinents dans des catégories spécifiques de machines sociales.

De ces différentes machines sociales et des cas de biais documentés, nous pouvons observer que chaque fois que des attributs sensibles sont visibles dans une machine sociale, il y a souvent des comportements biaisés parmi certains utilisateurs et des résultats inéquitables. Cela suggère qu'une bonne stratégie pour prévenir les biais est de concevoir des machines sociales de manière à dissimuler les attributs sensibles lorsqu'ils ne sont pas essentiels au fonctionnement de la machine sociale. La pratique de la révision en double aveugle (double-blind) dans les révisions par les pairs académiques est un bon exemple de cela. Un autre exemple peut être observé sur TaskRabbit, où la race et le genre des "taskers" de services sont visibles à travers leurs photos, ce qui peut conduire à une prise de décision biaisée par les utilisateurs. En revanche, sur Fiverr, la pratique courante d'utiliser des photos non humaines (comme des avatars) garantit que la race et le genre ne sont pas visibles, réduisant ainsi la probabilité de comportements discriminatoires basés sur ces attributs. Bien sûr, cela est efficace si d'autres éléments, tels que le nom de la personne, ne révèlent pas sa race ou son genre, et si les utilisateurs du système ne se rencontrent pas en personne (comme cela se produit généralement dans le contexte de Taskrabbit ou Uber).

Chapitre 5

Phénomènes de biais, préjudice et équité

Dans ce chapitre, nous nous penchons sur nos questions de recherche **(QR3)** et **(QR4)**, qui concernent une analyse normative des biais identifiés dans la section précédente :

> **(QR3)** En quel sens ces biais sont-ils considérés comme nuisibles ? et **(QR4)** Peut-on les relier à des attentes normatives exprimées sous la forme de critères techniques d'équité ?

Pour répondre à cette question, nous catégorisons les phénomènes rapportés dans la section 4.3 en un petit nombre de phénomènes plus abstraits, puis analysons chacun d'eux et les mettons en relation avec des concepts normatifs issus de l'équité algorithmique, à savoir les concepts de *préjudice* et les critères techniques d'équité.

5.1 Biais et préjudice

5.1.1 Définitions du préjudice socio-technique

Nous adaptons une taxonomie du préjudice définie par Shelby et al. [108], avec des définitions supplémentaires de Bird et al. [15] et Blodgett et al. [16]. Ces définitions ont été proposées pour les systèmes algorithmiques et décrivent le préjudice qui "émerge à travers l'interaction entre les systèmes techniques et les facteurs sociaux" [108], une description qui s'applique aux systèmes algorithmiques mais aussi aux machines sociales sans composantes algorithmiques particulièrement saillantes.

- **Préjudice allocatif** se produit lorsqu'un système algorithmique prive les membres d'un groupe désavantagé d'informations, d'opportunités ou de ressources et alloue ces ressources à d'autres groupes dominants. Le préjudice allocatif survient généralement dans le contexte de *systèmes de décision binaire automatisés*, c'est-à-dire des systèmes algorithmiques où les entrées sont associées à des personnes, et les sorties (résultats de la décision) incluent un résultat plus souhaitable (par exemple, la personne obtient un prêt bancaire) et un résultat "négatif" correspondant (par exemple, le prêt est refusé). De plus, il est généralement supposé qu'il existe une "bonne" décision pour chaque personne et que les décisions des systèmes algorithmiques ne sont pas toujours correctes.

- **Préjudice lié à la qualité de service** est causé par des disparités dans la performance des systèmes algorithmiques qui ne fournissent pas la même qualité de service à différents groupes de personnes [15].

- **Préjudices de représentation** sont causés par des représentations d'un groupe social sous un jour moins favorable que d'autres, ou des représentations qui les dénigrent, les sous-représentent ou ne reconnaissent pas leur existence du tout (adapté de Blodgett et al.[16]).

- **Préjudice interpersonnel** inclut la *diminution de la santé et du bien-être*, qui se produit lorsque la technologie est utilisée pour faciliter la violence envers les personnes (par exemple, usurpation d'identité, cyberintimidation, etc.) ou pour révéler des données privées sur les personnes *(violations de la vie privée)*.

- **Préjudices sociétaux** se réfère aux SSTs qui accélèrent l'ampleur du préjudice en utilisant la *mésinformation* (la diffusion d'informations trompeuses, qu'il y ait ou non intention de tromper), la *désinformation* (informations fausses) et la *malinformation* ("informations authentiques qui sont partagées avec l'intention de nuire") [63].

Tableau 5.1. *Le nombre d'articles décrivant les différents types de biais dans Wikipédia. Le total peut dépasser 85 car plusieurs articles discutent de plusieurs problèmes de biais.*

Phénomène de biais	sous-participation	sous-représentation	représentation biaisée	impact disparate	préjudices associés
inégalité de participation dans WP	✓				-
nombre de biographies		✓			représentatif, qualité de service
liens dans les biographies		✓			représentatif
biais lexical dans les biographies			✓		représentatif
catégorisation des pages pour les biographies			✓		représentatif
sous-représentation des intérêts des femmes		✓			représentatif
biais culturel		✓			représentatif
surveillance excessive des utilisateurs anonymes				✓	qualité de service, allocatif
nombre de critiques		✓		✓	représentatif
biais de notation			✓	✓	représentatif
biais de classement dans les recherches			✓	✓	allocatif
sentiment dans les critiques			✓		représentatif
discrépance dans la qualité du service				✓	qualité de service, allocatif
biais lexical dans les médias sociaux			✓		représentatif
inégalité de participation (jeux en ligne)	✓				-
disparité dans la représentation des genres (jeux en ligne)		✓	✓		représentatif, qualité de service
biais de genre dans la sélection des évaluateurs		✓			représentatif, allocatifs
taux d'acceptation des articles				✓	allocatif

5.1.2 Une catégorisation des phénomènes de biais

Les phénomènes identifiés dans notre revue systématique (section 4.3) peuvent être regroupés en quatre catégories suivantes : *sous-participation*, *sous-représentation*, *représentation biaisée* et *impact disparate*. Ces catégories sont décrites ci-dessous et liées aux types de préjudice socio-technique énumérés ci-dessus.

Cette analyse est résumée dans le tableau 5.1 : les phénomènes spécifiques sont listés verticalement, et les colonnes 2 à 5 correspondent aux quatre catégories ci-dessous, la dernière colonne indiquant les types de préjudice causés par les phénomènes.

Sous-participation Dans de nombreux articles, le terme *biais* se réfère à la disparité de participation entre différents groupes démographiques (par exemple, hommes et femmes) au sein des machines sociales [94, 132, 54, 87, 83, 134, 86, 128]. Pour Wikipédia, cela inclut les lecteurs, les écrivains et les éditeurs, tandis que dans les jeux en ligne, cela se réfère aux joueurs.

Bien que le manque de participation des femmes – ou de tout autre groupe – ne corresponde pas en soi à une définition du préjudice, il est symptomatique d'autres problèmes (par exemple, des phénomènes réduisant la participation des femmes) et peut contribuer à causer d'autres phénomènes qui correspondent à des définitions de préjudice.

Sous-représentation Le biais de sous-représentation inclut la représentation insuffisante d'un groupe démographique par rapport à un autre groupe démographique (par exemple, hommes versus femmes). Comme le montre le Tableau 5.1, la sous-représentation d'un groupe démographique englobe divers phénomènes. Dans le contexte de Wikipédia, les femmes et les personnes provenant de pays non occidentaux sont sous-représentées par le faible nombre de biographies et de liens vers ces biographies. Dans les jeux en ligne, les femmes sont sous-représentées parmi les personnages du jeu, et dans la révision par les pairs, elles sont sous-représentées parmi les évaluateurs.

La sous-représentation d'un groupe, illustrée par le nombre de biographies de membres du groupe, correspond à la définition du *préjudice de représentation*. De plus, si les biographies de Wikipédia sont considérées comme ayant pour effet d'augmenter la réputation et la visibilité en ligne des personnes, ce phénomène agit comme un "service" qui ne fournit pas le même service aux membres du groupe affecté qu'aux non-membres : le préjudice correspond alors à la définition du *préjudice lié à la qualité de service*. Lorsque ce service non seulement sous-représente un groupe, mais le prive également de ressources, il constitue un *préjudice allocatif*. Par exemple, le biais de genre dans la sélection des évaluateurs peut priver des individus d'opportunités ayant un impact sur le développement de leur carrière.

Représentation biaisée De nombreux phénomènes à travers les machines sociales aboutissent à des représentations biaisées de groupes démographiques

spécifiques, notamment les femmes et les minorités raciales : les membres de ces groupes sont fréquemment dépeints sous un jour négatif ou à travers une lentille stéréotypée, ce qui est la définition classique du *préjudice de représentation*. Les avis négatifs dans les places de marché en ligne, les biais linguistiques (lexicaux) dans Wikipédia et les médias sociaux, les représentations négatives de personnages féminins dans les jeux en ligne contribuent tous au *préjudice de représentation* envers les groupes affectés. De plus, les représentations biaisées peuvent perpétuer les stéréotypes et renforcer les préjugés existants dans le cadre des plateformes en ligne, correspondant ainsi à la définition du *préjudice sociétal*.

De plus, lorsque les biais dans les évaluations (Biais de notation) et les classements de recherche sur les places de marché en ligne conduisent à une visibilité réduite et à moins d'opportunités commerciales, ces représentations causent également un *préjudice allocatif*. En outre, cette visibilité inégale perpétue un cycle de désavantage, définissant des *préjudices sociétaux*. Il y a également eu des cas où des individus particuliers (plutôt que des groupes) ont été "attaqués" via Wikipédia [74], c'est-à-dire que leur biographie a été éditée de manière excessivement négative : ici, le préjudice est une forme de *préjudice interpersonnel* où la technologie a été utilisée comme une arme pour cibler la réputation d'une autre personne, et contribuer à l'information la plus dommageable dans un article Wikipédia peut être considéré comme de la *malinformation*, car l'objectif principal est de causer du tort.

Impact disparate L'impact disparate, tel que défini par la loi américaine, fait référence aux pratiques en matière d'emploi qui entraînent une discrimination involontaire. Cependant, dans le contexte des machines sociales, nous donnons à ce terme une signification plus large. Dans ces systèmes, l'impact disparate se produit lorsque les machines sociales ou les algorithmes désavantagent de manière disproportionnée certains groupes en fonction de caractéristiques telles que la race, le genre ou leur statut au sein d'une machine sociale spécifique. Cet impact découle de facteurs systémiques qui produisent des résultats inégaux.

Dans les marchés en ligne, les disparités de qualité de service conduisent à un accès inégal aux opportunités. Par exemple, sur des plateformes comme Airbnb, cela peut se manifester par des inégalités dans les opportunités d'hébergement ou dans les revenus locatifs. Ces disparités se traduisent par une prestation de service discriminatoire qui affecte négativement l'expérience client globale, ce qui correspond aux *préjudices liés à la qualité de service*. En outre, cela constitue un *préjudice allocatif* lorsqu'un groupe est systématiquement privé de services essentiels.

Un autre exemple se trouve dans la revue par les pairs académique, où les taux d'acceptation des articles peuvent favoriser de manière disproportionnée certains groupes. Si le processus de revue par les pairs désavantage systématiquement les chercheurs en fonction de leur genre, race ou affiliation institutionnelle, cela se traduit par des taux d'acceptation plus bas pour ces groupes. Cela affecte non seulement leur carrière académique, mais les prive

également d'opportunités de reconnaissance, de financement et de progression de carrière, illustrant ainsi un *préjudice allocatif*.

5.2 Biais et équité décisionnelle

Nous pouvons maintenant aborder la seconde partie de notre question de recherche, que nous reformulons comme suit : si le *biais* exprime un manque d'équité, peut-on également relier chaque biais à un critère d'équité technique défini pour les systèmes algorithmiques ?

Pour répondre à cette question, nous devons relier les biais observés aux modèles généraux utilisés dans l'analyse des systèmes algorithmiques. Le principal modèle de système algorithmique étudié en termes d'équité est celui des *systèmes de décision binaire automatisée*. Dans ces décisions, les entrées sont associées à des personnes, et les sorties (résultat de la décision) incluent généralement une issue plus désirable (par exemple, la personne obtient un prêt bancaire) et une "résultat négative" correspondante (par exemple, le prêt est refusé). De plus, il est généralement supposé qu'il existe une "correcte" décision pour chaque personne, et que les décisions d'un système algorithmique ne sont pas toujours correctes. L'existence de cette décision "correcte" de référence détermine en partie les critères d'équité applicables.

Parmi les types de biais décrits pour les machines sociales, ceux répertoriés dans le tableau 5.2 s'inscrivent dans le modèle des décisions binaires.

Tableau 5.2. *Phénomènes de biais correspondant au modèle des décisions binaires.*

Phénomène de biais	systèmes de décision binaire		vérité de terrain
	entrée	sortie	
nombre de biographies	personne	a/ n'a pas de bio	notabilité
liens dans les biographies	biographie	lié à/ non lié à	N/A
sous-représentation des intérêts des femmes	article	accepté/supprimé	N/A
biais culturel	information	exprimé/supprimé	N/A
surveillance excessive des utilisateurs anonymes	modification	acceptée/rejetée	règles de Wikipédia
nombre de critiques	personne	écrire/ne pas écrire une critique	N/A
biais de notation	personne	note élevée/faible	qualité du service
biais de classement dans les recherches	personne	classement élevé/faible	qualité du service
discrépance dans la qualité du service	personne	reçoit/ne reçoit pas de service	N/A
biais de genre dans la sélection des évaluateurs	évaluateur	sélectionné/non sélectionné	expertise et qualification
taux d'acceptation des articles	article	accepté/non accepté	qualité de la recherche

Tous les phénomènes de biais inclus sous *Sous-représentation* et *Impact Disparate*, à l'exception de la *Disparité de Représentation des Genres dans les Jeux en Ligne*, peuvent être modélisés par des décisions binaires. Pour les *biais de notation* et de*classement de recherche*, si le nombre de notations ou de classements dépasse deux, ceux-ci peuvent être modélisés comme des décisions avec plus de deux résultats possibles, mais où certains sont claire-

ment plus souhaitables que d'autres.

Dans les phénomènes modélisés par le modèle de décision binaire, l'entrée du système de décision est soit une personne, soit quelque chose appartenant à la personne (par exemple, une biographie, un article, une modification), ce qui rend l'acceptation ou le rejet de la décision directement impactant. En revanche, le phénomène de disparité de représentation des genres dans les jeux en ligne implique les personnages du jeu en tant qu'entrée du système de décision, ce qui le rend inadapté à la modélisation des décisions binaires.

Le tableau 5.2 illustre l'entrée et la sortie du modèle pour chaque phénomène. Par exemple, le processus d'écriture et d'ajout de biographies sur Wikipédia peut être modélisé comme un processus de décision binaire. Un instantané de Wikipédia représente un système de décision : étant donné une personne, la machine sociale de Wikipédia peut ou non lui avoir attribué une biographie. Idéalement, toutes les personnes notables auraient des biographies sur Wikipédia, tandis que les personnes non notables n'en auraient pas, ce qui indique qu'un concept de décision correcte vs incorrecte peut également être défini pour chaque personne.

Un autre exemple est celui de la *surveillance excessive*, qui relève à la fois de la *Sous-représentation* et de *l'Impact Disparate*. La surveillance excessive se réfère à un processus de décision binaire où une modification devrait être acceptée (si elle est conforme aux règles de Wikipédia) ou révoquée. Chaque entrée de décision est associée à un éditeur, qui subira une issue négative si sa modification est révoquée.

De nombreux critères techniques ont été proposés pour évaluer l'équité des décisions binaires. Ici, nous nous intéressons aux critères *d'équité de groupe*, c'est-à-dire ceux qui définissent l'équité par rapport à un groupe social identifiable. Dans ce qui suit, nous décrivons formellement les critères d'équité de groupe les plus en vue et discutons de leur applicabilité aux biais mentionnés ci-dessus. Nous utilisons les notations suivantes : l'appartenance d'une personne à un groupe est représentée par une variable aléatoire $A \in \{0, 1\}$, la décision "correcte" pour cette personne est représentée par une variable aléatoire $Y \in \{0, 1\}$, et la décision rendue par le système est représentée par $\hat{Y} \in \{0, 1\}$.

5.2.1 Parité démographique

L'un des critères d'équité les plus simples, également appelé indépendance ou parité statistique [135, 137], la parité démographique exige que la probabilité pour une personne d'obtenir un résultat positif du système de décision $(\hat{Y} = 1)$ soit indépendante de l'attribut sensible :

$$P[\hat{Y} = 1|A = 0] = P[\hat{Y} = 1|A = 1] \tag{5.1}$$

La parité démographique exprime que les membres du groupe considéré (par exemple, les femmes) devraient recevoir un résultat positif au même taux que ceux du groupe complémentaire (c'est-à-dire les hommes). Par exemple, dans le *nombre de biographies*, l'utilisation de la parité démographique

comme critère de référence en matière d'équité signifie ignorer la "correction" des décisions, c'est-à-dire si les résultats du processus sont conformes aux règles de Wikipédia (les biographies publiées sont celles de personnes notables, les modifications acceptées respectent les règles, etc.). Lorsque le nombre de biographies de femmes est simplement comparé à celui des biographies d'hommes (c'est-à-dire que seulement 17,9 % des biographies sont des biographies de femmes[121]), l'attente normative implicite est qu'il devrait être de 50 %, ce qui signifie que les hommes et les femmes devraient avoir des biographies à des taux égaux : en d'autres termes, l'attente normative est la parité démographique.

5.2.2 Equal opportunity

Également appelée séparation et parité du taux positif [59], ce critère définit l'équité en fonction de la probabilité d'obtenir un résultat favorable du système de décision, mais cette fois conditionnel à ce que ce résultat soit la décision correcte :

$$P[\hat{Y} = 1|Y = 1, A = 0] = P[\hat{Y} = 1|Y = 1, A = 1] \tag{5.2}$$

5.2.3 Equalized odds

étend *equal opportunity* pour inclure des taux d'erreur égaux : la probabilité d'obtenir un résultat favorable est égale entre les groupes, à la fois

pour les personnes pour lesquelles c'est la décision correcte et pour celles pour lesquelles le résultat négatif serait la décision correcte :

$$P[\hat{Y} = 1|Y = y, A = 0] = P[\hat{Y} = 1|Y = y, A = 1], \forall y \in 0, 1 \tag{5.3}$$

Contrairement à la parité démographique, l'égalité des chances et l'égalité des chances prennent en compte la décision *correcte* pour déterminer si les décisions sont équitables envers un groupe social. Par exemple, si nous considérons le processus de publication de biographies sur Wikipédia, l'application *equal opportunity* signifie vérifier si les femmes *notables* (ou les membres notables du groupe considéré) obtiennent des biographies au même taux que les hommes *notables*. Cette évaluation du biais différera de la parité démographique si la proportion de personnes notables est différente d'un groupe à l'autre. L'analyse de Wagner et al. [126], concluant qu'il n'y a pas de biais contre les femmes malgré la grande disparité dans le nombre de biographies d'hommes et de femmes, s'explique par cette formulation de l'équité. Adams et al. [1] considèrent le même concept d'équité (avec une "vérité de base" différente pour la notabilité) et obtiennent le résultat opposé. Par conséquent, pour appliquer Equal Opportunity et Equalized Odds, une vérité de base est requise. Dans le tableau 5.2, pour les processus de décision qui ont une vérité de base, nous pouvons appliquer des métriques d'équité telles que la parité démographique, equalized odds, ou predictive parity (ainsi que d'autres). Cependant, s'il n'existe aucune vérité de base disponible, la

parité démographique est la seule métrique applicable.

Pour illustrer avec l'exemple du *nombre de biographies* : sous l'attente de *equalized odds*, l'équité est atteinte si les critères de notabilité sont appliqués uniformément à tous les groupes. Cependant, sous l'attente de *parité démographique*, cette application uniforme n'est pas suffisante. Cela suggère que le critère de "vérité de base" lui-même peut être biaisé (par exemple, les hommes atteignent la "notabilité" (au sens de Wikipédia) à un taux plus élevé que les femmes). Dans ce contexte, atteindre l'équité nécessiterait de modifier les critères de notabilité afin que les taux de base soient égaux entre les groupes, rendant ainsi les deux critères d'équité équivalents.

5.3 Biais et équité dans la représentation

L'autre perspective sur l'équité dans les systèmes algorithmiques repose sur le concept de *représentations*, sans aucun modèle particulier du comportement du système : dans cette perspective, l'équité est définie par les représentations des groupes sociaux, produites par ou utilisées dans le système, où une représentation équitable est celle qui ne cause aucun *préjudice représentationnel* [28, 108], un concept qualitatif.

Nous avons regroupé les phénomènes où ce concept d'équité est pertinent sous le concept de *représentations biaisées des membres d'un groupe*, et ils consistent principalement en des *stéréotypes négatifs et un langage dégradant* dans la représentation des groupes sociaux. Bien qu'il n'existe

pas de définitions établies des mesures d'équité, les techniques utilisées pour détecter ces biais permettent d'inférer l'attente normative implicite d'équité.

Dans la plupart des machines sociales, les représentations d'un groupe apparaissent généralement sous forme textuelle, et les biais de représentation sont souvent mesurés en sélectionnant des mots ou des concepts qui représentent un stéréotype particulier, et en comptant leur occurrence dans les textes associés aux différents groupes (par exemple, les biographies des hommes et des femmes, ou les publications sur les réseaux sociaux concernant les membres du groupe) [126]. L'attente normative implicite est que les stéréotypes affectant un groupe social donné ne devraient pas être détectables dans les articles décrivant les personnes de ce groupe, ce qui est cohérent avec une forme *d'équité contrefactuelle*, similaire à un concept défini pour la classification de texte [50]. En d'autres termes, étant donné la description d'une personne ou d'un événement lié à un groupe social défavorisé, comment la description changerait-elle si la personne appartenait à un autre groupe social, ou si l'événement était associé à un autre groupe social ? L'équité exigerait que les changements ne soient pas liés aux stéréotypes concernant le groupe social.

Dans un cadre plus général, des techniques existent pour mesurer si un corpus (un grand ensemble de textes) est affecté par des biais indésirables, tels que des stéréotypes de genre. Ces biais sont véhiculés par la sémantique du texte, et dans les systèmes algorithmiques modernes de traitement automatique des langues (NLP) la sémantique est généralement encodée dans

la représentation linguistique interne du système, c'est-à-dire un modèle de langage statistique construit à partir d'un grand corpus de textes naturels. La technique de modélisation linguistique la plus courante, connue sous le nom de *word embedding*, consiste à mapper des mots dans des espaces vectoriels de haute dimension de manière à ce que les mots ayant des significations similaires soient rapprochés les uns des autres dans l'espace vectoriel.

Les biais peuvent être mesurés dans les word embeddings en utilisant une technique adaptée du *Implicit Association Test* [56] utilisé pour identifier les biais chez les humains : essentiellement, le test mesure si un mot se référant à un groupe social est plus similaire dans l'espace sémantique à des termes agréables ou désagréables, ce qui indique si le modèle de langage est biaisé positivement ou négativement à l'égard de ce groupe [20]. En plus de capturer le sentiment général envers un groupe, des mesures similaires peuvent capturer des stéréotypes et des biais sociétaux tels que l'idée que les tâches ménagères sont principalement réservées aux femmes [18]. La technique a également été étendue pour évaluer les biais dans des phrases complètes en utilisant des *sentence encodings* au lieu des word embeddings individuels [76, 118] : les tests pour les mots sont connus sous le nom de tests WEAT, et les tests d'encodage de phrases sont appelés tests SEAT.

Chapitre 6

La représentation des femmes dans Wikipédia : une étude de cas des politiciennes américaines

Ce chapitre aborde la **(QR5)** en explorant comment la définition de mesures spécifiques d'équité peut être utilisée pour identifier les sources de biais au sein des machines sociales. Nous nous concentrons sur les mesures d'équité et proposons un cadre pour comprendre les origines du biais.

6.1 La sous-représentation des femmes

6.1.1 Travaux connexes

De nombreux articles comparent le nombre de biographies d'hommes et de femmes [126, 122, 134, 129, 133] (les autres identités de genre ne sont pas étudiées ici non plus) et constatent que les femmes sont significativement sous-représentées : en juillet 2019, seulement 17,9 % des biographies sur Wikipédia étaient des biographies de femmes [122]. Cependant, ces décomptes absolus offrent une compréhension limitée de l'effet de Wikipédia lui-même : la question la plus pertinente est de savoir si les femmes *notables* sont sous-représentées sur Wikipédia. Plusieurs auteurs considèrent que la représentation équitable des genres se traduit par une probabilité égale pour les hommes et les femmes notables d'avoir une biographie sur Wikipédia. En 2011, Reagle et Rhue [96] ont collecté des listes de personnes notables à partir de sources biographiques (listes de personnalités influentes compilées par des magazines, dictionnaires biographiques, etc.) et ont évalué la représentation de ces personnes notables sur Wikipédia. Ils ont découvert que les hommes notables avaient au moins deux fois plus de chances que les femmes d'avoir une biographie sur Wikipédia. En 2016, Wagner et al. [127] évaluent la notoriété des hommes et des femmes sur Wikipédia en la comparant à des mesures de notoriété internes et externes, telles que le nombre d'éditions linguistiques contenant une biographie et le volume de recherches Google pour

la personne. Selon le critère des éditions linguistiques, ils constatent que les femmes sur Wikipédia sont en moyenne de 4 % à 12 % plus notables que leurs homologues masculins, ce qui suggère qu'il existe un seuil de notoriété plus élevé pour qu'une femme obtienne une biographie. Les valeurs exactes ne peuvent pas être comparées aux conclusions de Reagle et Rhue car la méthode d'analyse ne comptabilise pas le nombre de personnes notables absentes de Wikipédia.

L'étude la plus similaire à la nôtre est celle d'Adams et al. [1], qui examine les biographies de sociologues américains. En utilisant des définitions bibliométriques de la notoriété, ils constatent que pour des niveaux de notoriété similaires, les hommes ont au moins deux fois plus de chances d'avoir une biographie sur Wikipédia.

Toutes ces études s'accordent à dire que les femmes sont sous-représentées sur Wikipédia, mais il n'existe pas d'évaluation concluante (au sens de fournir une quantification exacte) de ce biais pour l'ensemble des biographies. Notre analyse actuelle contribuera à cette analyse générale en étudiant un nouvel ensemble de biographies, et il est intéressant de noter que pour cette population particulière, la sous-représentation des femmes est bien inférieure aux valeurs rapportées dans les travaux précédents.

6.1.2 Méthodologie

Dans le but d'évaluer le biais spécifiquement introduit dans Wikipédia, par opposition aux différences de genre dans la société que Wikipédia décrit,

nous avons deux objectifs : évaluer la sous-représentation des femmes et déterminer si leur représentation est affectée par des stéréotypes de genre.

Pour le premier objectif, nous suivons une approche similaire à celle d'Adams et al. [1], qui évaluent la représentation des sociologues dans les universités américaines à travers les genres et les ethnies : en nous concentrant sur un ensemble de personnes partageant une occupation, les taux de représentation peuvent être comparés avec un critère de notoriété cohérent et mesurable. Dans cette étude, nous considérons un ensemble différent de personnes : les représentants élus dans les législatures des 50 États américains. Nous évaluons la probabilité qu'un représentant homme ou femme ait une biographie et comparons la distribution des longueurs des biographies.

Nous utilisons ensuite le même ensemble de données pour évaluer si la représentation d'une personne est influencée par des stéréotypes de genre.

6.1.3 Probabilité d'avoir une biographie

Dans notre ensemble de données, le nombre total de politiciens américains est de 6096. Parmi eux, 2 001 sont des femmes, 4 087 sont des hommes, 6 sont des femmes transgenres, et 2 sont non-binaires. En raison du petit nombre des deux derniers groupes, nous les excluons de notre analyse. Le pourcentage de femmes dans cet ensemble de politiciens est de 32,82 %. Le tableau 6.1 contient des informations statistiques sur le nombre de politiciens américains par genre.

	Politiciens américains		
	dans la société	avec biographie Wikipédia	sans biographie Wikipédia
femmes	2001	1774	227
hommes	4087	3685	402
total	6088	5459	629

Tableau 6.1. *Informations statistiques sur le nombre de politiciens américains selon le sexe (hommes et femmes)*

Sur ces 6096 politiciens américains, 5459 personnes ont des biographies sur Wikipédia, dont 1774 femmes et 3685 hommes. Par conséquent, le pourcentage de femmes politiciennes ayant une biographie sur Wikipédia est de 88,7 %. Pour les hommes, le pourcentage est de 90,2 %.

Nous pouvons modéliser la situation de la manière suivante : Wikipédia, en tant que machine sociale, classe les personnes comme notables (elles obtiennent une biographie sur Wikipédia) ou non notables (elles n'ont pas de biographie). Comme ces politiciens sont principalement notables en raison de leur rôle de représentants d'État élus, nous pouvons considérer que la vérité terrain est que tous sont notables (selon la politique de notabilité de Wikipédia[1]). Les taux d'erreur (faux négatifs) sont donc de 11,3 % pour les femmes et de 9,8 % pour les hommes.

Nous avons effectué un test du chi-squared pour évaluer l'association potentielle entre le genre et le fait d'avoir une biographie sur Wikipédia. Le genre a été considéré comme la variable *d'exposition* tandis que le fait d'avoir une biographie sur Wikipédia a servi de *résultat*, en utilisant le tableau 6.1.

[1]WP:NSUBPOL: https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)/Subnational_politicians

La statistique du chi-squared a donné une valeur de 3,14, avec une p-value correspondante de 0,077. Cela indique une signification au niveau de 0,1, c'est-à-dire une signification statistique très modérée.

6.1.4 Longueur des biographies

L'analyse précédente n'évalue que la probabilité qu'une personne ait une biographie. Nous comparons maintenant la longueur des biographies des hommes et des femmes, en évaluant si les femmes sont sous-représentées à cet égard. La figure 6.1 montre les distributions des longueurs des biographies pour les hommes (en bleu) et les femmes (en rouge). En moyenne, les biographies des femmes sont 2,92 % plus courtes (en termes de nombre de mots) que celles des hommes. Cependant, à la figure 6.2, il est évident que les diagrammes de la *Empirical cumulative distribution function* (CDF) et de la *probability density function* (PDF), sont presque identiques pour les hommes et les femmes. Cela indique une distribution des longueurs des biographies presque identique pour les deux genres.

6.2 Représentations biaisées

6.2.1 Travaux connexes

La notion de biais de contenu exprime le fait que les biographies des personnes sont façonnées par des stéréotypes liés à leurs caractéristiques

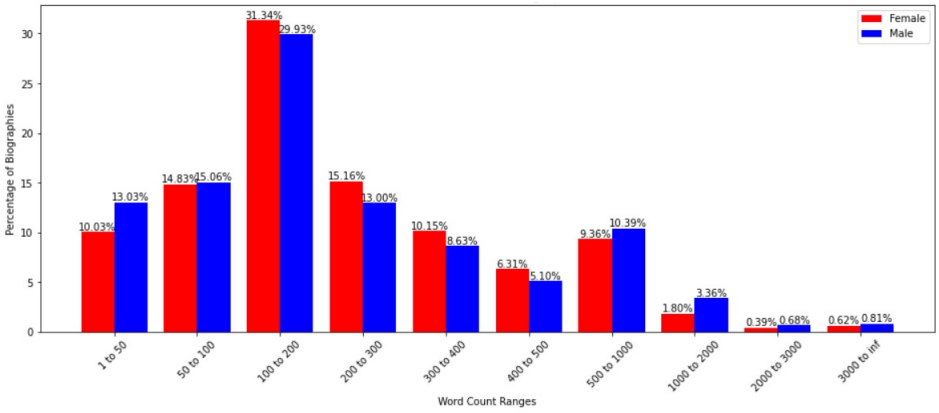
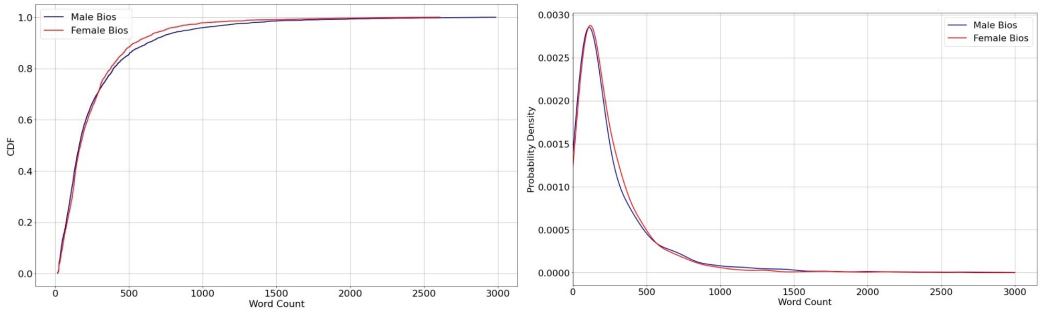


Figure 6.1. *Le pourcentage de biographies en fonction du nombre de mots dans les biographies.*



(a) *CDF* ***(b)*** *PDF*

Figure 6.2. *Fonction de Empirical cumulative distribution (CDF) et fonction de probability density (PDF) du nombre de mots pour les biographies d'hommes et de femmes.*

démographiques, telles que le genre ou la race. Selon la journaliste Ann Finkbeiner [42], les femmes scientifiques sont souvent décrites de manière à souligner leur genre et leur famille plutôt que leurs réalisations (par exemple, en mentionnant la profession de leur mari ou les arrangements pour la garde des enfants). Inspirés par ces observations, Wagner et al. [127] ont examiné le *topical bias* dans le contenu biographique de Wikipédia en utilisant *pointwise mutual information* (PMI) pour tester l'hypothèse selon

laquelle certains sujets sont surreprésentés dans les biographies de femmes : ces sujets sont le genre (par exemple, homme, femme, monsieur, madame), les relations amoureuses (par exemple, marié, divorcé, couple, épouse), et la famille (par exemple, enfants, mère). Ils constatent également un léger biais linguistique concernant l'utilisation de termes abstraits à connotation positive ou négative : les biographies de femmes contiennent plus de termes abstraits à connotation négative, tandis que les biographies d'hommes contiennent plus de termes abstraits à connotation positive, ce qui est interprété comme un biais inconscient des éditeurs masculins, qui décrivent les hommes comme faisant partie de "leur propre" (et véhiculent donc un sentiment généralement plus positif à propos des hommes), tandis que les femmes sont décrites comme faisant partie d'un groupe "autre" pour lequel les aspects positifs sont considérés comme spécifiques et les aspects négatifs comme plus généraux.

Bamman et Smith [7] utilisent un modèle à variable latente pour analyser les événements décrits dans les biographies de Wikipédia, et constatent que plusieurs types d'événements biographiques sont très inégalement répartis entre les biographies d'hommes et de femmes : par exemple, les biographies d'hommes sont beaucoup plus susceptibles de décrire l'engagement dans l'armée, tandis que les biographies de femmes sont beaucoup plus susceptibles de décrire la naissance d'un enfant.

Bien que ces analyses révèlent des différences de contenu significatives entre les biographies d'hommes et de femmes, les aspects thématiques pour-

raient être causés par des personnes adoptant des rôles traditionnels (par exemple, devenir soldat ou participer à des concours de beauté) plus que par une description biaisée. Ce qui rend cette question difficile à répondre est qu'il n'existe pas de "représentation du monde réel" évidente avec laquelle comparer les observations, contrairement au cas de la représentation quantitative des genres discuté plus haut.

Une technique a été proposée par Field et al.[41] pour contrôler la "représentation du monde réel" : l'idée est de construire des ensembles de données en associant chaque biographie d'un ensemble étudié à une autre biographie décrivant une personne ayant des caractéristiques très similaires, à l'exception de l'attribut étudié. Cela permet de comparer des personnes de genres différents mais dont les "histoires de vie" sont très similaires. La méthode a été utilisée pour comparer les portraits dans Wikipédia de personnes ayant des orientations sexuelles différentes [93], mais n'a pas été utilisée pour analyser les biais de genre.

6.2.2 Méthodologie

Notre second objectif est d'évaluer le biais de contenu, c'est-à-dire de déterminer si la représentation des hommes et des femmes est fortement influencée par des stéréotypes de genre, et comment cela change lorsque nous contrôlons les différences sous-jacentes dans les "histoires de vie", des personnes, c'est-à-dire comment leurs professions et événements de vie se conforment aux rôles de genre traditionnels.

À cette fin, nous étudions à nouveau la représentation des hommes et des femmes dans notre ensemble de données de politiciens américains. Bien que nous ne contrôlions pas entièrement les "différences dans les histoires de vie" des personnes, cela fournit un certain niveau de contrôle dans la mesure où les personnes décrites sont toutes des politiciens élus. Nous cherchons d'abord à déterminer si les biographies d'hommes et de femmes sont *systématiquement différentes*.

De plus, au lieu de rechercher des formes spécifiques de biais informées par une expertise du domaine (par exemple, des termes abstraits à connotation positive étant utilisés pour les hommes plus que pour les femmes et vice-versa pour la connotation négative), nous tentons d'évaluer l'ampleur globale de la différence de représentations : notre hypothèse est que si le genre de la personne décrite dans une biographie peut être identifié avec une grande précision (en ignorant les indications explicites telles que les noms ou pronoms genrés), alors nous pouvons dire que les biographies sont systématiquement différentes, et cette précision fournit en outre une indication de si le biais est plus ou moins important dans des ensembles spécifiques de biographies. Dans ces expériences, nous entraînons des classificateurs sur des représentations sémantiques de biographies en texte intégral, à savoir des embeddings de documents. Nous comparons également la précision d'un classificateur automatique avec la capacité d'un lecteur humain à distinguer les genres, ainsi que des mesures géométriques sur les ensembles de vecteurs d'embeddings.

Dans un second temps, nous tentons d'extraire les *phrases* qui sont les plus forts indicateurs de genre. Les études précédentes reposent souvent sur le décompte de mots pour identifier les biais de genre, mais nous soutenons qu'en supprimant le contexte de chaque mot, nous ne pouvons pas déterminer si sa présence (ou son absence) est due à un élément significatif de "l'histoire de vie" de la personne (éventuellement lié à la conformité aux rôles de genre traditionnels) ou à une description biaisée au sein de Wikipédia. En revanche, les phrases peuvent nous donner des indicateurs plus significatifs de la manière dont les hommes sont décrits différemment des femmes.

Afin d'extraire les phrases les plus associées aux biographies d'hommes ou de femmes, nous entraînons des classificateurs d'apprentissage automatique sur l'ensemble de données des phrases trouvées dans les biographies, représentées par des embeddings de phrases. Nous extrayons ensuite les phrases qui sont les plus significativement masculines ou féminines, c'est-à-dire les plus éloignées de la frontière de classification dans l'espace vectoriel.

6.2.3 Classification des biographies par genre

Notre objectif ici est de fournir une mesure quantitative des différences systématiques entre les biographies d'hommes et de femmes. Notre hypothèse est que la précision maximale d'un classificateur entraîné à classer une biographie comme étant celle d'un homme ou d'une femme peut être utilisée comme mesure de substitution de la difficulté inhérente à distinguer les biographies d'hommes et de femmes, et indirectement pour mesurer à quel

point les hommes et les femmes sont décrits différemment. Nous notons que nous nous intéressons principalement à comparer la précision des classificateurs sur différents ensembles de données, plutôt qu'à découvrir la précision maximale exacte. Par conséquent, nous n'avons pas nécessairement besoin d'un classificateur optimal.

Comme entrée des modèles de classification, nous représentons les biographies sous forme d'embeddings de documents (documents embeddings; DE). En utilisant un modèle DE pré-entraîné[2], nous transformons chaque biographie en un vecteur de 96 dimensions V. Nous avons entraîné plusieurs types de classificateurs (logistic regression, support vector machine (SVM), gaussian naive bayes, decision tree, random forest) et nous avons finalement sélectionné un classificateur SVM avec un kernel polynomial de niveau 2, qui a la plus haute précision. Nous avons effectué ce test sur deux ensembles de données. Le premier ensemble contient 51 116 biographies sélectionnées aléatoirement parmi toutes les biographies disponibles sur Wikipédia. Le second ensemble est notre ensemble de données de politiciens américains décrit à la section 6.1.4. Dans le second ensemble de données, nous avons également supprimé les indications explicites de genre, y compris les pronoms (he, she, his, her, ...) et les termes spécifiques au genre tels que "wife", "husband" or "sister". Nous avons remplacé ces termes par des équivalents neutres tels que "they", "their", "partner", "sibling", etc.

Bien que l'utilisation des embeddings de documents crée déjà une représentation

[2]Modèles de la bibliothèque Spacy: https://spacy.io/models/en

de la sémantique du texte et n'inclut pas explicitement ces termes genrés, nous les supprimons afin de garantir que la représentation est vraiment centrée sur l'histoire de vie globale et évite de coder le genre à partir d'indications explicites. Ce changement représente environ deux à trois points de pourcentage dans la précision du classificateur, selon l'ensemble de données.

Résultats : Le classificateur SVM est capable de séparer les biographies en catégories masculines et féminines dans le premier ensemble de données avec une précision de 90 % et dans le second ensemble de données avec une précision de 68 %. Ce résultat révèle deux points importants. Tout d'abord, il indique que les hommes et les femmes sont décrits différemment et qu'il existe des différences indéniables dans les biographies des hommes et des femmes. Cependant, cela n'indique pas nécessairement une description biaisée et peut également révéler des différences sous-jacentes entre les histoires de vie des hommes notables et des femmes notables. La baisse de précision, de 90 % à 68 % (soit 22 %), indique qu'une partie substantielle de ces disparités disparaît lorsque nous contrôlons les professions des personnes, qui sont corrélées avec les rôles de genre traditionnels.

6.2.4 Classification humaine

En plus de la classification utilisant l'apprentissage automatique, nous avons également demandé à un expert humain de tenter de classifier un échantillon de biographies et d'identifier les caractéristiques saillantes. Il con-

vient de noter que dans l'échantillon donné à l'expert, les noms des personnes avaient été retirés de leurs biographies, ainsi que toute indication explicite de genre (pronoms, etc., comme décrit précédemment). Pour un échantillon de 75 biographies, l'expert a été invité à deviner le genre et à indiquer son niveau de confiance sur trois niveaux : 0 pour une simple supposition, 1 pour une probabilité que ce soit un certain genre, et 2 pour une certitude totale. Les détails des suppositions de l'expert humain sont présentés dans le Tableau 6.2. Par exemple, le chiffre 10 dans la colonne $f-0$ et la ligne *femme* montre que 10 biographies de femmes ont été correctement reconnues par l'expert avec un niveau de confiance de 0, tandis que le chiffre 5 dans la colonne $f-0$ et la ligne *homme* indique que 5 biographies d'hommes ont été incorrectement identifiées comme étant celles de femmes avec un niveau de confiance de 0. Selon le Tableau 6.2, la précision du diagnostic de l'expert humain est de 69 %, tandis que les classificateurs d'apprentissage automatique ont atteint une précision de 68 %..

genre réel	détection du genre par l'expert					
	f-0	f-1	f-2	m-0	m-1	m-2
femme	10	7	6	8	6	3
homme	5	1	0	5	15	9

Tableau 6.2. *Détails des suppositions des experts humains. La première colonne montre le genre réel des biographies.*

6.2.5 Distance vectorielle moyenne

Nous corroborons les résultats de l'expérience de classification en comparant les vecteurs moyens des hommes (V_m) et des femmes (V_f) à travers deux ensembles de données. Cela implique de calculer le vecteur moyen pour les biographies des femmes (noté W) et des hommes (noté M), puis de mesurer la distance euclidienne entre ces vecteurs.

Pour l'ensemble de données couvrant toutes les professions, la valeur de la distance euclidienne est de 0,32, tandis que pour l'ensemble de données se concentrant spécifiquement sur les politiciens, elle est de 0,22. Cela indique qu'il existe une différence systématique dans la manière dont les hommes et les femmes sont décrits, en partie due aux rôles différents que les hommes et les femmes jouent dans la société. Cependant, même lorsque nous contrôlons la profession (en ne considérant que les politiciens de niveau étatique aux États-Unis), la différence entre les biographies des hommes et des femmes, mesurée par la distance dans l'espace vectoriel des embeddings de documents, reste significative.

Nous validons la signification statistique de la distance en utilisant le test de randomisation suivant. Selon l'hypothèse nulle, les biographies sont tirées d'une seule distribution, et la distance observée (0,22 dans l'espace vectoriel) est due à la composition aléatoire des deux échantillons.

Pour tester cela, nous avons combiné 522 biographies d'hommes et 522 biographies de femmes en un seul groupe, et avons répété le processus de

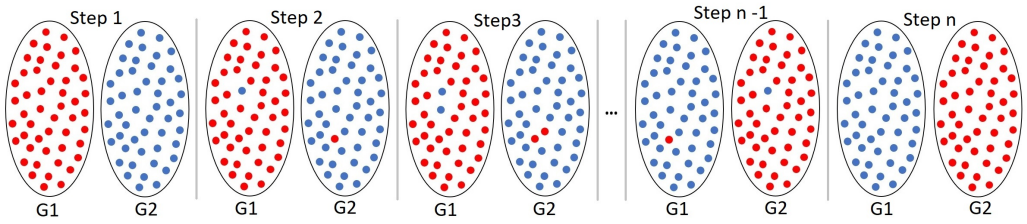
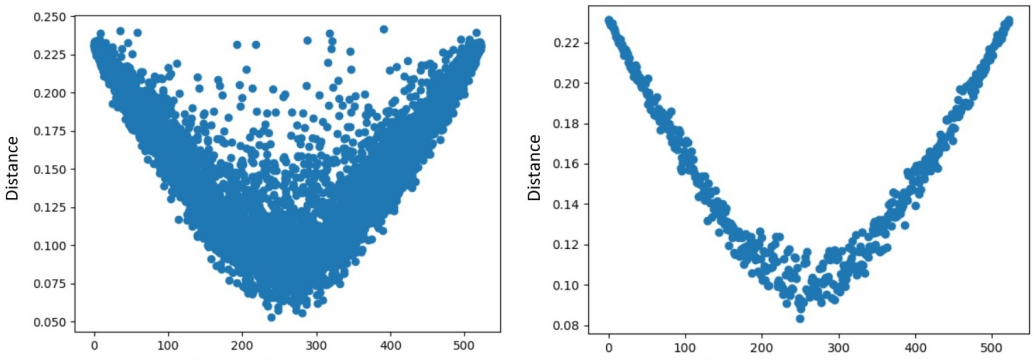


Figure 6.3. *Expérience avec la composition systématique de groupes combinant des biographies d'hommes et de femmes. Les points rouges représentent les biographies de femmes et les points bleus représentent les biographies d'hommes dans les groupes.*

séparation aléatoire de ce groupe en deux sous-groupes de taille égale $G1$ et $G2$, puis mesuré la distance entre les vecteurs moyens des deux groupes. Sur 100 000 essais, nous avons obtenu une distance supérieure ou égale à 0,22 seulement 173 fois, indiquant une valeur p de 0,00173, ou une signification au niveau 0,002.

En outre, nous avons créé une visualisation des résultats de l'essai en les traçant en fonction de la composition hommes/femmes des groupes. En composant systématiquement un groupe $G1$ avec k hommes et $n-k$ femmes, et k femmes et $n-k$ hommes dans le groupe $G2$, pour k variant de 0 à 522 comme indiqué dans la Figure 6.3, et en répétant cela dix fois aléatoirement, nous calculons et traçons les distances en fonction de la composition du groupe, comme indiqué dans la Figure 6.4. Le graphique montre clairement que les différences observées sont principalement dues à la composition en termes de genre, et ne sont que légèrement affectées par le caractère aléatoire des échantillons.



(a) Distance moyenne incluant les 10 répétitions *(b) Moyenne des distances moyennes des répétitions*

Figure 6.4. *La distance entre le vecteur moyen de G1 et le vecteur moyen de G2 pour les politiciens américains.*

6.2.6 Indications les plus fortes du genre

L'expérience précédente indique que la précision pour identifier les hommes et les femmes diminue en contrôlant la profession et les mots liés au genre, mais n'atteint pas 50 %. Il existe donc des indications permettant à un humain ou à un classificateur d'identifier les hommes ou les femmes. Notre objectif ici est d'extraire les indicateurs les plus forts qu'une biographie appartient à un homme ou à une femme, et d'évaluer si ces termes indiquent un biais ou plutôt des différences sous-jacentes dans les histoires de vie des personnes. Les entrées de notre classificateur dans l'expérience précédente étaient des embeddings de documents, où les dimensions individuelles des vecteurs ne sont pas facilement interprétables. Ici, nous revenons à une analyse au niveau des mots et des phrases.

Nous commençons par produire des nuages de mots mettant en évidence

les différences entre les biographies d'hommes et de femmes, comparables à ceux produits par Wagner et al. [127], à l'exception que nous comparons les nuages de mots pour un échantillon aléatoire de biographies avec l'ensemble de données spécifique aux biographies de politiciens américains (Figure 6.5).

(a) Nuage de mots de tous les hommes *(b) Nuage de mots des politiciens*

(c) Nuage de mots de toutes les femmes *(d) Nuage de mots des politiciennes*

Figure 6.5. *Mots les plus représentés dans les biographies d'hommes et de femmes, provenant d'un échantillon aléatoire de biographies (à gauche) et de notre ensemble de données sur les politiciens américains (à droite).*

Les poids des mots sont basés sur le ratio d'occurrences des mots entre les deux genres. Le nuage de mots extrait des biographies d'hommes sélectionnés aléatoirement (Figure 6.5a) contient des termes comme "karting", "snooker", et "featherweight", associés à des sports dominés par les hommes. À l'inverse, des mots tels que "bodice" et "sorority", indicatifs de

sujets spécifiques aux femmes, sont plus fréquents dans les biographies de femmes (Figure 6.5a).

En comparant les nuages de mots des politiciens (Figures 6.5b et 6.5d), il existe encore quelques différences thématiques, probablement dues à la description du parcours des politiciens (leurs activités avant d'être élus). Un contraste notable réside dans les différentes tailles des mots "husband" et "wife", suggérant que les biographies des femmes discutent davantage de leur mari que les biographies des hommes ne discutent de leur femme, comme le constatent Wagner et al. [127]. Cependant, même cette différence peut être partiellement expliquée par les déséquilibres sociaux sous-jacents du pouvoir. Les hommes tendant à avoir un statut social plus élevé que les femmes, si une telle différence existe dans un couple, il est plus probable que le mari d'une femme soit notable plutôt que l'inverse. Par exemple, la biographie de Michelle Obama parlera beaucoup de son mari très notable, Barack, bien plus que sa biographie ne parlera d'elle, et l'observation inverse peut être faite pour les biographies de Theresa May, l'ancienne Première ministre du Royaume-Uni, et de son mari Phillip May. Les stéréotypes de genre et les biais sociaux peuvent expliquer pourquoi les dirigeants mondiaux sont majoritairement des hommes avec des épouses moins notables, plutôt que des femmes avec des maris moins notables, avec un effet évident sur les occurrences de mots comme "wife" et "husband" dans leurs biographies.

Cette expérience nous conduit à deux observations. Premièrement, il est difficile de conclure à partir de ces nuages de mots dans quelle mesure les

différences observées sont attribuables à des différences dans les histoires de vie, ou à un biais de genre dans les descriptions. Une seconde observation est que bien que l'expérience de classification révèle des différences beaucoup plus faibles dans le second ensemble de données, cette différence quantitative n'est pas clairement visible lorsque nous comparons les nuages de mots, soulignant l'importance d'une analyse quantitative.

Pour évaluer les différences au niveau des phrases, nous employons à la fois des classificateurs d'apprentissage automatique et l'expert humain pour identifier les phrases qui permettent de distinguer efficacement si une biographie appartient à une femme ou à un homme. Nous extrayons l'ensemble complet des phrases apparaissant dans toutes les biographies, les marquons comme appartenant à une biographie d'homme ou de femme, et les convertissons en embeddings de phrases. Nous entraînons ensuite un classificateur SVM sur l'ensemble, en tentant de classifier chaque phrase comme "dite à propos d'un homme" ou "dite à propos d'une femme". Enfin, nous extrayons les phrases qui apparaissent les plus éloignées de la frontière de classification comme de forts indicateurs de genre.

Séparément, dans notre expérience de "classification humaine" nous avons demandé à l'expert humain d'identifier les phrases qui semblaient être des indications de genre. Globalement, la collection de phrases obtenues de cette manière était principalement liée aux rôles de genre traditionnels des hommes et des femmes, en termes de parcours, d'occupation et d'intérêts. Le Tableau 6.3 présente des exemples de phrases qui ont été identifiées comme

des indications probables de genre.

Phrases	Différences sociales	Genre détecté
received a Bachelor of Arts	intérêt	femme
the person has focused on topics including the disproportionate disappearance and domestic violence rates of Indigenous peoples in Montana	intérêt	femme
the person is an American politician and children's book author from Georgia	occupation	femme
While attending college, the person worked as a nanny for four years on the Upper West Side	occupation	femme
they has served the physical therapy community in leadership positions at Emory Healthcare and Children's Hospital of Atlanta.	occupation	femme
American business person from ...	occupation	homme
they was commissioned a second lieutenant in the United States Army	occupation	homme
The person enlisted in the United States Army Reserve in 2011 and is currently serving in the Judge Advocate General (JAG) Corps	occupation	homme
the person participated in the War in Afghanistan	occupation	homme
the person was sent to South Korea for about a month to participate in joint military training	occupation	homme
member of the Newark Portuguese Sports Club	intérêt	homme
served as a volunteer firefighter	occupation	homme

Tableau 6.3. *Les phrases que l'expert humain a utilisées comme critère pour déterminer le genre des biographies.*

Par exemple, l'implication dans l'art, l'écriture de livres pour enfants,

ou le travail comme nourrice suggèrent une biographie de femme, tandis que des rôles comme pompier, service militaire, ou expérience de combat en Afghanistan indiquent une biographie d'homme. Les différences observées semblent donc être associées aux normes sociétales et aux rôles de genre traditionnels concernant les intérêts et les occupations généralement attribués aux hommes et aux femmes, plutôt qu'à un quelconque biais de genre évident.

6.2.7 Résumé

Des travaux antérieurs ont identifié que dans les biographies de Wikipédia, les femmes sont sous-représentées et représentées de manière biaisée. Bien qu'il soit problématique que le nombre de biographies de femmes sur Wikipédia soit très faible, cela peut en partie s'expliquer par le fait que les rôles que les femmes ont tendance à adopter dans la société (par choix ou non) sont moins susceptibles de les rendre "notables". Cependant, ce qui est encore plus problématique, c'est que même lorsque nous essayons de contrôler la notoriété, les femmes restent fortement sous-représentées.

Notre analyse d'un ensemble de données de biographies de représentants au niveau des États américains contribue à la compréhension globale de ce phénomène. Nous constatons que les hommes et les femmes ont presque autant de chances d'avoir une biographie : les taux sont de 89 % pour les femmes et de 91 % pour les hommes. Cela contraste fortement avec la situation rapportée par Adams et al. [1] dans le monde universitaire : pour des niveaux de notoriété comparables, les professeurs hommes étaient plus

de deux fois plus susceptibles d'avoir une biographie qu'une professeure.

Une interprétation de cette différence est que dans un contexte politique, il peut y avoir un certain niveau de volonté institutionnelle pour que les représentants élus aient une biographie, ce qui s'appliquerait aux hommes comme aux femmes. Un critère clair de notoriété (être un représentant élu dans un cadre familier pour de nombreux contributeurs de Wikipédia) signifie qu'il y aura moins d'effet "glass-ceiling" empêchant les femmes d'être considérées comme notables.

En ce qui concerne les représentations biaisées des femmes, la plupart des travaux antérieurs rapportant des *biais lexicaux* sont basés sur des comptages de mots (ou de bigrammes), qui ne parviennent pas à capturer le contexte dans lequel les mots sont utilisés. Il est probable que les différences proviennent en partie des différences dans les "histoires de vie" entre les hommes et les femmes, et éliminer les biais dans la représentation des hommes et des femmes ne supprimera pas entièrement ces différences.

Nous avons montré que la performance des classificateurs et la distance entre les vecteurs d'embeddings de documents peuvent capturer les différences d'une manière qui les rend quantifiables. Lorsque les biographies décrivent des personnes notables pour les mêmes raisons, leurs histoires de vie sont similaires et la capacité d'un classificateur à les distinguer (hommes vs femmes) diminue de manière significative. De telles métriques seront utiles pour surveiller les niveaux de biais au fil du temps et à travers différents contextes.

Chapitre 7

Relations de causalité : une étude de cas de Wikipédia

Après avoir expliqué les biais observés en termes de dommages sociotechniques et d'équité, nous nous penchons maintenant sur les causes de ces différents phénomènes, en particulier dans le contexte de Wikipédia en tant qu'étude de cas. Nous examinons notre troisième question de recherche :

(QR6) Existe-t-il des relations de causalité claires entre ces phénomènes problématiques ?

L'objectif de cette question est d'identifier la structure causale de la situation : comment les différents phénomènes problématiques observés dans la machine sociale complexe s'influencent-ils mutuellement ? Comment pourrait-on influencer la situation globale si l'on tentait de réduire l'un des phénomènes ?

7.1 Sous-participation

Phénomène : les membres d'un groupe participent moins à Wikipédia que les membres d'autres groupes.

Causes : dans le cas des femmes, une grande variété de causes a été évoquée pour expliquer leur faible participation à Wikipédia. En particulier, Sue Gardner, ancienne directrice exécutive de Wikipédia, a identifié en 2011 neuf raisons à cela [48], dont la plupart sont discutées par d'autres études examinées dans le cadre de cette recherche. Ces raisons peuvent être résumées par les trois causes générales suivantes : (i) *self-exclusion*, (ii) *the unwelcoming atmosphere of Wikipedia*, et (iii) *the poor design of Wikipedia's interface*.

La première cause, *self-exclusion*, couvre des raisons purement sociales pour lesquelles les femmes ne peuvent pas, ou choisissent de ne pas participer à Wikipédia. Par "purely social" nous entendons que ces causes ne sont pas liées à la conception ou à l'atmosphère de Wikipédia elle-même : la moindre confiance en soi des femmes par rapport aux hommes, leurs vies déjà chargées (travail et une part disproportionnée des tâches domestiques).

La deuxième cause générale, que nous appelons *unwelcoming atmosphere of Wikipedia*, se réfère à des facteurs sociaux spécifiquement dans la manière dont ils se manifestent au sein de la communauté des éditeurs : la culture conflictuelle [22, 94], et une atmosphère parfois misogyne ou "sexuelle" atmosphere [48], qui sont généralement perçues différemment par les

hommes et les femmes.

Enfin, la troisième cause est liée aux aspects plus techniques de Wikipédia, y compris son interface utilisateur graphique et les protocoles d'interaction attendus avec les autres membres de la communauté. Un aspect important de cette cause générale est qu'elle est la seule qui puisse être directement abordée par des modifications techniques de Wikipédia elle-même.

Les causes de la sous-participation des autres groupes à Wikipédia n'ont pas été bien étudiées. Nous conjecturons qu'elles sont liées à des facteurs similaires : des facteurs d' *self-exclusion* (par exemple, la barrière linguistique), *unwelcoming atmosphere of Wikipedia* (si la communauté existante de contributeurs est principalement composée d'un groupe avec des normes sociales très différentes des leurs), et possiblement des facteurs liés à l'interface utilisateur.

En plus de ces facteurs, deux autres facteurs doivent être ajoutés, liés à d'autres phénomènes de biais sur Wikipédia. Premièrement, le problème de la "surveillance excessive" (over-policing) conduit clairement à une participation moindre des membres d'un groupe touché par ce phénomène. Comme nous ne connaissons pas la composition démographique des utilisateurs anonymes (par exemple, en termes de genre), nous ne pouvons pas déduire comment "over-policing" des utilisateurs anonymes peut affecter la participation des groupes définis par d'autres attributs (par exemple, les femmes ou les minorités visibles), mais il existe une possibilité claire qu'une relation de causalité existe ici. Deuxièmement, nous conjecturons que la sous-représentation des

intérêts d'un groupe sur Wikipédia pourrait également affecter la participation de ce groupe à Wikipédia, en renforçant l'idée que Wikipédia n'est pas "fait pour eux".

7.2 Sous-représentation des membres d'un groupe

Phénomène : les membres d'un groupe sont sous-représentés dans le contenu des articles de Wikipédia, que ce soit dans les biographies ou par des mentions dans d'autres articles.

Causes : la sous-représentation des femmes sur Wikipédia est liée à trois causes : la première est la sous-participation des femmes : en supposant que les femmes soient plus susceptibles que les hommes de s'intéresser à écrire sur d'autres femmes, il est logique qu'une communauté de contributrices plus petite crée moins de biographies de femmes ou de mentions de femmes dans d'autres articles, par rapport à la situation hypothétique où le pool de contributeurs serait plus équilibré. Des phénomènes similaires se produisent probablement pour d'autres groupes sociaux, comme les citoyens d'un pays donné ou les membres d'un groupe culturel très uni : nous nous attendons à ce que les membres du groupe soient plus susceptibles que les autres d'écrire sur leurs semblables. Une deuxième raison pertinente est l'influence des règles de *notoriété* sur Wikipédia. Comme l'indique le débat sur la question de savoir si les femmes sont réellement sous- ou surreprésentées parmi les personnes

notables [126, 1], les règles de notoriété de Wikipédia définissent (de manière vague) qui a droit à une biographie sur Wikipédia. Ces règles de notoriété se combinent également avec les processus de suppression de Wikipédia pour supprimer certaines des biographies existantes, et affectent ainsi indirectement la représentation des groupes.

Enfin, nous notons qu'une autre relation de causalité indirecte peut exister entre la sous-participation des membres d'un groupe et leur sous-représentation : les participants les plus actifs de Wikipédia contribuent à façonner les normes et politiques de Wikipédia, y compris les règles de notoriété. Un groupe pour lequel les critères de notoriété "externes" (par exemple, les références dans les journaux imprimés d'un certain pays) sont biaisés peut reconnaître que se baser sur ces critères de notoriété est une mauvaise idée, et pousser pour un changement. Nous conjecturons donc l'existence d'une relation de causalité supplémentaire entre la sous-participation des membres d'un groupe et les biais des règles de notoriété envers ce groupe.

7.3 Sous-représentation des intérêts et perspectives d'un groupe

Phénomène : peu d'articles discutent de sujets d'intérêt particulier pour un groupe, ou les perspectives du groupe sont sous-représentées dans les articles en général.

Causes : comme c'est le cas pour la sous-représentation d'un groupe, la sous-participation du groupe entraînerait également une sous-représentation des intérêts et perspectives du groupe. En général, les biais sociétaux se retrouvent également dans les contenus de Wikipédia, et contribuent à faire taire les perspectives des autres groupes : ici, l'effet est que les membres d'autres groupes, au lieu de simplement ne pas inclure la perspective du groupe dans un article, peuvent activement la supprimer [82].

7.4 Représentation biaisée d'un groupe

Phénomène : le contenu des articles transmet des représentations biaisées des membres de groupes sociaux particuliers.

Causes : dans ce cas, la raison dominante de ces biais est les biais sociétaux présents dans la société en général : le texte produit par "crowdsource" reflète naturellement les biais présents dans l'esprit de ses auteurs. Cependant, on peut s'attendre à ce que les membres du groupe concerné veuillent corriger cette situation, en éditant les articles biaisés afin de réduire leur biais ; en ce sens, le manque de participation d'un groupe est également susceptible de contribuer à des représentations biaisées de ce groupe.

7.5 Biais de surveillance excessive

Phénomène : les modifications apportées par un groupe particulier (utilisateurs anonymes) sont plus étroitement surveillées que celles d'autres groupes (utilisateurs enregistrés).

Causes : selon De Laat [33], ce phénomène se produit lorsque les éditeurs patrouillent les nouvelles modifications et interprètent mal l'interface des outils de signalement automatisés : conscients que les groupes anonymes sont plus susceptibles de vandaliser les articles ou de produire des contributions "mauvaises" (selon les politiques de Wikipédia), ils prêtent une attention particulière aux modifications apportées par les utilisateurs anonymes, plutôt que de simplement se fier au logiciel de signalement, qui considère déjà l'anonymat des utilisateurs comme un facteur de risque.

7.6 Résumé

Les différentes relations causales discutées dans les sous-sections précédentes sont illustrées à la figure 7.1. Les causes sont marquées par des rectangles bleus et sont représentées en deux groupes : (i) "social causes" (non directement liées à la partie technique de Wikipédia) sur le côté gauche de la figure et (ii) "Wikipedia's technical causes" qui incluent l'ensemble des processus, règles et outils techniques (y compris les bots) en haut de la figure. La cause

à l'intersection des "social causes" et des causes techniques de Wikipédia en haut à gauche, *unwelcoming atmosphere of Wikipedia* est à la fois une cause sociale (directement associée aux attitudes des personnes) et une cause technique, dans le sens où elle relève du champ d'action de Wikipédia en tant que SST.

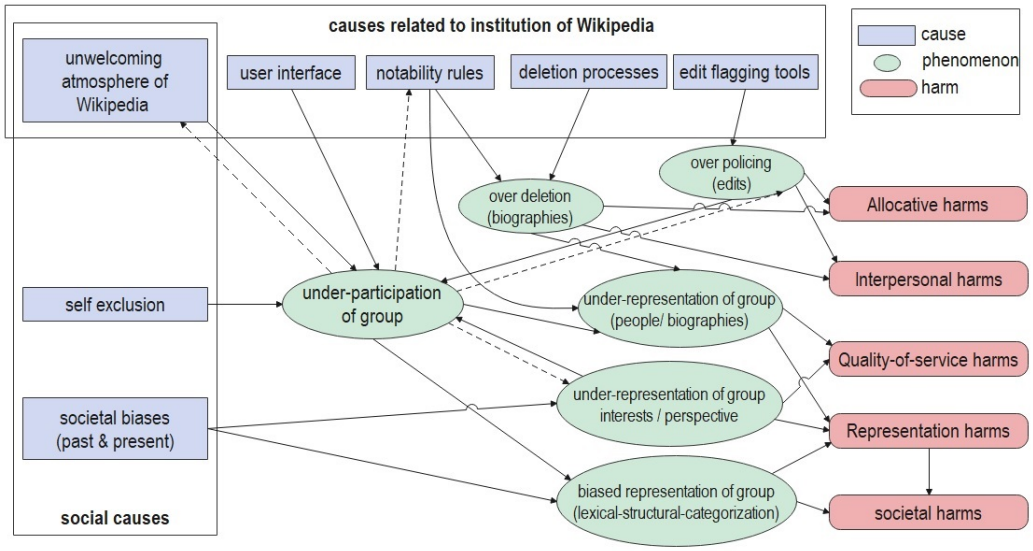


Figure 7.1. *Réseau de relations causales pour les phénomènes problématiques de Wikipedia.*

Les phénomènes de biais sont montrés avec des ellipses vertes, et les préjudices avec des rectangles rouges arrondis. Les flèches indiquent les relations causales documentées dans la littérature ; les flèches en pointillés représentent des relations causales supplémentaires que nous n'avons pas vues explicitement formulées dans la littérature, mais que nous croyons vraies. Les vérifier empiriquement fait partie des travaux futurs. Comme le montre la

figure 7.1, le phénomène de *sous-participation d'un groupe*, en plus d'être un phénomène de biais en soi, est l'une des causes de la formation de deux autres phénomènes : la *sous-représentation du groupe* et la *représentation biaisée du groupe*.

Chapitre 8

Une refonte du système de catégorisation de Wikipédia pour réduire les biais systémiques

Pour rappel, le troisième objectif de cette recherche est d'explorer comment la conception des machines sociales peut être modifiée afin d'atténuer les biais. Dans le chapitre 7, nous avons analysé les relations causales entre différents phénomènes de biais et examiné quels phénomènes, et dans quelle mesure, sont les plus efficaces pour engendrer des préjudices. En nous appuyant sur ces bases, ce chapitre se concentre sur les stratégies de réduction des biais à travers des interventions de conception.

Comme étude de cas, nous nous concentrons spécifiquement sur la catégorisation des articles biographiques dans Wikipédia — des entrées concernant des individus — où les catégories jouent un rôle essentiel dans la manière dont les personnes sont regroupées, perçues et contextualisées. Étant collectivement édité par des milliers de contributeurs, le contenu de Wikipédia est reconnu pour refléter une variété de biais sociétaux du monde réel, comme discuté au chapitre 4, notamment les biais de genre et les biais géographiques [126, 12]. Cependant, il existe très peu de recherches sur les biais du système de catégorisation.

Un cas notable de biais de genre dans les catégories de Wikipédia a été soulevé par Yanisky-Ravid et Mittelman dans un article de 2015 [132] : en 2013, la romancière américaine Amanda Filipacchi a remarqué que sa biographie avait été retirée de la catégorie des *American novelists* pour être placée dans la sous-catégorie des *American female novelists*. Bien que cette catégorisation soit correcte et pertinente, il n'existait pas de catégorie correspondante pour les *:* les hommes restaient dans la catégorie plus générale. Cette asymétrie (qui a depuis été corrigée) posait problème car elle renforçait l'idée que les hommes représentent le groupe par défaut ou non marqué, tandis que les femmes sont perçues comme des exceptions nécessitant une étiquette particulière.

En dehors du cas de Yanisky-Ravid et Mittelman [132] à notre connaissance, seules quelques recherches ont été menées sur l'édition catalane de Wikipédia [23, 39], lesquelles ont révélé que les identités masculines étaient

systématiquement considérées comme la norme et ont également constaté que les identités de genre diverses étaient mal représentées. Par coïncidence, leur rapport approfondi [39] s'est conclu par une recommandation visant à exploiter Wikidata afin d'améliorer la cohérence de la représentation des genres. Notre travail démontre, au moyen d'une preuve de concept, que cette solution est réalisable et peut permettre de traiter un éventail plus large de biais.

8.1 Analyse

8.1.1 Définir le biais dans le contexte des catégories de Wikipédia

Nous définissons le biais comme l'accentuation systématique ou disproportionnée de certains attributs démographiques — tels que le genre, la race ou la religion — d'une manière qui renforce les stéréotypes ou marginalise certains groupes.

Le système de catégorisation de Wikipédia est guidé par des politiques et des pratiques éditoriales visant à garantir la cohérence et la neutralité. Les directives pertinentes (WP:PEOPLECAT[1]) sont restées en grande partie inchangées depuis la première version rédigée en 2007. En théorie, ces directives devraient assurer la neutralité et l'équité. Toutefois, elles sont

[1] https://en.wikipedia.org/w/index.php?title=Wikipedia:PEOPLECAT

appliquées de manière inégale et influencées par le consensus éditorial, les précédents historiques et les débats communautaires. Cette incohérence contribue aux biais, même lorsque les décisions de catégorisation sont prises de bonne foi.

Un exemple illustratif est le cas historique de la catégorie "American novelists", évoqué dans l'introduction. Le fait de créer une sous-catégorie distincte pour les romancières, à partir d'une catégorie de romanciers implicitement masculine, revenait à exclure les femmes de l'identité plus large de "novelist" et à les marquer comme des exceptions. Ce schéma de catégorisation a été largement critiqué et a finalement été corrigé par la communauté, qui a créé une catégorie correspondante de "American male novelists".

Ce cas illustre un biais social connu sous le nom d' "othering" [113], c'est-à-dire la pratique consistant à "identifier des personnes par une caractéristique qui diffère d'un état normatif perçu, alors qu'elle est non pertinente."[2]. L'othering renforce les frontières sociales entre un 'nous' et un 'eux,' où les groupes marqués (par exemple, les femmes, les personnes noires, les minorités religieuses) sont considérés comme des exceptions par rapport à un groupe par défaut non marqué, généralement blanc, masculin et occidental.

Les directives de Wikipédia ne mentionnent pas le terme "othering", mais utilisent celui de "ghettoization" pour désigner la situation où les membres d'un groupe démographique (défini par leur origine ethnique, leur genre ou leur handicap) sont uniquement classés dans une catégorie basée sur cet

[2] https://en.wikipedia.org/wiki/Wikipedia:Othering

attribut. Cette règle vise à prévenir le risque de placer les individus exclusivement dans des sous-catégories identitaires sans également les inclure dans des catégories plus larges, professionnelles ou thématiques, ce qui revient à les isoler. Cela implique que des catégories fondées sur de tels attributs (par ex. *American women novelists*) sont acceptables tant que les personnes de ces catégories peuvent aussi être trouvées dans des catégories qui ne sont pas basées sur un attribut sensible (par ex. *American novelists by state / American novelists from New York*). Autrement dit, l'acceptabilité de ces catégories ne dépend pas de l'existence d'une catégorie pour le groupe dominant (les hommes, dans le cas des romanciers américains), et permet de mettre en avant les groupes sous-représentés, dans le sens discuté par Young et al.[134].

Le mécanisme de Wikipédia qui permet à une personne d'apparaître dans plusieurs catégories est le concept de *non-diffusing subcategories*: : ces catégories permettent de classer une personne à la fois dans la sous-catégorie et dans la catégorie parente, ou encore dans d'autres sous-catégories de cette même catégorie parente. Bien que les catégories de Wikipédia définies par un attribut sensible soient généralement explicitement marquées comme *non-diffusing*, le respect de la règle de la "ghettoization" exige que les biographies listées dans la sous-catégorie soient également présentes dans la catégorie parente, ou dans une autre sous-catégorie qui n'est pas définie par un attribut sensible. Une forme de biais systématique peut être observée lorsque les violations de cette règle affectent principalement les membres de groupes

sous-représentés.

Un autre enjeu important dans ce contexte est de savoir si l'attribut démographique selon lequel une personne est catégorisée est pertinent pour sa notoriété. Les directives de Wikipédia précisent que les catégories doivent reposer sur des caractéristiques définissant la notoriété du sujet et éviter la surcatégorisation ou l'accent inutile sur des aspects tels que le genre, la race ou la religion, à moins que ceux-ci ne soient centraux pour le sujet. Par exemple, étiqueter quelqu'un comme 'Black Canadian comedian' soulève la question suivante : son origine ethnique est-elle significative pour son identité professionnelle, ou ne sert-elle qu'à le marquer comme "different"? Les directives de Wikipédia tentent de trouver un équilibre entre visibilité et pertinence en limitant les catégories d'*intersection* aux seules considérées comme significatives ou historiquement pertinentes.

Un exemple est celui de la catégorie débattue *Jewish mathematicians*, qui a finalement été supprimée après un débat animé, l'intersection entre le fait d'être juif et celui d'être mathématicien n'ayant pas été jugée notable en soi. Comme l'a fait remarquer un contributeur : "There is no Jewish way of doing mathematics"[3].

Comme il est difficile de déterminer si une catégorie d'intersection particulière est pertinente ou notable, il existe un risque que cette directive soit également appliquée de manière incohérente d'un groupe démographique à

[3] https://en.wikipedia.org/wiki/Wikipedia:Categories_for_discussion/Log/2007_May_14#Category:Jewish_mathematicians

un autre. Nous constatons, à partir de notre analyse, que des catégories d'intersection moins pertinentes sont tout de même utilisées pour mettre en valeur les réalisations des membres de groupes sous-représentés et leur donner de la visibilité, sans qu'il y ait beaucoup de débats à ce sujet.

8.1.2 Méthodologie d'analyse

Afin d'évaluer la manière dont les biais se manifestent dans le système de catégorisation de Wikipédia, nous avons mené une analyse à grande échelle des catégories liées aux biographies de la Wikipédia en anglais. Notre attention s'est portée sur les catégorisations fondées sur l'identité — en particulier celles liées au genre et à l'ethnicité — où l'application incohérente des directives peut engendrer des biais systémiques. Avant de présenter nos résultats, nous exposons l'approche utilisée pour collecter, traiter et évaluer les données pertinentes.

Pour analyser les biais et les incohérences liés au genre et à l'origine ethnique, nous avons suivi un processus structuré de revue manuelle. Nous avons extrait de la Wikipédia en anglais toutes les catégories associées aux articles biographiques, en capturant leur structure hiérarchique et relationnelle. Les catégories de Wikipédia forment un graphe orienté, où la plupart des catégories fonctionnent à la fois comme sous-catégories et comme catégories parentes. Après la collecte des données, nous avons procédé à un prétraitement pour assurer la cohérence et réduire le bruit : nous avons supprimé les doublons ainsi que les catégories redondantes. Après cette étape,

l'ensemble de données final comptait 206,847 entrées uniques. Nous avons normalisé les données en convertissant tous les textes en minuscules et en uniformisant les conventions de nommage. Nous avons également écarté les enregistrements incomplets ou ambigus afin de préserver l'intégrité des données.

Ensuite, nous avons quantifié la présence de différents groupes en comptant la fréquence d'apparition des identifiants de genre ou de sexe (*men, women, male, female*) dans les libellés des catégories. Nous avons ensuite sélectionné un échantillon représentatif de catégories de Wikipédia liées aux professions, aux événements historiques et aux figures notables. Pour chaque catégorie, nous avons recueilli les entrées et documenté les attributs démographiques des individus, tels que le genre, l'origine ethnique et l'origine géographique (lorsqu'elle était disponible).

Bien que l'analyse ait impliqué une inspection manuelle, elle a été guidée par un cadre de codage cohérent qui nous a permis de suivre et de comparer systématiquement la représentation au sein des catégories.

8.1.3 Biais de genre dans les catégories

Un schéma clé mis en évidence dans notre analyse est le traitement inégal du genre dans le système de catégorisation de Wikipédia, où certains groupes sont plus explicitement marqués par le genre que d'autres.

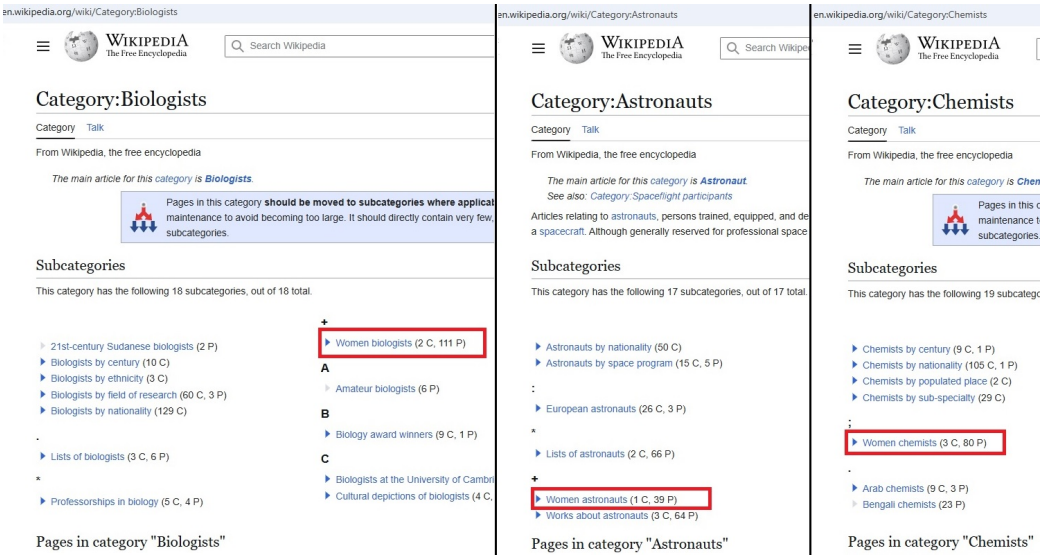


Figure 8.1. *Page de catégorie Wikipédia : Biologists, Astronauts et Chemists, les sous-catégories réservées aux femmes mises en évidence.*

Par exemple, comme l'illustre la Figure 8.1, les catégories telles que *Biologists*, *Astronauts* et *Chemists* comprennent des sous-catégories spécifiques comme *Women biologists*, *Women astronauts* et *Women chemists*, mais aucune sous-catégorie équivalente pour les hommes. Cela suggère que l'identité masculine est considérée comme la norme, tandis que les femmes sont marquées comme des exceptions. Ce schéma est largement répandu : nous avons identifié 11,049 catégories comportant un identifiant féminin sans contrepartie masculine, contre seulement 4,355 catégories comportant un identifiant masculin sans contrepartie féminine.

Les catégories existant uniquement pour les hommes (par exemple *Kuwaiti male weightlifters*) apparaissent presque exclusivement dans le domaine du sport, où la division systématique en groupes masculins/féminins

est courante. Dans de nombreuses combinaisons pays/sport, peu de sportifs sont suffisamment notables pour avoir une biographie sur Wikipédia (par exemple, un seul haltérophile koweïtien, homme ou femme, possède une biographie), et il s'agit fréquemment d'hommes. Cela produit une situation où la catégorie parente (par ex. *Kuwaiti weightlifters*) ne comporte qu'une sous-catégorie *male*.

En revanche, de nombreuses catégories explicites pour les femmes existent en dehors du domaine sportif et, comme dans les exemples ci-dessus (Figure 8.1), elles servent à mettre en valeur les réalisations des femmes dans des domaines historiquement dominés par les hommes [134], ce qui constitue une catégorisation acceptable selon les directives de Wikipédia.

Cependant, le biais se manifeste de deux manières. Tout d'abord, de nombreuses catégories réservées aux femmes violent la condition dite de "ghettoization", c'est-à-dire qu'elles contiennent des pages qui ne sont accessibles par aucune autre catégorie liée à la profession concernée. Par exemple, la catégorie *Women chemists* contient les pages de 79 personnes, dont 18 — soit près d'un quart — ne sont accessibles par aucun autre chemin de catégorisation depuis la catégorie parente *Chemist*. Toutes appartiennent à au moins une autre catégorie (par ex. *1991 deaths* ou *University of the West Indies alumni*), mais dans leur domaine professionnel de la chimie, ces femmes ne peuvent être trouvées qu'à travers leur genre. Il est important de noter que ce n'est pas faute de possibilités de catégorisation, puisque des catégories existent (ou pourraient être créées) pour les chimistes

de différentes spécialités, nationalités ou périodes historiques (par ex. *20th century chemists*).

Comme il n'existe pas de catégories explicites pour les hommes chimistes, cette situation indésirable — où une personne n'est atteignable dans sa catégorie professionnelle que via une sous-catégorie basée sur le genre — ne se produit pas pour les hommes. En tant que situation indésirable qui affecte de manière disproportionnée les femmes, cela peut clairement être décrit comme un biais de genre systématique.

Deuxièmement, le marquage explicite des femmes peut se justifier dans les professions dominées par les hommes, car il permet de mettre en valeur leurs réalisations en tant que minorité dans le domaine. Toutefois, on pourrait s'attendre à ce que la situation soit inversée dans les professions dominées par les femmes : dans ces cas, les catégories réservées aux femmes devraient constituer la norme, tandis que celles réservées aux hommes devraient être explicitement marquées comme l'exception. Or, ce n'est pas ce que l'on observe. Un exemple frappant est le nombre disproportionné de catégories liées aux "women nurses" par rapport à celles consacrées aux hommes exerçant la même profession. La surreprésentation des femmes dans les catégories liées aux soins infirmiers sur Wikipédia peut être mise en perspective par les données biographiques. Nous avons mené une analyse quantitative à partir de données structurées issues de Wikidata, en récupérant et comptant les biographies d'individus ayant pour profession "nurse" et dont le genre était renseigné. L'ensemble de données comprenait 7,031 femmes et

574 hommes identifiés comme infirmiers ou infirmières. Cela indique que les femmes représentent plus de 92% de l'ensemble des biographies d'infirmiers, ce qui correspond aux tendances observées dans le monde du travail, où les soins infirmiers sont une profession largement féminisée.

Cependant, bien que des catégories telles que "male nurses" et "men in nursing" existent, elles sont très largement dépassées par une multitude de catégories (plus de vingt) explicitement dédiées aux femmes dans les rôles infirmiers. Cela implique que, même dans une profession où les femmes sont de loin majoritaires, elles continuent à être explicitement désignées comme si elles constituaient une exception. De plus, il s'agit d'un cas de "intersection category", et l'on peut se demander si cette intersection est véritablement *notable* et mérite une catégorie spécifique. En d'autres termes, existe-t-il une manière spécifiquement féminine d'exercer la profession d'infirmier ? Nous répondrions par la négative. Dans ce cas, la surcatégorisation des femmes ne sert qu'à renforcer l'idée selon laquelle le fait d'être une femme serait une caractéristique centrale de l'identité professionnelle des infirmières.

En conclusion, l'examen des catégories genrées révèle un schéma répandu d'*othering* des femmes, avec des milliers de catégories les désignant comme étant d'une certaine manière inhabituelles ou particulières. Si certaines remplissent l'objectif utile de mettre en valeur leurs réalisations dans des domaines professionnels dominés par les hommes, cet avantage est terni par le fait qu'elles sont également explicitement marquées dans des domaines où elles constituent une écrasante majorité — ce qui rend l'objectif et la

pertinence de cette catégorisation discutables. De plus, en raison de cette catégorisation, une proportion significative des pages concernant les femmes enfreint les directives de Wikipédia sur la "ghettoization",en n'étant accessibles dans leur catégorie professionnelle qu'en tant que professionnelles *women*.

8.1.4 Biais raciaux dans les catégories

"race" comme on l'appelle parfois) est un autre attribut sensible explicitement couvert par les directives de catégorisation de Wikipédia.

Comme pour le genre, on observe ici une nette divergence entre la manière dont les identités blanches / caucasiennes sont décrites par rapport aux autres groupes ethniques. La majorité des biographies de Wikipédia concernent des personnes originaires d'Europe occidentale et d'Amérique du Nord [12], et la plupart des éditeurs proviennent également de pays occidentaux, en particulier dans l'édition en langue anglaise. Dans ce contexte, il n'est pas surprenant de constater que les identités *white* sont considérées comme l'identité par défaut (encore davantage que les hommes étant le genre "par défaut"). Cela se manifeste dans des catégories où les groupes ethniques autres que blancs sont largement signalés, incluant black, arab, latin/hispanic et d'autres. Nous notons également que l'identité juive est fréquemment utilisée de manière interchangeable comme catégorisation ethnique et religieuse.

Comme pour les femmes, de nombreuses catégories spécifiques à une

ethnie sont définies à l'intérieur de catégories professionnelles, sans lien apparent avec le domaine professionnel. Alors que la catégorie de *Jewish mathematicians* a été supprimée comme non pertinente, on peut s'interroger sur la raison pour laquelle il existe des catégories comme *Arab chemists* ou *Black British academics*.

Un examen détaillé de catégories telles que *Black Canadian people by occupation* illustre une répartition étendue des individus par domaine professionnel — allant des universitaires, activistes et acteurs aux scientifiques et écrivains. Il est à noter que ces identificateurs raciaux sont présents même dans des professions où la race n'est pas directement pertinente. Des exemples spécifiques incluent "Black British academics", "Black Canadian actors" et "Black French musicians," qui mettent l'accent sur l'identité raciale en parallèle des rôles professionnels ou sociaux.

Comme pour les femmes, cela crée des situations où des personnes catégorisées ne figurent que dans des sous-catégories spécifiques à leur appartenance ethnique. Par exemple, un examen détaillé de la catégorie *Arab chemists*, contenant 32 biographies, révèle que deux d'entre elles ne sont accessibles (depuis la catégorie *Chemists*) que dans cette sous-catégorie particulière.

En revanche, on observe une absence flagrante de catégories racialisées équivalentes pour les individus blancs — telles que *White British actors* ou *White Canadian writers*. Dans les rares cas où le terme "White" est explicitement mentionné, il apparaît surtout dans des contextes idéologiquement

chargés, tels que *White nationalists* ou *White supremacists.*

Encore une fois, les catégories ethniques servent vraisemblablement à mettre en valeur les réalisations de personnes issues de groupes minoritaires dans des domaines dominés par les blancs. Cependant, cette asymétrie dans la catégorisation indique que la blancheur est traitée comme l'identité non marquée, par défaut, tandis que les autres identités ethniques ou raciales sont positionnées comme exceptionnelles ou déviant de la norme. Ces catégories risquent d'enfermer les individus issus de minorités dans une lecture strictement raciale de leur identité et de renforcer les stéréotypes. Le regroupement des individus selon l'identité raciale dans des contextes professionnels peut involontairement présenter leurs réussites comme étant liées principalement à leur identité raciale plutôt qu'à leurs compétences, leurs contributions ou leur expertise. Cela contraste avec l'absence de catégories racialisées pour les individus blancs, dont les réalisations sont généralement présentées sans référence à la race, leur permettant d'exister uniquement dans leurs domaines professionnels ou culturels.

8.2 Refonte proposée : un système de catégorisation automatisé

Afin de répondre aux problèmes de biais et d'incohérence dans le système actuel de catégorisation manuelle de Wikipédia, nous proposons de passer à un cadre automatisé et systématique. Ce système repensé s'appuie sur les

données structurées disponibles dans Wikidata afin d'automatiser la catégorisation des individus en fonction de leurs attributs.

8.2.1 Conception du système

Dans le système proposé, les catégories fondées sur des attributs démographiques ne sont pas créées manuellement par les éditeurs de Wikipédia. Au contraire, un algorithme de catégorisation organise dynamiquement les individus dans des catégories en fonction de leurs attributs, automatiquement extraits de Wikidata. Ces attributs, tels que le genre, la profession, la religion ou la nationalité, sont saisis par les éditeurs de Wikidata et accessibles de manière indirecte depuis Wikipédia.

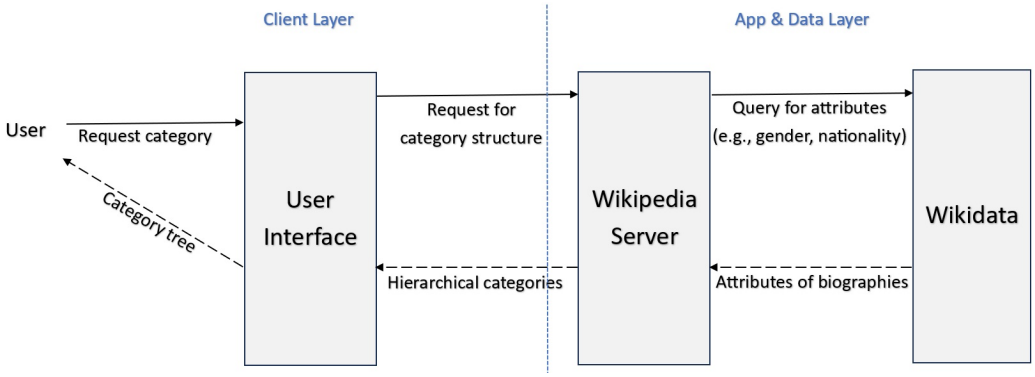


Figure 8.2. *Architecture proposée de notre système de catégorisation.*

La Figure 8.2 illustre l'architecture du système proposé, qui répartit les responsabilités entre une couche client et une couche application-données. Le processus commence avec l'utilisateur, qui interagit avec l'interface utilisateur dans un navigateur web pour demander une catégorie spécifique (par

ex. *Architects*). L'interface utilisateur transmet cette requête au serveur Wikipédia, qui agit comme l'unité centrale de traitement pour la construction des catégories. Le serveur Wikipédia interroge ensuite Wikidata afin de récupérer les attributs structurés pertinents — tels que la religion, le genre, la nationalité et d'autres — pour les individus associés à la catégorie demandée.

Ces attributs sont dynamiquement traités par le moteur de catégorisation sur le serveur Wikipédia afin de générer une structure hiérarchique de catégories cohérente. Cette structure est ensuite renvoyée à l'interface utilisateur, qui l'affiche à l'écran sous la forme d'un arbre de catégories.

8.2.1.1 Logique de sous-catégorisation :

Pour opérationnaliser ce cadre, nous avons conçu un algorithme de sous-catégorisation pas à pas qui construit dynamiquement les hiérarchies de catégories en fonction des attributs disponibles dans Wikidata. Le processus est le suivant :

Étape 1 : Lancement de la requête de catégorie Lorsqu'un utilisateur recherche des biographies en utilisant un filtre basé sur une catégorie (par ex., `occupation = "architect"`), l'interface utilisateur envoie une requête au serveur Wikipédia afin d'obtenir toutes les biographies pertinentes associées à cette catégorie.

Étape 2 : Demande des attributs des biographies Pour récupérer toutes les entités (biographies) avec l'occupation spécifiée (par ex., architectes),

le serveur Wikipédia envoie une requête à Wikidata et collecte les valeurs d'attributs structurés correspondantes pour chaque entité.

Étape 3 : Extraction des attributs À partir des enregistrements de Wikidata, le système extrait les attributs prédéfinis et leurs valeurs, tels que : `Gender/Sex = {man, woman, non-binary}`[4], `Religion = {Christian, Muslim, Jewish, ...}`, `Language = {English, French, Spanish, ...}`, `Nationality={...}`, `Ethnicity= {...}`, `Country= {...}`, et d'autres.

Étape 4 : Regroupement des attributs et comptage des fréquences Le système regroupe les biographies selon les différentes valeurs d'attributs (par ex., `man`, `woman`, `non-binary` pour le genre ; `Muslim`, `Christian`, `Jewish` pour la religion) et compte la fréquence de chaque groupe.

Étape 5 : Génération de la hiérarchie des catégories Une hiérarchie imbriquée est construite de sorte qu'à chaque niveau, une valeur d'attribut spécifique (par ex., `gender/sex = woman`) serve à former des sous-catégories. Les niveaux suivants incorporent des combinaisons d'autres attributs extraits (par ex., Religion, Ethnicité), comme illustré à la Figure 8.3. Cette approche permet la création de catégories telles que *Women Architects*, *Muslim Women Architects*.

Étape 6 : Rendu dans l'interface utilisateur La structure hiérarchique générée est renvoyée au client et affichée comme un arbre de catégories

[4]Nous notons que Wikidata dispose d'un seul attribut, P21, pour "sex or gender". Des valeurs comme "trans man" permettent de décrire à la fois le sexe et l'identité de genre dans un seul attribut, mais rendent les requêtes sur un seul composant plus difficiles.

navigable, permettant aux utilisateurs d'explorer les biographies selon de multiples dimensions identitaires.

Navigation L'arbre de catégories donne accès aux catégories basées sur des combinaisons d'attributs et, pour chaque catégorie, à toutes les biographies correspondant à cette combinaison de valeurs d'attributs. Par exemple, si une biographie inclut une valeur de genre (par ex., man, woman ou non-binary), elle sera catégorisée en conséquence. Par construction, toutes les catégories sont non diffusantes : la biographie d'une *muslim woman architect* sera accessible sous *Architects*, *Women architects*, *Muslim architects*, et *Muslim women architects*.

8.2.2 Implémentation du prototype

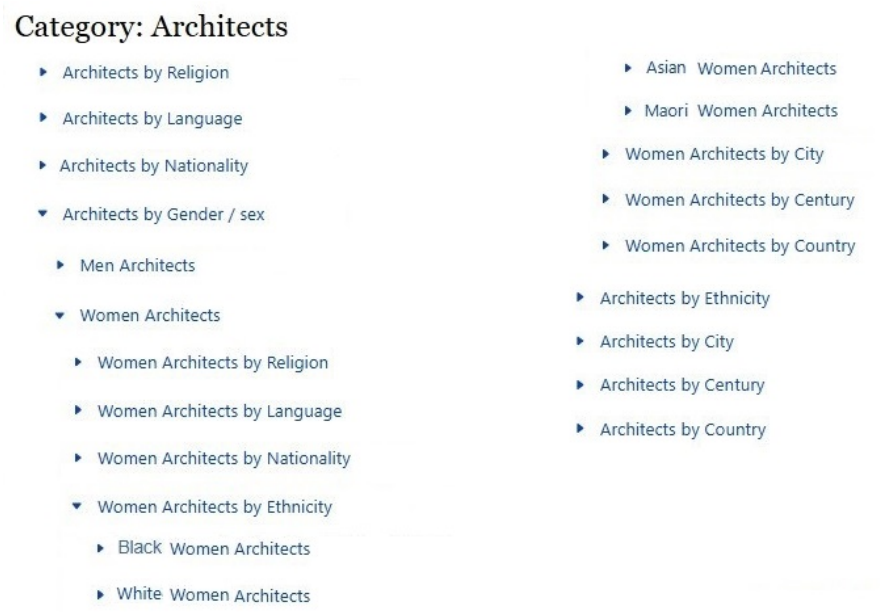


Figure 8.3. *Exemple de notre catégorisation hiérarchique pour l'occupation Architecte sur Wikipédia.*

Afin de valider la faisabilité de cette conception, nous avons développé une implémentation prototype. Pour ce prototype, nous avons sélectionné aléatoirement 88,939 biographies et extrait des attributs clés depuis Wikidata, notamment l'occupation, le genre/sexe, la religion, la langue, la nationalité, l'ethnicité, la ville, le siècle et le pays. Notre système de catégorisation repensé fonctionne hors ligne, mais il pourrait être intégré à Wikipédia en établissant une connexion en temps réel avec Wikidata, afin de mettre à jour dynamiquement les catégories en fonction des valeurs d'attributs.

La Figure 8.3 illustre cette approche de catégorisation hiérarchique en

utilisant l'occupation *Architect* comme exemple. Lorsqu'un utilisateur effectue une recherche par occupation et sélectionne *Architects*, il est dirigé vers une page listant toutes les biographies pertinentes. L'expansion de la catégorie *Architects* à l'aide de l'icône en forme de triangle révèle une série de sous-catégories, incluant *Architects by Gender/Sex*. En développant davantage ce nœud, on obtient des catégories spécifiques au genre, telles que *Men Architects* et *Women Architects*. La sélection d'une catégorie spécifique comme *Women Architects* révèle ensuite d'autres sous-catégories, subdivisées par nationalité, religion, langue ou ethnicité. Celles-ci donnent à leur tour accès à des catégories basées sur des valeurs d'attributs spécifiques (par ex., *Canadian Women Architects*, *White Women Architects*).

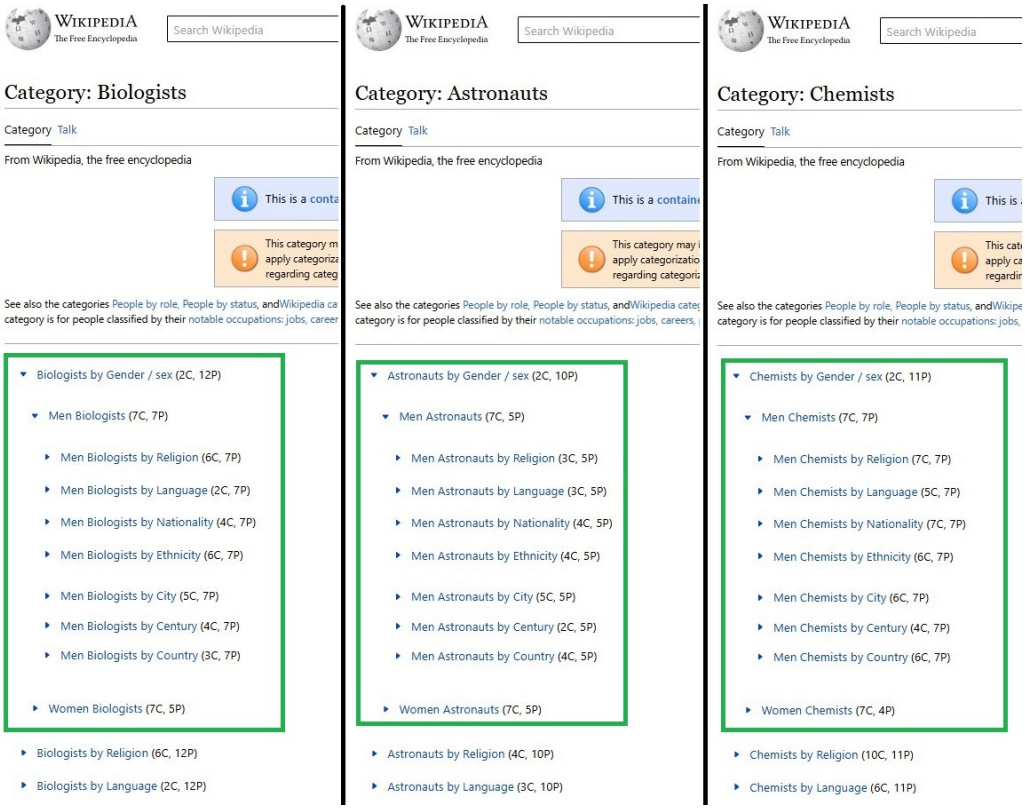


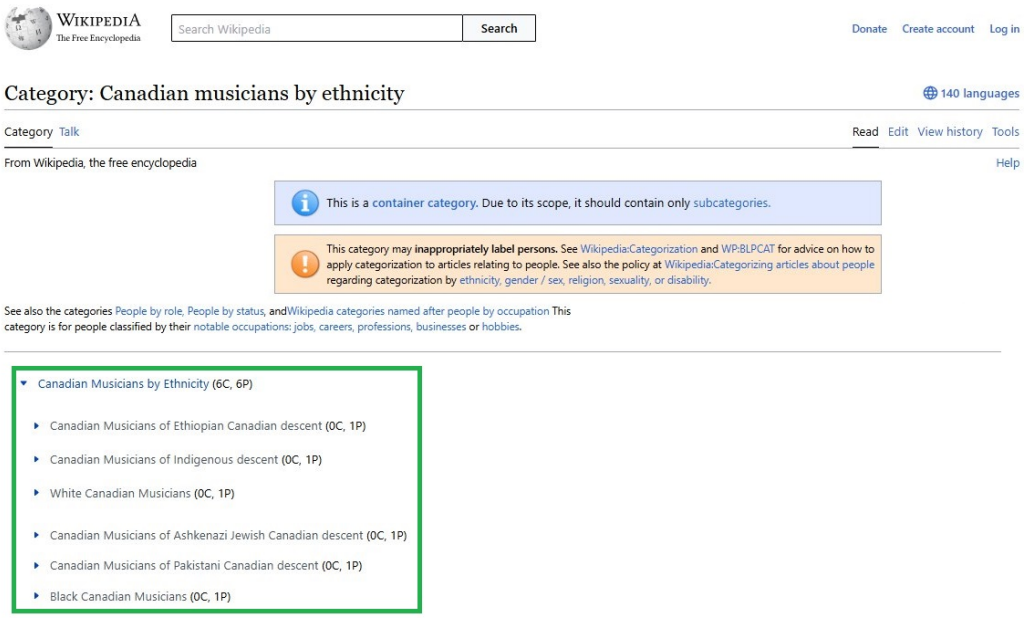
Figure 8.4. *Vue des catégories repensée : Biologistes, Astronautes et Chimistes.*

Comme le montre la figure 8.1, alors que les catégories originales de Wikipédia pour les *Architects*, *Biologists* et *Astronauts* comprenaient des sous-catégories pour les femmes mais pas pour les hommes, notre interface repensée rend visibles les deux identités de genre, comme illustré dans la figure 8.4. Les hommes et les femmes apparaissent dans leurs sous-catégories respectives, car notre jeu de données, basé sur un sous-ensemble de biographies de Wikipédia, inclut des individus identifiés comme *man* et *woman*. Si des individus d'autres identités de genre étaient présents lors de l'étape

initiale de récupération, notre système les catégoriserait également, assurant ainsi une représentation complète et inclusive au-delà du cadre binaire.

De plus, à l'intérieur des sous-catégories *men* et *women*, il existe souvent d'autres subdivisions basées sur des attributs tels que le pays, la religion, et d'autres encore. Dans notre système repensé, de manière similaire à la catégorisation actuelle de Wikipédia, les utilisateurs peuvent explorer les sous-catégories au sein des catégories plus larges basées sur le genre. Par exemple, sous la catégorie *Male biologists*, les utilisateurs peuvent trouver des sous-catégories comme *Male biologists by country* ou *Male biologists by religion*. Cette structure permet d'afficher les individus selon des combinaisons d'attributs. Comme dans le système actuel, lorsque les utilisateurs cliquent sur le petit triangle à côté d'une catégorie, ses sous-catégories sont révélées. En cliquant directement sur le nom d'une catégorie, on accède à une page dédiée listant les individus qui appartiennent à cette catégorie.

(a) Page de catégorie Wikipedia originale



(b) Vue repensée

Figure 8.5. *Vue originale et repensée de Canadian Musicians by ethnicity.*

La figure 8.5 montre comment cette refonte modifie la catégorie Wikipedia *Canadian people by ethnicity* et sa sous-catégorie *Canadian Musicians by ethnicity*. Les sous-catégories de *Canadian Musicians by ethnicity*, telles que

Black Canadian musicians, *Canadian musicians of Asian descent*, *Jewish Canadian musicians* et *Indigenous Canadian musicians*, sont listées — chacune appliquant des étiquettes raciales, ethniques ou religieuses de manière sélective. Cette étiquetage inégal renforce l'idée que seules certaines identités doivent être explicitement marquées, tandis que d'autres — notamment les identités blanches ou chrétiennes — restent non marquées.

Dans notre refonte, illustrée à la figure 8.5b, toutes les identités ethniques sont étiquetées de manière égale, ce qui permet de mettre en valeur les réalisations des minorités ethniques, et l'étiquetage du groupe majoritaire — les personnes blanches, dans ce cas — évite de marginaliser les membres des minorités ethniques. En standardisant le traitement de tous les groupes démographiques, notre cadre évite l'étiquetage sélectif, maintient des conventions de dénomination cohérentes et respecte la complexité des identités croisées. De plus, notre refonte résout le problème de la logique de catégorisation incohérente en appliquant une structure uniforme à tous les marqueurs d'identité — qu'ils soient raciaux, ethniques, religieux ou nationaux — garantissant que les catégories soient à la fois cohérentes et équitables.

8.2.3 Avantages du système automatisé

Création dynamique de sous-catégories : Les sous-catégories seront générées dynamiquement en fonction des attributs définis, sans biais ni intervention manuelle. Par exemple :

- Dans la catégorie des architectes, s'il existe des femmes architectes, elles apparaîtront dans la sous-catégorie "women Architects" sans nécessiter la création manuelle de la sous-catégorie.
- De même, dans la catégorie des infirmiers, si un homme infirmier est présent, le système créera automatiquement et remplira une sous-catégorie "Men Nurses".

Équité et atténuation des biais : Le système automatisé garantit l'équité en appliquant des règles de catégorisation cohérentes à tous les individus. Cette approche empêche la plateforme d'introduire de nouveaux biais ou d'exacerber les biais sociétaux existants. De plus, le recours à des données structurées réduit le risque que des décisions éditoriales subjectives influencent les désignations de catégories.

Amélioration de la mesure des biais : La nouvelle conception permet une meilleure compréhension des biais sociétaux en fournissant des données transparentes sur la manière dont les individus sont catégorisés. Par exemple, en analysant la répartition des catégories par genre, religion ou autres attributs, les chercheurs peuvent obtenir des informations sur la représentation et les biais dans le contenu de Wikipedia.

Scalabilité et efficacité : L'automatisation du processus de catégorisation réduit considérablement la charge de travail des contributeurs. Au lieu de passer du temps à organiser et éditer manuellement les catégories,

les éditeurs peuvent se concentrer sur l'amélioration de la qualité et de la précision du contenu.

Adaptabilité : Le système peut facilement intégrer de nouveaux attributs ou des modifications dans Wikidata sans nécessiter de mises à jour manuelles importantes.

8.3 Résumé

Dans ce chapitre, nous avons exploré les biais dans le système de catégorisation de Wikipedia, en nous concentrant sur les biais de genre et raciaux. Le genre et l'ethnicité sont largement reconnus comme des attributs *sensibles*, c'est-à-dire qu'ils définissent des groupes fréquemment discriminés. En analysant les relations catégorie-sous-catégorie de Wikipedia, nous avons identifié des schémas de biais et des incohérences structurelles qui compromettent l'objectif de la plateforme de fournir des informations équilibrées et équitables.

Le principal schéma de biais présent dans les catégories biographiques de Wikipedia est que les femmes et les minorités ethniques sont fréquemment étiquetées dans des contextes professionnels où un tel étiquetage est largement non pertinent. Par exemple, dans les catégories basées sur les professions telles que *Architects* ou *Chemists*, des sous-catégories basées sur le genre et l'ethnicité sont créées, par exemple *Women chemists* ou *Arab Chemists*. Dans de nombreux cas, ces catégories ne sont pas significatives

ou *notable*, dans le sens où les femmes ou les personnes arabes exercent l'architecture et la chimie de manière comparable aux personnes d'autres groupes démographiques. En revanche, les catégories ethniques peuvent être plus significatives pour des professions davantage influencées par la culture : on peut imaginer que *Arab chefs* ou *African-american musicians* présentent un intérêt encyclopédique en tant que groupes.

Bien que ces catégories explicites offrent une certaine visibilité aux membres de groupes sous-représentés, ce schéma, où les groupes majoritaires ne sont pas marqués, risque également de créer un message négatif selon lequel les réalisations professionnelles des membres des groupes sous-représentés sont principalement liées à leur genre ou à leur identité raciale plutôt qu'à leurs compétences, contributions ou expertise. Par exemple, cela peut suggérer qu'une femme astronaute est *notable* uniquement parce qu'elle fait partie des rares femmes dans la profession, plutôt que parce qu'elle est une astronaute compétente.

Les directives de Wikipedia reconnaissent ce risque dans une certaine mesure et recommandent que les biographies de femmes, de membres de minorités ethniques ou de personnes handicapées soient également catégorisées dans des catégories plus générales qui ne sont pas définies par leur genre, leur ethnie ou leur handicap. Comme le processus de catégorisation de Wikipedia est manuel, dans de nombreux cas cette directive n'est pas respectée, et puisque la catégorisation explicite est principalement effectuée pour les femmes et les minorités, ce problème affecte de manière dispropor-

tionnée les membres de ces groupes, entraînant une forme de préjudice de représentation(*representational harm*) [108].

Les résultats soulignent l'importance de développer des systèmes d'organisation des connaissances plus transparents et objectifs sur les plateformes collaboratives telles que Wikipedia. Nous avons présenté une proposition visant à automatiser le cadre de catégorisation des biographies en utilisant des données structurées provenant de Wikidata. En créant systématiquement des catégories pour toutes les valeurs d'attribut personnel, y compris celles identifiant les groupes majoritaires, les différents groupes sont mis sur un pied d'égalité, tout en maintenant la visibilité des groupes sous-représentés.

Notre prototype fonctionnel démontre la faisabilité de l'idée et montre que les biais sociétaux peuvent parfois être atténués en concevant mieux les systèmes d'information qui médiatisent la collaboration humaine.

Les recherches futures pourraient prolonger ce travail en développant les détails pour mettre pleinement en œuvre le système au sein de Wikipedia et en évaluant l'impact à long terme du cadre repensé sur la précision du contenu et l'engagement des utilisateurs. Cela fournirait des informations précieuses sur l'efficacité de ces interventions.

Chapitre 9

Conclusion et travaux futurs

Dans cette recherche, notre objectif général est de mieux comprendre comment les considérations d'équité (telles que définies dans le contexte de l'*équité algorithmique*) s'appliquent dans le contexte plus large des *machines sociales*.

L'équité n'a pas été étudiée de manière générale et systématique à travers les social machines, et lorsque des questions liées à l'équité (par exemple le biais de genre) ont été analysées dans des social machines spécifiques, les phénomènes problématiques ont été décrits à l'aide d'analyses et de métriques ad hoc, rendant difficile la comparaison des différents problèmes ou la généralisation de l'analyse au cadre plus large des social machines. Pour remédier à cette lacune, la thèse apporte trois contributions principales.

Premièrement, nous avons réalisé une revue systématique des biais affectant un ensemble diversifié de social machines, en catégorisant les phénomènes

nuisibles et en les reliant aux concepts d'équité établis dans les contextes algorithmiques. Deuxièmement, nous avons mené une étude de cas sur le biais de genre dans les biographies de Wikipedia, en appliquant les concepts d'équité algorithmique pour analyser à la fois la sous-représentation et la représentation biaisée. Troisièmement, nous avons exploré comment les interventions de conception peuvent atténuer les biais dans les social machines, en développant un prototype de système visant à repenser l'infrastructure de catégorisation de Wikipedia de manière plus cohérente et équitable.

Ensemble, ces contributions fournissent à la fois un cadre conceptuel pour comprendre l'équité dans les social machines et des démonstrations concrètes de la manière dont les biais peuvent être mesurés, analysés et atténués. Dans la suite, nous discutons en détail des objectifs et des contributions.

9.1 Contributions

Cette thèse a défini trois objectifs principaux, chacun examiné à travers des questions de recherche ciblées. Ci-dessous, nous passons en revue ces objectifs et mettons en évidence les résultats qui y répondent directement.

9.1.1 Contribution 1 : Revue systématique de l'équité dans les machines sociales

Objectif 1 : Mieux comprendre les problématiques liées à l'équité à travers les machines sociales, qu'elles aient ou non des causes algorithmiques clairement identifiables.

Pour atteindre cet objectif, nous avons réalisé une revue systématique des biais affectant un ensemble diversifié et représentatif de machines sociales. Grâce à cette approche, nous avons cherché à obtenir une vue d'ensemble de l'équité dans les machines sociales. Dans notre enquête, nous avons étudié la manière dont les concepts liés à l'équité sont appliqués dans ces contextes. Nous avons identifié et catégorisé les phénomènes décrits comme des biais envers des groupes démographiques identifiables, les cadrant de manière normative comme nuisibles avec des significations spécifiques et en référence aux concepts d'équité initialement définis pour les systèmes algorithmiques.

Nous avons constaté que les principaux phénomènes nuisibles que nous avons identifiés peuvent être décrits comme la sous-représentation ou la représentation biaisée des groupes défavorisés, causant un préjudice de représentation envers ces groupes, ainsi que des processus décisionnels entraînant un préjudice d'allocation ou de qualité de service envers les groupes défavorisés. Dans ce dernier cas, nous avons montré que des métriques d'équité initialement définies pour les systèmes algorithmiques telles que *demographic parity* ou *equalized odds* peuvent être utilisées pour décrire les biais en termes signifi-

catifs et génériques.

Notre revue a également révélé que des phénomènes similaires se produisent pour plusieurs groupes défavorisés, bien qu'ils aient été analysés principalement en fonction de leur impact sur les femmes. Nous émettons l'hypothèse que les biais de genre pourraient être plus répandus que d'autres types de biais (par exemple, les biais raciaux), mais surtout qu'ils sont probablement plus faciles à mesurer dans les données collectées à partir des social machines, en raison des indications explicites de genre dans le texte (par exemple, les pronoms). Cela suggérerait que les biais signalés comme affectant un groupe démographique pourraient servir d'indications pour rechercher d'autres biais lors de futurs audits des mêmes systèmes. De plus, notre analyse met en évidence des biais affectant des groupes qui ne peuvent être définis ou qui ne sont pertinents que dans le contexte des social machines : cela a des implications supplémentaires pour la conception des social machines : toute catégorisation des utilisateurs dans le contexte de la social machine crée implicitement un groupe qui peut être affecté par des biais. Bien que ce groupe ne soit pas explicitement protégé par les lois anti-discrimination, les corrélations entre l'appartenance à ce groupe et les groupes protégés peuvent entraîner des biais involontaires à l'égard de ces groupes protégés.

9.1.2 Contribution 2 : Étude de cas sur le biais de genre dans les biographies de Wikipedia

Objectif 2 : Comprendre le rôle que joue la social machine elle-même dans le biais — si elle se contente de refléter les biais sociétaux existants ou contribue à les produire ou les amplifier.

Pour atteindre cet objectif, nous avons revisité le problème du biais de genre dans les biographies de Wikipedia avec une nouvelle analyse fondée sur les concepts d'équité algorithmique. Notre recherche applique des analyses novatrices pour examiner si les femmes sont sous-représentées et/ou représentées de manière biaisée, et tente de démêler les biais sociaux existants des biais supplémentaires introduits par Wikipedia en tant que social machine.

Bien qu'il semble problématique que très peu de biographies Wikipedia concernent des femmes, cela peut s'expliquer en partie par le fait que les rôles que les femmes ont tendance à adopter dans la société (par choix ou non) sont moins susceptibles de les rendre "notables". Cependant, il est encore plus problématique que, même lorsque nous essayons de contrôler la notabilité, les femmes restent fortement sous-représentées.

Notre analyse d'un ensemble de données sur les biographies de représentants au niveau des États américains a montré que les hommes et les femmes ont presque autant de chances d'avoir une biographie : les taux sont de 89% pour les femmes et de 91% pour les hommes. Cela contraste fortement avec

la situation rapportée par Adams et al. [1] dans le monde académique : pour des niveaux comparables de notabilité, les hommes professeurs avaient plus du double de chances d'avoir une biographie par rapport à une femme professeure.

Une interprétation de cette divergence est que, dans un contexte politique, il peut exister un certain niveau de volonté institutionnelle pour que les représentants élus aient une biographie — ce qui s'appliquerait tant aux hommes qu'aux femmes — et qu'un critère clair de notabilité (être un représentant élu dans un cadre familier à de nombreux contributeurs de Wikipedia) signifie qu'il y aura moins d' "glass-ceiling effect" empêchant les femmes d'être considérées comme notables.

Concernant les représentations biaisées des femmes, la plupart des travaux précédents rapportant un *lexical bias* se basent sur des comptages de mots (ou de bigrammes), ce qui ne permet pas de capturer le contexte dans lequel les mots sont utilisés. Nous avons déterminé si la représentation dans les social machines reflète honnêtement notre société biaisée ou si ces systèmes socio-techniques génèrent leurs propres biais, similaires aux biais algorithmiques. Il est probable que les représentations biaisées des femmes découlent en partie des différences dans les "histoires de vie" entre hommes et femmes, et éliminer les biais dans la représentation des hommes et des femmes ne supprimera pas entièrement ces différences.

Nous avons montré que la performance des classificateurs et la distance entre les vecteurs d'*embedding* de documents peuvent capturer les différences

d'une manière quantifiable. Lorsque les biographies décrivent des personnes notables pour les mêmes raisons, leurs histoires de vie sont similaires et la capacité d'un classificateur à les différencier (hommes vs. femmes) diminue de manière significative. De telles métriques seront utiles pour surveiller les niveaux de biais au fil du temps et selon les contextes.

9.1.3 Contribution 3 : Interventions de conception pour l'atténuation des biais

Objectif 3 : Comprendre comment la conception d'une social machine peut être modifiée pour atténuer les biais.

Dans la plupart des cas, les phénomènes de biais résultent d'interactions complexes entre de nombreux utilisateurs d'un système informatique et ne peuvent pas être directement attribués à un composant algorithmique unique. Par conséquent, aucun algorithme spécifique ne peut être modifié ou remplacé pour éliminer ces biais. Cependant, les personnes impliquées dans la production des résultats problématiques interagissent selon des processus et des normes définis au sein de chaque social machine, et basent leurs décisions sur les informations rendues visibles par son infrastructure informatique. Par exemple, le système peut permettre aux utilisateurs de télécharger une photo d'eux-mêmes et autoriser ou non les autres à voir cette photo et à en déduire si la personne est un homme noir ou une femme blanche.

Cela implique que la conception de la social machine, incluant à la fois

son infrastructure informatique et les processus par lesquels les personnes interagissent, peut être modifiée afin d'atténuer les biais. Pour répondre à notre troisième objectif, nous avons exploré comment ces biais peuvent être réduits par des interventions de conception et avons développé et mis en œuvre un système de catégorisation repensé pour les biographies de Wikipedia.

Nous avons étudié les biais dans le système de catégorisation de Wikipedia, en nous concentrant sur les disparités de genre et raciales. En analysant la structure catégorie-sous-catégorie des entrées biographiques, nous avons observé un schéma où les femmes et les minorités ethniques sont disproportionnellement étiquetées selon des marqueurs d'identité — tels que "Women chemists" ou "Arab chemists" — notamment dans des contextes professionnels où ces distinctions sont souvent peu pertinentes. Cela contraste avec les groupes majoritaires (typiquement blancs, masculins, occidentaux), qui sont généralement inclus dans des catégories non marquées et par défaut comme "Chemists" ou "Architects", renforçant ainsi leur perception de normalité.

Bien que les catégories explicites basées sur l'identité puissent améliorer la visibilité des groupes sous-représentés, elles risquent également de suggérer que les individus sont notables principalement en raison de leur identité, plutôt que de leurs réalisations ou de leur expertise. Cette asymétrie crée un préjudice de représentation, perpétuant les stéréotypes et marginalisant certaines communautés. Bien que les directives de Wikipedia recommandent de placer les individus à la fois dans des catégories générales et des catégories spécifiques à l'identité, le caractère manuel du système entraîne souvent une

application incohérente, affectant de manière disproportionnée les femmes et les minorités.

Pour remédier à ces incohérences structurelles, nous avons introduit un système de catégorisation automatisé utilisant des données structurées provenant de Wikidata. Ce système assigne les individus à des catégories en fonction de leurs attributs personnels de manière cohérente et symétrique, incluant les catégories pour les groupes majoritaires lorsque cela est applicable. Le prototype fonctionnel démontre la faisabilité d'une telle refonte et met en évidence comment des changements réfléchis dans l'infrastructure numérique peuvent réduire les biais systémiques et promouvoir une représentation plus équitable sur des plateformes collaboratives comme Wikipedia.

9.2 Publications dérivées

Notre revue des biais dans Wikipedia, présentée dans les chapitres 4, 5 et 7, a été publiée dans l'article suivant, qui a reçu le **Prix du Meilleur Article** :

1. M. S. Damadi and A. Davoust. "Fairness in Socio-Technical Systems: A Case Study of Wikipedia". In: *Proceedings of the 29th International Conference, CollabTech.* 2023.

Ce travail a été étendu à une sélection représentative de social machines, fournissant une exploration complète des concepts d'équité dans les social machines, et a été publié dans l'article de journal suivant :

2. M. S. Damadi and A. Davoust. "Fairness in Social Machines: A Systematic Review." *Journal of Information, Communication and Ethics in Society,* 2025.

De plus, notre analyse d'un ensemble de données sur les biographies de représentants au niveau des États américains, mentionnée dans le chapitre 6, a été publiée sous forme d'article de conférence :

3. M. S. Damadi and A. Davoust. "Evaluating gender bias in Wikipedia using document embeddings". In: *Proceedings of the 36th canadian conference on artificial intelligence.* 2024.

Enfin, le dernier article s'appuie directement sur les analyses présentées dans les chapitres précédents, proposant un cadre de catégorisation automatisé basé sur les attributs, discuté au chapitre 8, afin de remédier aux incohérences structurelles et d'atténuer les biais dans le système de catégorisation de Wikipedia.

4. M. S. Damadi and A. Davoust. "A Redesign of Wikipedia's Categorization System to Reduce Systematic Bias". *submitted in CoopIS,* July 2025.

9.3 Travaux futurs

Bien que cette recherche apporte de nouvelles perspectives sur l'équité dans les social machines et démontre la faisabilité d'interventions de concep-

tion, plusieurs limites subsistent, ouvrant la voie à de futures explorations.

Premièrement, bien que notre revue ait adopté une approche systématique couvrant un large éventail de social machines, notre étude de cas détaillée et notre analyse empirique se sont concentrées principalement sur Wikipedia. Ce choix a été motivé par l'accessibilité de la plateforme, la richesse de ses données structurées et l'attention académique antérieure. Bien que nous ayons également examiné d'autres domaines — notamment les marchés en ligne, les plateformes de médias sociaux, les jeux en ligne multijoueurs et les mondes virtuels, ainsi que les systèmes de crowdsourcing — notre travail empirique s'est concentré sur la mesure et l'atténuation des biais au sein de Wikipedia. Les recherches futures pourraient prolonger ce travail en développant des méthodes pour mesurer systématiquement les biais sur d'autres plateformes (par exemple, TaskRabbit, Fiverr, Uber, Airbnb, Reddit, YouTube ou la révision par les pairs) et en concevant et évaluant des interventions visant à atténuer ces biais dans la pratique.

Deuxièmement, notre travail empirique s'est principalement centré sur la mesure et l'atténuation des biais de genre et, dans une moindre mesure, des biais raciaux. D'autres formes de biais — tels que ceux liés au handicap, à l'âge ou à l'origine géographique — restent peu explorées dans les social machines, en partie à cause de l'absence d'attributs explicites ou facilement mesurables dans les jeux de données disponibles. Les recherches futures pourraient explorer des moyens de détecter et d'atténuer ces formes de biais moins visibles, par exemple en exploitant des techniques avancées de traite-

ment automatique des langues capables de capturer des schémas nuancés de représentation au-delà des pronoms de genre ou des marqueurs d'identité explicites.

En élargissant le champ d'investigation au-delà de Wikipedia et en étendant la gamme de dimensions démographiques étudiées, les travaux futurs pourront s'appuyer sur nos contributions pour favoriser des social machines plus équitables, transparentes et inclusives.

Bibliographie

[1] Julia Adams, Hannah Brückner, and Cambria Naslund. Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". *Socius*, 5:2378023118823946, 2019.

[2] Rishi Ahuja and Ronan C Lyons. The silent treatment: discrimination against same-sex relations in the sharing economy. *Oxford Economic Papers*, 71(3):564–576, 2019.

[3] Hind Almerekhi, Supervised by Bernard J Jansen, and co-supervised by Haewoon Kwak. Investigating toxicity across multiple reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020*, pages 294–298, 2020.

[4] Guadalupe Alvarez, Aileen Oeberst, Ulrike Cress, and Laura Ferrari. Linguistic evidence of in-group bias in english and spanish wikipedia articles about international conflicts. *Discourse, Context & Media*, 35:100391, 2020.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2016.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacksawelcome@1230. 2016.

[7] David Bamman and Noah A Smith. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376, 2014.

[8] Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1):1–34, 2021.

[9] Arsh Banerjee, Pierce Maloney, and Christian Ronda. Unsupervised discovery of implicit gender bias: A new analysis of reddit and fitocracy. Technical report, Princeton University, 2023.

[10] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.

[11] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

[12] Pablo Beytía. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Companion Proceedings of the Web Conference 2020*, pages 806–810, 2020.

[13] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR, 2018.

[14] Marie Biolková, Tom Moore, Karen Schindler, Karl Swann, Andy Vail, Lindsay Flook, Helen Dick, Greg Fitzharris, Christopher A Price, and Norah Spears. Investigation of potential gender bias in the peer review system at reproduction. *Learned Publishing*, 36(1):25–30, 2023.

[15] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[16] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*, 2020.

[17] Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. 2018.

[18] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to

homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[19] Audrey L Brehm. Navigating the feminine in massively multiplayer online games: gender in world of warcraft. *Frontiers in psychology*, 4:903, 2013.

[20] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[21] Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

[22] Justine Cassell. Editing wars behind the scenes. *New York Times*, 2011.

[23] Miquel Centelles and Núria Ferran-Ferrer. Assessing knowledge organization systems from a gender perspective: Wikipedia taxonomy and wikidata ontologies. *Journal of Documentation*, 80(7):124–147, 2024.

[24] Kelsey C Chappetta and Joan M Barth. Gaming roles versus gender roles in online gameplay. *Information, Communication & Society*, 25(2):162–183, 2022.

[25] Zhenpeng Chen, Jie M Zhang, Max Hort, Federica Sarro, and Mark Harman. Fairness testing: A comprehensive survey and analysis of trends. *arXiv preprint arXiv:2207.10223*, 2022.

[26] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[27] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. Activity sensing in the wild: a field trial of ubifit garden. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1797–1806, 2008.

[28] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

[29] Mir Saeed Damadi and Alan Davoust. Fairness in socio-technical systems: A case study of wikipedia. In *Proceedings of the 29th International Conference, CollabTech*, 2023.

[30] Mir Saeed Damadi and Alan Davoust. Fairness in social machines: A systematic reviews. 2024.

[31] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[32] Paul B De Laat. The use of software tools and autonomous bots against vandalism: eroding wikipedia's moral order? *Ethics and Information Technology*, 17:175–188, 2015.

[33] Paul B de Laat. Profiling vandalism in wikipedia: A schauerian approach to justification. *Ethics and Information Technology*, 18:131–148, 2016.

[34] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[35] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics*, 9(2):1–22, 2017.

[36] Benjamin G Edelman and Michael Luca. Digital discrimination: The case of airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054), 2014.

[37] Siân Evans, Jacqueline Mabey, and Michael Mandiberg. Editing for equality: The outcomes of the art+ feminism wikipedia edit-a-thons. *Art Documentation: Journal of the Art Libraries Society of North America*, 34(2):194–203, 2015.

[38] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM conference on web science*, pages 87–96, 2019.

[39] Núria Ferran Ferrer, Miquel Centelles Velilla, Yessica Maciá Martínez, Juan-José Boté-Vericad, and Julià Minguillón. Dones de categoria: Análisi del biaix de génere a les categories de viquipèdia: Informe de diagnosi técnica, posicionament académic i proposta de millora del sistema d'organització del coneixement de viquipèdia. Technical report, Programa d'Igualtat de Génere de la Xarxa Vives d'Universitats, 2024.

[40] Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. Discovering and categorising language biases in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 140–151, 2021.

[41] Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635, 2022.

[42] Ann Finkbeiner. What i'm not going to do: Do media have to talk about family matters? *DoubleXScience*, 2013.

[43] Will Fleisher. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 480–490, 2021.

[44] Eitan Frachtenberg and Kelly S McConville. Metrics and methods in the evaluation of prestige bias in peer review: A case study in computer systems conferences. *Plos one*, 17(2):e0264131, 2022.

[45] Jon Froehlich, Tawanna Dillahunt, Predrag Klasnja, Jennifer Mankoff, Sunny Consolvo, Beverly Harrison, and James A Landay. Ubigreen: investigating a mobile tool for tracking and supporting green transportation habits. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1043–1052, 2009.

[46] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[47] Gege Gao, Aehong Min, and Patrick C Shih. Gendered design bias: gender differences of in-game character choice and playing style in league of legends. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 307–317, 2017.

[48] Sue Gardner. Nine reasons women don't edit wikipedia (in their own words). Sue Gardner's blog, 2011.

[49] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020.

[50] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through

robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA, 2019. Association for Computing Machinery.

[51] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[52] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial discrimination in transportation network companies. *Journal of Public Economics*, 190:104205, 2020.

[53] R Stuart Geiger. The lives of bots. *arXiv preprint arXiv:1810.09590*, 2018.

[54] Ruediger Glott and Rishab Ghosh. Analysis of wikipedia survey data. topic: Age and gender differences. *Retrieved November*, 14:2014, 2010.

[55] Mark Graham, Ralph K Straumann, and Bernie Hogan. Digital divisions of labor and informational magnetism: Mapping participation in wikipedia. *Annals of the Association of American Geographers*, 105(6):1158–1178, 2015.

[56] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[57] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933, 2017.

[58] Moritz Hardt and Solon Barocas. Fairness in machine learning. In *Neural Information Processing Symposium, Tutorials Track*, 2017.

[59] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[60] Jim Hendler and Tim Berners-Lee. From the semantic web to social machines: A research challenge for ai on the world wide web. *Artificial intelligence*, 174(2):156–161, 2010.

[61] Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

[62] James D Ivory. Still a man's game: Gender representation in online reviews of video games. *Mass communication & society*, 9(1):103–114, 2006.

[63] Shawn Janzen, Caroline Orr, and Sara-Jayne Terp. Cognitive security and resilience: A social ecological model of disinformation and other harms with applications to covid-19 vaccine information behaviors. 2022.

[64] Nicolas Jullien. What we know about wikipedia: A review of the literature analyzing the project (s). 2012.

[65] Venoo Kakar, Joel Voelz, Julia Wu, and Julisa Franco. The visible host: Does race guide airbnb rental rates in san francisco? *Journal of Housing Economics*, 40:25–40, 2018.

[66] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[67] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[68] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[69] Kyungwon Lee, Anne-Marie Hakstian, and Jerome D Williams. Creating a world where anyone can belong anywhere: Consumer equality in the sharing economy. *Journal of Business Research*, 130:221–231, 2021.

[70] Jun Li, Dennis Zhang, and Ruomeng Cui. A better way to fight discrimination in the sharing economy. *Harvard Business Review*, page 27, 2017.

[71] Randall M Livingstone. Population automation: An interview with wikipedia bot pioneer ram-man. *First Monday*, 2016.

[72] Anya Marchenko. The impact of host race and gender on prices on airbnb. *Journal of Housing Economics*, 46:101635, 2019.

[73] Sara Marjanovic, Karolina Stańczak, and Isabelle Augenstein. Quantifying gender biases towards politicians on reddit. *PloS one*, 17(10):e0274317, 2022.

[74] Brian Martin. Persistent bias on wikipedia: Methods and responses. *Social Science Computer Review*, 36(3):379–388, 2018.

[75] Paolo Massa and Federico Scrinzi. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–9, 2012.

[76] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[77] Nathalie D McKenzie, Raymond Liu, Alden V Chiu, Mariana Chavez-MacGregor, Dean Frohlich, Sarfraz Ahmad, and Carolyn B Hendricks.

Exploring bias in scientific peer review: an asco initiative. *JCO oncology practice*, 18(12):791–799, 2022.

[78] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[79] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. "the sum of all human knowledge": A systematic review of scholarly research on the content of w ikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245, 2015.

[80] Arvind Mewada, Rupesh Kumar Dewang, Paritosh Goldar, and Sushil Kumar Maurya. Sentibert: A novel approach for fake review detection incorporating sentiment features with contextual features. In *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, pages 230–235, 2023.

[81] Santiago M Mola-Velasco. Wikipedia vandalism detection through machine learning: Feature review and new proposals: Lab report for pan at clef 2010. *arXiv preprint arXiv:1210.5560*, 2012.

[82] Danielle A Morris-O'Connor, Andreas Strotmann, and Dangzhi Zhao. The colonization of wikipedia: evidence from characteristic editing be-

haviors of warring camps. *Journal of Documentation*, (ahead-of-print), 2022.

[83] FA Nielsen. Wikipedia research and tools: Review and comments. *Working draft. Informatics and Mathematical Modelling, Technical University of Denmark. URL: http://www2. imm. dtu. dk/pubdb/views/publication_details. php*, 2019.

[84] Mathias Wullum Nielsen, Christine Friis Baker, Emer Brady, Michael Bang Petersen, and Jens Peter Andersen. Weak evidence of country-and institution-related status bias in the peer review of abstracts. *Elife*, 10:e64561, 2021.

[85] Pablo Noriega and Dave De Jonge. Electronic institutions: the ei/eide framework. In *Social coordination frameworks for social technical systems*, pages 47–76. Springer, 2016.

[86] Kamala O Norris. Gender stereotypes, aggression, and computer games: An online survey of women. *Cyberpsychology & Behavior*, 7(6):714–727, 2004.

[87] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12):2381–2403, 2014.

[88] Cory Ondrejka. Aviators, moguls, fashionistas and barons: Economics and ownership in second life. *Available at SSRN 614663*, 2004.

[89] Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39. IEEE, 2020.

[90] Luca Oneto, Anna Siri, Gianvittorio Luria, and Davide Anguita. Dropout prediction at university of genoa: a privacy preserving data driven approach. In *ESANN*, 2017.

[91] Zacharoula Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, 17(4):49–64, 2014.

[92] Petros Papapanagiotou, Alan Davoust, Dave Murray-Rust, Areti Manataki, Max Van Kleek, Nigel Shadbolt, and Dave Robertson. Social machines for all. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1208–1212, 2018.

[93] Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490, 2021.

[94] Nebojša Ratković and Ivana Madžarević. Women's participation in wikipedia: Cross-border balkan perspective. *Área Abierta*, 21(2):237–253, 2021.

[95] Charu Rawat. An ethical reflection on user privacy and transparency of algorithmic blocking systems in the wikipedia community.

[96] Joseph Reagle and Lauren Rhue. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21, 2011.

[97] Stephanie Rennick, Melanie Clinton, Elena Ioannidou, Liana Oh, Charlotte Clooney, Edward Healy, and Seán G Roberts. Gender bias in video game dialogue. *Royal Society Open Science*, 10(5):221095, 2023.

[98] Lauren Rhue. An overview of crowd-based markets and racial discrimination. 2018.

[99] David Robertson and Fausto Giunchiglia. Programming the social computer. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120379, 2013.

[100] Richard Rogers, Emina Sendijarevic, et al. Neutral or national point of view? a comparison of srebrenica articles across wikipedia's language versions, 2012.

[101] Günter Ropohl. Philosophy of socio-technical systems. *Society for Philosophy and Technology Quarterly Electronic Journal*, 4(3):186–194, 1999.

[102] Shrikant Saxena and Shweta Jain. Exploring and mitigating gender bias in recommender systems with explicit feedback. *arXiv preprint arXiv:2112.02530*, 2021.

[103] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68, 2019.

[104] Anna Severin, Joao Martins, Rachel Heyard, François Delavy, Anne Jorstad, and Matthias Egger. Gender and other potential biases in peer review: cross-sectional analysis of 38 250 external peer review reports. *BMJ open*, 10(8):e035058, 2020.

[105] Nigel Shadbolt, Kieron O'Hara, David De Roure, and Wendy Hall. *The theory and practice of social machines*. Springer, 2019.

[106] Nigel Shadbolt, Max Van Kleek, and Reuben Binns. The rise of social machines: the development of a human/digital ecosystem. *IEEE Consumer Electronics Magazine*, 5(2):106–111, 2016.

[107] Nigel R Shadbolt, Daniel A Smith, Elena Simperl, Max Van Kleek, Yang Yang, and Wendy Hall. Towards a classification framework for social machines. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 905–912, 2013.

[108] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. Sociotechnical harms: Scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791*, 2022.

[109] Kao Si, Yiwei Li, Chao Ma, and Feng Guo. Affiliation bias in peer review and the gender gap. *Research Policy*, 52(7):104797, 2023.

[110] Riyaz Sikora and Liangjun You. Effect of reputation mechanisms and ratings biases on traders' behavior in online marketplaces. *Journal of organizational computing and electronic commerce*, 24(1):58–73, 2014.

[111] Judith Simon, Pak Hang Wong, and Gernot Rieder. Algorithmic bias and the value sensitive design approach. *Internet Policy Review*, 9(4):1–16, 2020.

[112] Flaminio Squazzoni, Giangiacomo Bravo, Mike Farjam, Ana Marusic, Bahar Mehmani, Michael Willis, Aliaksandr Birukou, Pierpaolo Dondio, and Francisco Grimaldo. Peer review and gender bias: A study on 145 scholarly journals. *Science advances*, 7(2):eabd0299, 2021.

[113] Jean-François Staszak. Other/otherness. international encyclopedia of human geography. *Strange Times, My Dear: The PEN Anthology of Contemporary Iranian Literature*, 2008.

[114] Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*, 2021.

[115] Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology*, 73(6):811–819, 2022.

[116] Steven Tadelis. The economics of reputation and feedback systems in e-commerce marketplaces. *IEEE Internet Computing*, 20(1):12–19, 2015.

[117] Steven Tadelis. Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340, 2016.

[118] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in neural information processing systems*, 32, 2019.

[119] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. Effects of algorithmic flagging on fairness: Quasi-experimental evidence from wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021.

[120] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3):1–40, 2017.

[121] Francesca Tripodi. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, page 14614448211023772, 2021.

[122] Francesca Tripodi. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, 25(7):1687–1707, 2023.

[123] Milena Tsvetkova, Taha Yasseri, Eric T Meyer, J Brian Pickering, Vegard Engen, Paul Walland, Marika Lüders, Asbjørn Følstad, and George Bravos. Understanding human-machine networks: a cross-disciplinary survey. *ACM Computing Surveys (CSUR)*, 50(1):1–35, 2017.

[124] Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. Digital discrimination in sharing economy a requirements engineering perspective. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 204–214. IEEE, 2020.

[125] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

[126] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463, 2015.

[127] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24, 2016.

[128] Dmitri Williams, Nicole Martins, Mia Consalvo, and James D Ivory. The virtual census: Representations of gender, race and age in video games. *New media & society*, 11(5):815–834, 2009.

[129] Zena Worku, Taryn Bipat, David W McDonald, and Mark Zachry. Exploring systematic bias through article deletions on wikipedia from a behavioral perspective. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–22, 2020.

[130] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, 2020.

[131] Hong Xie and John CS Lui. Incentive mechanism and rating system design for crowdsourcing systems: Analysis, tradeoffs and inference. *IEEE Transactions on Services Computing*, 11(1):90–102, 2016.

[132] Shlomit Yanisky-Ravid and Amy Mittelman. Gender biases in cyberspace: A two-stage model, the new arena of wikipedia and other websites. *Fordham Intell. Prop. Media & Ent. LJ*, 26:381, 2015.

[133] Amber Young, Ari D Wigdor, and Gerald Kane. It's not what you think: Gender bias in information about fortune 1000 ceos on wikipedia. 2016.

[134] Amber G Young, Ariel D Wigdor, and Gerald C Kane. The gender bias tug-of-war in a co-creation community: Core-periphery tension on wikipedia. *Journal of Management Information Systems*, 37(4):1047–1072, 2020.

[135] Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Training fair classifiers. In *AISTATS'17: 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[136] M Zandpour. The gender gap on wikipedia. 2020.

[137] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[138] Lei Zheng, Christopher M Albano, Neev M Vora, Feng Mai, and Jeffrey V Nickerson. The roles bots play in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019.

Annexe A : Articles sur Wikipédia

[1] Julia Adams, Hannah Brückner, and Cambria Naslund. Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". *Socius*, 5:2378023118823946, 2019.

[2] Kouatzin Aguilar-Morales, Gustavo Aguirre-Suarez, Brigham Bowles, Angel Lee, and Van Charles Lansingh. Wikipedia, friend or foe regarding information on diabetic retinopathy? a content analysis in the world's leading 19 languages. *PLoS One*, 16(10):e0258246, 2021.

[3] Guadalupe Alvarez, Aileen Oeberst, Ulrike Cress, and Laura Ferrari. Linguistic evidence of in-group bias in english and spanish wikipedia articles about international conflicts. *Discourse, Context & Media*, 35:100391, 2020.

[4] Samuel Baltz. Reducing bias in wikipedia's coverage of political scientists. *PS: Political Science & Politics*, 55(2):439–444, 2022.

[5] Pablo Beytía. The positioning matters: Estimating geographical bias in the multilingual record of biographies on wikipedia. In *Companion Proceedings of the Web Conference 2020*, pages 806–810, 2020.

[6] Pablo Beytía and Hans-Peter Müller. Towards a digital reflexive sociology: Using wikipedia's biographical repository as a reflexive tool. *Poetics*, 95:101732, 2022.

[7] Pablo Beytía Reyes and Claudia Wagner. Visibility layers: a framework for systematising the gender gap in wikipedia content. *Internet Policy Review*, 11(1):1–22, 2022.

[8] Carwil Bjork-James. New maps for an inclusive wikipedia: decolonial scholarship and strategies to counter systemic bias. *New Review of Hypermedia and Multimedia*, 27(3):207–228, 2021.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[10] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. The gender gap in wikipedia talk pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[11] Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

[12] Florencia Claes and Luis Deltell. Wikipedia in spanish. behaviour of the spanish-speaking community in relation to collaborative work on the internet. 2021.

[13] Sanmay Das, Allen Lavoie, and Malik Magdon-Ismail. Manipulation among the arbiters of collective intelligence: How wikipedia administrators mold public opinion. *ACM Transactions on the Web (TWEB)*, 10(4):1–25, 2016.

[14] Paul B De Laat. The use of software tools and autonomous bots against vandalism: eroding wikipedia's moral order? *Ethics and Information Technology*, 17:175–188, 2015.

[15] Paul B de Laat. Profiling vandalism in wikipedia: A schauerian approach to justification. *Ethics and Information Technology*, 18:131–148, 2016.

[16] Hannah Devinney, Anton Eklund, Igor Ryazanov, and Jingwen Cai. Developing a multilingual corpus of wikipedia biographies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 285–294, 2023.

[17] Pierpaolo Dondio, Niccolò Casnici, and Flaminio Squazzoni. Unpacking the structure of knowledge diffusion in wikipedia: Local biases, noble prizes and the wisdom of crowds. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 17–25, 2015.

[18] Angela Fan and Claire Gardent. Generating full length wikipedia biographies: The impact of gender bias on the retrieval-based generation of women biographies. *arXiv preprint arXiv:2204.05879*, 2022.

[19] Núria Ferran-Ferrer, Juan-José Boté-Vericad, and Julià Minguillón. Wikipedia gender gap: a scoping review. *Profesional de la información/Information Professional*, 32(6), 2023.

[20] Núria Ferran-Ferrer, Patricia Castellanos-Pineda, Julià Minguillón, and Julio Meneses. The gender gap on the spanish wikipedia: Listening to the voices of women editors. *Profesional de la Informacion*, 30(5), 2021.

[21] Núria Ferran-Ferrer, Marc Miquel-Ribé, Julio Meneses, and Julià Minguillón. The gender perspective in wikipedia: A content and participation challenge. In *Companion Proceedings of the Web Conference 2022*, pages 1319–1323, 2022.

[22] Anjalie Field, Chan Young Park, Kevin Z Lin, and Yulia Tsvetkov. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635, 2022.

[23] Fabian Flöck, Denny Vrandečić, and Elena Simperl. Towards a diversity-minded wikipedia. In *Proceedings of the 3rd International Web Science Conference*, pages 1–8, 2011.

[24] Heather Ford, Tamson Pietsch, and Kelly Tall. Gender and the invisibility of care on wikipedia. *Big Data & Society*, 10(2):20539517231210276, 2023.

[25] Ruediger Glott and Rishab Ghosh. Analysis of wikipedia survey data. topic: Age and gender differences. *Retrieved November*, 14:2014, 2010.

[26] Jan Grabowski and Shira Klein. Wikipedia's intentional distortion of the history of the holocaust. *The Journal of Holocaust Research*, 37(2):133–190, 2023.

[27] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174, 2015.

[28] Shane Greenstein and Feng Zhu. Do experts or crowd-based models produce more bias? evidence from encyclopædia britannica and wikipedia. *Mis Quarterly*, 2018.

[29] Sneh Gupta and Kulveen Trehan. Twitter reacts to absence of women on wikipedia: a mixed-methods analysis of# visiblewikiwomen campaign. *Media Asia*, 49(2):130–154, 2022.

[30] Karl Gustafsson. International reconciliation on the internet? ontological security, attribution and the construction of war memory narratives in wikipedia. *International Relations*, 34(1):3–24, 2020.

[31] Scott A Hale. Multilinguals and wikipedia editing. In *Proceedings of the 2014 ACM conference on Web science*, pages 99–108, 2014.

[32] Aaron Halfaker and R Stuart Geiger. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–37, 2020.

[33] Toni Hermoso Pulido. Simple wikidata analysis for tracking and improving biographies in catalan wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 582–583, 2021.

[34] Nina Hood and Allison Littlejohn. Hacking history: Redressing gender inequities on wikipedia through an editathon. *International Review of Research in Open and Distributed Learning*, 19(5), 2018.

[35] Christoph Hube. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 717–721, 2017.

[36] María Sefidari Huici. Equidad de conocimiento y sesgos: Un análisis cuantitativo del contenido destacado en la portada de wikipedia. *IC Revista Científica de Información y Comunicación*, (19):141–163, 2022.

[37] Nicolas Jullien. What we know about wikipedia: A review of the literature analyzing the project (s). *Available at SSRN 2053597*, 2012.

[38] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, 2007.

[39] Max Klein and Piotr Konieczny. Wikipedia in the world of global gender inequality indices: What the biography gender gap is measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*, pages 1–2, 2015.

[40] Piotr Konieczny. Decision making in the self-evolved collegiate court: Wikipedia's arbitration committee and its implications for self-governance and judiciary in cyberspace. *International Sociology*, 32(6):755–774, 2017.

[41] Mackenzie Emily Lemieux, Rebecca Zhang, and Francesca Tripodi. "too soon" to count? how gender and race cloud notability considerations on wikipedia. *Big Data & Society*, 10(1):20539517231165490, 2023.

[42] Ang Li, Rosta Farzan, and Claudia Lopez. Collaboration in editing world wide news in four wikipedia language communities. In *X Latin American Conference on Human Computer Interaction*, pages 1–4, 2021.

[43] Shlomit Aharoni Lir. Strangers in a seemingly open-to-all website: the gender bias in wikipedia. *Equality, Diversity and Inclusion: An International Journal*, 40(7):801–818, 2021.

[44] Randall M Livingstone. Places on the map and in the cloud: Representations of locality and geography in wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 211–212, 2011.

[45] Brian Martin. Persistent bias on wikipedia: Methods and responses. *Social Science Computer Review*, 36(3):379–388, 2018.

[46] Carlos Martinez-Ortiz, Marijn Koolen, Floor Buschenhenke, and Karina van Dalen-Oskam. Beyond the book: Linking books to wikipedia. In *2015 IEEE 11th International Conference on e-Science*, pages 12–21. IEEE, 2015.

[47] Franziska Martini. Notable enough? the questioning of women's biographies on wikipedia. *Feminist Media Studies*, pages 1–17, 2023.

[48] Paolo Massa and Federico Scrinzi. Exploring linguistic points of view of wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 213–214, 2011.

[49] Paolo Massa and Federico Scrinzi. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–9, 2012.

[50] Julie McDonough Dolmaya. Expanding the sum of all human knowledge: Wikipedia, translation and linguistic justice. *The Translator*, 23(2):143–157, 2017.

[51] Zachary J McDowell and Matthew A Vetter. *Wikipedia and the Representation of Reality*. Taylor & Francis, 2022.

[52] Cristina Menghini, Aris Anagnostopoulos, and Eli Upfal. Wikipedia polarization and its effects on navigation paths. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 6154–6156. IEEE, 2019.

[53] Amanda Menking and Jon Rosenberg. Wp: Not, wp: Npov, and other stories wikipedia tells us: A feminist critique of wikipedia's epistemology. *Science, Technology, & Human Values*, 46(3):455–479, 2021.

[54] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. "the sum of all human knowledge": A systematic

review of scholarly research on the content of w ikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245, 2015.

[55] Julià Minguillón, Julio Meneses, Eduard Aibar, Núria Ferran-Ferrer, and Sergi Fàbregues. Exploring the gender gap in the spanish wikipedia: Differences in engagement and editing practices. *Plos one*, 16(2):e0246702, 2021.

[56] Yukta Misra, Bhawna Dahiya, Sejal Rathi, and Rishabh Kaushal. Bias detection in indian narratives on wikipedia. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2023.

[57] Jonathan T Morgan, Robert M Mason, and Karine Nahon. Negotiating cultural values in social media: A case study from wikipedia. In *2012 45th Hawaii International Conference on System Sciences*, pages 3490–3499. IEEE, 2012.

[58] Danielle A Morris-O'Connor, Andreas Strotmann, and Dangzhi Zhao. The colonization of wikipedia: evidence from characteristic editing behaviors of warring camps. *Journal of Documentation*, 79(3):784–810, 2023.

[59] Danielle Morris-O'Connor, Andreas Strotmann, and Dangzhi Zhao. Editorial behaviors for biasing wikipedia articles. *Proceedings of the*

Association for Information Science and Technology, 59(1):226–234, 2022.

[60] Aileen Oeberst, Ina von der Beck, Christina Matschke, Toni Alexander Ihme, and Ulrike Cress. Collectively biased representations of the past: Ingroup bias in wikipedia articles about intergroup conflicts. *British Journal of Social Psychology*, 59(4):791–818, 2020.

[61] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. Wikipedia in the eyes of its beholders: A systematic review of scholarly research on wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, 65(12):2381–2403, 2014.

[62] Chan Young Park, Xinru Yan, Anjalie Field, and Yulia Tsvetkov. Multilingual contextual affective analysis of lgbt people portrayals in wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 479–490, 2021.

[63] Nebojša Ratković and Ivana Madžarević. Women's participation in wikipedia: Cross-border balkan perspective. *Area Abierta*, 21(2), 2021.

[64] Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. Analysing timelines of national histories across wikipedia editions: A comparative computational approach. In *Pro-*

ceedings of the International AAAI Conference on Web and Social Media, volume 11, pages 210–219, 2017.

[65] Nigel Shadbolt, Kieron O'Hara, David De Roure, and Wendy Hall. *The theory and practice of social machines*. Springer, 2019.

[66] Nigel Shadbolt, Max Van Kleek, and Reuben Binns. The rise of social machines: the development of a human/digital ecosystem. *IEEE Consumer Electronics Magazine*, 5(2):106–111, 2016.

[67] Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. Wikibio: a semantic resource for the intersectional analysis of biographical events. *arXiv preprint arXiv:2306.09505*, 2023.

[68] Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on wikipedia. *arXiv preprint arXiv:2106.01601*, 2021.

[69] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. Effects of algorithmic flagging on fairness: quasi-experimental evidence from wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021.

[70] Claire Timperley. The subversive potential of wikipedia: A resource for diversifying political science content online. *PS: Political Science & Politics*, 53(3):556–560, 2020.

[71] Francesca Tripodi. Ms. categorized: Gender, notability, and inequality on wikipedia. *New Media & Society*, 25(7):1687–1707, 2023.

[72] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463, 2015.

[73] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5:1–24, 2016.

[74] Tom Willaert and Guido Roumans. Nitpicking online knowledge representations of governmental leadership. the case of belgian prime ministers in wikipedia and wikidata. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 30(1):1–41, 2020.

[75] Zena Worku, Taryn Bipat, David W McDonald, and Mark Zachry. Exploring systematic bias through article deletions on wikipedia from a behavioral perspective. In *Proceedings of the 16th International Symposium on Open Collaboration*, pages 1–22, 2020.

[76] Linsi Xia, Naomi Yamashita, and Toru Ishida. Analysis on multilingual discussion for wikipedia translation. In *2011 Second International Conference on Culture and Computing*, pages 104–109. IEEE, 2011.

[77] Puyu Yang and Giovanni Colavizza. Polarization and reliability of news sources in wikipedia. *arXiv preprint arXiv:2210.16065*, 2022.

[78] SpoerriA YasseriT and KertészJ GrahamM. 2014the most controversial topics in wikipedia: a multilingual and geographical analysis. inglobal wikipedia: international and cross-cultural issues in online collaboration.

[79] Zining Ye, Xinran Yuan, Shaurya Gaur, Aaron Halfaker, Jodi Forlizzi, and Haiyi Zhu. Wikipedia ores explorer: Visualizing trade-offs for designing applications with machine learning api. In *Designing Interactive Systems Conference 2021*, pages 1554–1565, 2021.

[80] Amber Young, Ari D Wigdor, and Gerald Kane. It's not what you think: Gender bias in information about fortune 1000 ceos on wikipedia. 2016.

[81] Amber G Young, Ariel D Wigdor, and Gerald C Kane. The gender bias tug-of-war in a co-creation community: Core-periphery tension on wikipedia. *Journal of Management Information Systems*, 37(4):1047–1072, 2020.

[82] Olga Zagovora, Fabian Flöck, and Claudia Wagner. "(weitergeleitet von journalistin)" the gendered presentation of professions on wikipedia. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 83–92, 2017.

[83] Dangzhi Zhao and Andreas Strotmann. Editorial practices and systematic conscious bias on wikipedia: An initial test with articles on traditional chinese medicine. In *ISSI*, pages 1985–1990, 2019.

[84] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media*, 29:100205, 2022.

[85] Xiang Zheng, Jiajing Chen, Erjia Yan, and Chaoqun Ni. Gender and country biases in wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology*, 74(2):219–233, 2023.

[86] Yiwei Zhou, Elena Demidova, and Alexandra I Cristea. Who likes me more? analysing entity-centric language-specific bias in multilingual wikipedia. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 750–757, 2016.

Annexe B : Articles sur les marchés en ligne

[1] Rishi Ahuja and Ronan C Lyons. The silent treatment: discrimination against same-sex relations in the sharing economy. *Oxford Economic Papers*, 71(3):564–576, 2019.

[2] Sihem Amer-Yahia, Shady Elbassuoni, Ahmad Ghizzawi, Ria Mae Borromeo, Emilie Hoareau, and Philippe Mulhem. Fairness in online jobs:{A} case study on taskrabbit and google. In *International Conference on Extending Database Technologies (EDBT)*, 2020.

[3] Shahar Ayal, Daphna Bar-Haim, and Moran Ofir. Behavioral biases in peer-to-peer (p2p) lending. *Behavioral finance: The coming of age*, pages 367–400, 2018.

[4] Janis Cloos. Employer review platforms–do the rating environment and platform design affect the informativeness of reviews? theory, evidence, and suggestions. *mrev management revue*, 32(3):152–181, 2021.

[5] Ruomeng Cui, Jun Li, and Dennis J Zhang. Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on airbnb. *Management Science*, 66(3):1071–1094, 2020.

[6] Raymond Fisman and Michael Luca. Fixing discrimination in online marketplaces. *Harvard Business Review*, 94(12):88–95, 2016.

[7] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1914–1933, 2017.

[8] Clare J Hooper, Brian Bailey, Hugh Glaser, and James Hendler. Social machines in practice: Solutions, stakeholders and scopes. In *Proceedings of the 8th ACM Conference on Web Science*, pages 156–160, 2016.

[9] Venoo Kakar, Joel Voelz, Julia Wu, and Julisa Franco. The visible host: Does race guide airbnb rental rates in san francisco? *Journal of Housing Economics*, 40:25–40, 2018.

[10] Qing Ke. Sharing means renting? an entire-marketplace analysis of airbnb. In *Proceedings of the 2017 ACM on web science conference*, pages 131–139, 2017.

[11] Kyungwon Lee, Anne-Marie Hakstian, and Jerome D Williams. Creating a world where anyone can belong anywhere: Consumer equality in the sharing economy. *Journal of Business Research*, 130:221–231, 2021.

[12] Ning F Ma, Veronica A Rivera, Zheng Yao, and Dongwook Yoon. "brush it off": How women workers manage and cope with bias and harassment in gender-agnostic gig platforms. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2022.

[13] Anya Marchenko. The impact of host race and gender on prices on airbnb. *Journal of Housing Economics*, 46:101635, 2019.

[14] Arvind Mewada, Rupesh Kumar Dewang, Paritosh Goldar, and Sushil Kumar Maurya. Sentibert: A novel approach for fake review detection incorporating sentiment features with contextual features. In *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, pages 230–235, 2023.

[15] Lauren Rhue. An overview of crowd-based markets and racial discrimination. 2018.

[16] Riyaz Sikora and Liangjun You. Effect of reputation mechanisms and ratings biases on traders' behavior in online marketplaces. *Journal of organizational computing and electronic commerce*, 24(1):58–73, 2014.

[17] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. Does fair ranking improve minority outcomes? understanding the interplay of human

and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 989–999, 2021.

[18] Steven Tadelis. The economics of reputation and feedback systems in e-commerce marketplaces. *IEEE Internet Computing*, 20(1):12–19, 2015.

[19] Steven Tadelis. Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340, 2016.

[20] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3):1–40, 2017.

[21] David Tsurel, Michael Doron, Alexander Nus, Arnon Dagan, Ido Guy, and Dafna Shahaf. E-commerce dispute resolution prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1465–1474, 2020.

[22] Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. Digital discrimination in sharing economy a requirements engineering perspective. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 204–214. IEEE, 2020.

[23] Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. Analysis of non-discrimination policies in the sharing economy. In *2021 IEEE Inter-*

national Conference on Software Maintenance and Evolution (ICSME), pages 104–113. IEEE, 2021.

[24] Miroslav Tushev, Fahimeh Ebrahimi, and Anas Mahmoud. A systematic literature review of anti-discrimination design strategies in the digital sharing economy. *IEEE Transactions on Software Engineering*, 48(12):5148–5157, 2022.

[25] Ruhai Wu and Chun Qiu. When karma strikes back: A model of seller manipulation of consumer reviews in an online marketplace. *Journal of Business Research*, 155:113316, 2023.

[26] Hong Xie and John CS Lui. Incentive mechanism and rating system design for crowdsourcing systems: Analysis, tradeoffs and inference. *IEEE Transactions on Services Computing*, 11(1):90–102, 2016.

Annexe C : Articles sur les réseaux sociaux

[1] Lütfiye Seda Mut Altın and Horacio Saggion. A novel corpus for automated sexism identification on social media. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 10–15, 2024.

[2] Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021.

[3] Thales Bertaglia, Katarina Bartekova, Rinske Jongma, Stephen Mccarthy, and Adriana Iamnitchi. Sexism in focus: An annotated dataset of youtube comments for gender bias research. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 22–28, 2023.

[4] Mert Can Cakmak, Nitin Agarwal, Obianuju Okeke, Ugochukwu

Onyepunuka, and Billy Spann. Evaluating bias and fairness in ai: An analysis of youtube's recommendation algorithm and its impact on geopolitical discourse. 2024.

[5] Nathaniel Ming Curran. Discrimination in the gig economy: The experiences of black online english teachers. *Language and Education*, 37(2):171–185, 2023.

[6] Tyler McCoy Gay, Oluyemi TO Farinu, and Monisha Issano Jackson. "from all sides": Black-asian reddit communities identify and expand experiences of the multiracial microaggression taxonomy. *Social Sciences*, 11(4):168, 2022.

[7] Miriam-Linnea Hale and André Melzer. Girls vs. men-the prevalence of gender-biased language in popular youtube videos. In *International Conference on Entertainment Computing*, pages 176–186. Springer, 2023.

[8] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. Measuring misinformation in video search platforms: An audit study on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020.

[9] Rachna Jain, Ashish Kumar, Anand Nayyar, Kritika Dewan, Rishika Garg, Shatakshi Raman, and Sahil Ganguly. Explaining sentiment analysis results on social media texts through visualization. *Multimedia Tools and Applications*, 82(15):22613–22629, 2023.

[10] Jae Yeon Kim, Jaeung Sim, and Daegon Cho. Identity and status: When counterspeech increases hate speech reporting and why. *Information Systems Frontiers*, 25(5):1683–1694, 2023.

[11] Jack LaViolette and Bernie Hogan. Using platform signals for distinguishing discourses: The case of men's rights and men's liberation on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 323–334, 2019.

[12] Sara Marjanovic, Karolina Stańczak, and Isabelle Augenstein. Quantifying gender biases towards politicians on reddit. *PloS one*, 17(10):e0274317, 2022.

[13] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr, and Virgílio Almeida. Analyzing right-wing youtube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM conference on web science*, pages 323–332, 2018.

[14] Rutul D Patel, Priya Dave, Justin Loloi, Samantha Freeman, Nathan Feiertag, Mustufa Babar, and Kara L Watts. Gender bias in youtube videos describing common urology conditions. *Urology*, 169:256–266, 2022.

[15] Maria Priestley and Alex Mesoudi. Do online voting patterns reflect evolved features of human cognition? an exploratory empirical investigation. *PloS one*, 10(6):e0129703, 2015.

[16] Julian A Rodriguez. Lgbtq incorporated: Youtube and the management of diversity. *Journal of Homosexuality*, 70(9):1807–1828, 2023.

[17] Samantha R Rosenthal, Kelsey A Gately, Samantha G Philippe, Allyson B Baker, Monique P Dawes, and Jennifer E Swanberg. Everyday discrimination and mental health among sexual and gender minority adults: The moderating role of self-compassion. *Psychology of Sexual Orientation and Gender Diversity*, 2023.

[18] Joni Salminen, Soon-Gyo Jung, and Bernard J Jansen. Detecting demographic bias in automatically generated personas. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[19] Rupak Sarkar and Ashiqur R KhudaBukhsh. Are chess discussions racist? an adversarial hate speech data set (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15881–15882, 2021.

[20] Shrikant Saxena and Shweta Jain. Exploring and mitigating gender bias in recommender systems with explicit feedback. *arXiv preprint arXiv:2112.02530*, 2021.

[21] Shrikant Saxena and Shweta Jain. Exploring and mitigating gender bias in book recommender systems with explicit feedback. *Journal of Intelligent Information Systems*, pages 1–22, 2024.

[22] Rishi Shounak, Sayantan Roy, Vivek Kumar, and Vivek Tiwari. Reddit comment toxicity score prediction through bert via transformer based architecture. In *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0353–0358. IEEE, 2022.

[23] Sunny Shrestha. Design, determination, and evaluation of gender-based bias mitigation techniques for music recommender systems. Master's thesis, University of Denver, 2023.

[24] Almog Simchon, William J Brady, and Jay J Van Bavel. Troll and divide: the language of online polarization. *PNAS nexus*, 1(1):pgac019, 2022.

[25] Marudan Sivagurunathan, David M Walton, Tara Packham, Richard G Booth, and Joy C MacDermid. Discourses around male ipv related systemic biases on reddit. *Journal of Interpersonal Violence*, 37(19-20):NP17834–NP17859, 2022.

[26] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59, 2017.

[27] Mike Thelwall and Emma Stuart. She's reddit: A source of statistically significant gendered interest information? *Information processing & management*, 56(4):1543–1558, 2019.

[28] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. *arXiv preprint arXiv:2307.09209*, 2023.

[29] Qunfang Wu, Louisa Kayah Williams, Ellen Simpson, and Bryan Semaan. Conversations about crime: Re-enforcing and fighting against platformed racism on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–38, 2022.

[30] Diyi Yang and Scott Counts. Understanding self-narration of personally experienced racism on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[31] Zizhong Zhang, Haixin Mu, and Sida Huang. Playing to save sisters: how female gaming communities foster social support within different cultural contexts. *Journal of Broadcasting & Electronic Media*, 67(5):693–713, 2023.

Annexe D : Articles sur les jeux en ligne

[1] Audrey L Brehm. Navigating the feminine in massively multiplayer online games: gender in world of warcraft. *Frontiers in psychology*, 4:903, 2013.

[2] Kelsey C Chappetta and Joan M Barth. Gaming roles versus gender roles in online gameplay. *Information, Communication & Society*, 25(2):162–183, 2022.

[3] Gege Gao, Aehong Min, and Patrick C Shih. Gendered design bias: gender differences of in-game character choice and playing style in league of legends. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 307–317, 2017.

[4] James D Ivory. Still a man's game: Gender representation in online reviews of video games. *Mass communication & society*, 9(1):103–114, 2006.

[5] Daniel Madden, Yuxuan Liu, Haowei Yu, Mustafa Feyyaz Sonbudak, Giovanni M Troiano, and Casper Harteveld. "why are you playing games? you are a girl!": Exploring gender biases in esports. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15, 2021.

[6] Kamala O Norris. Gender stereotypes, aggression, and computer games: An online survey of women. *Cyberpsychology & Behavior*, 7(6):714–727, 2004.

[7] Archana Parashar and S Niranjana Varma. Gendered harassment and bias in online gaming: Experiences of female gamers. In *Cyberfeminism and Gender Violence in Social Media*, pages 226–244. IGI Global, 2023.

[8] Stephanie Rennick, Melanie Clinton, Elena Ioannidou, Liana Oh, Charlotte Clooney, Edward Healy, and Seán G Roberts. Gender bias in video game dialogue. *Royal Society Open Science*, 10(5):221095, 2023.

[9] Cuihua Shen, Rabindra Ratan, Y Dora Cai, and Alex Leavitt. Do men advance faster than women? debunking the gender performance gap in two massively multiplayer online games. *Journal of Computer-Mediated Communication*, 21(4):312–329, 2016.

[10] Jing Sun. Gender in chinese video games. *The International Encyclopedia of Gender, Media, and Communication*, pages 1–5, 2020.

[11] Kellie Vella, Madison Klarkowski, Selen Turkay, and Daniel Johnson. Making friends in online games: gender differences and designing for greater social connectedness. *Behaviour & Information Technology*, 39(8):917–934, 2020.

[12] Dmitri Williams, Nicole Martins, Mia Consalvo, and James D Ivory. The virtual census: Representations of gender, race and age in video games. *New media & society*, 11(5):815–834, 2009.

Annexe E: Articles sur Crowdsourcing

[1] Sawsan Abdel-Razig and Halah Ibrahim. Outsider bias: how your name influences the peer review process, 2023.

[2] Michael Allen. No gender bias in peer review, claims study. *Physics World*, 34(2):10i, 2021.

[3] G Bazoukis. Prestige bias—an old, untreated enemy of the peer-review process. *Hippokratia*, 24(2):94, 2020.

[4] Marie Biolková, Tom Moore, Karen Schindler, Karl Swann, Andy Vail, Lindsay Flook, Helen Dick, Greg Fitzharris, Christopher A Price, and Norah Spears. Investigation of potential gender bias in the peer review system at reproduction. *Learned Publishing*, 36(1):25–30, 2023.

[5] Ángel Borja. Is there gender bias in the peer-review process in several elsevier's marine journals? *Marine pollution bulletin*, 96(1-2):1–2, 2015.

[6] Lutz Bornmann and Hans-Dieter Daniel. Selection of research fellowship recipients by committee peer review. reliability, fairness and predictive validity of board of trustees' decisions. *Scientometrics*, 63(2):297–320, 2005.

[7] Lutz Bornmann and Hans-Dieter Daniel. Reviewer and editor biases in journal peer review: an investigation of manuscript refereeing at angewandte chemie international edition. *Research evaluation*, 18(4):262–272, 2009.

[8] Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics*, 2(4):280–287, 2008.

[9] Michael Callaham. Gender bias and peer review: annals seeks greater diversity. *Annals of Emergency Medicine*, 74(6):736–739, 2019.

[10] Cibele Cássia-Silva, Barbbara Silva Rocha, Luisa Fernanda Liévano-Latorre, Mariane Brom Sobreiro, and Luisa Maria Diele-Viegas. Overcoming the gender bias in ecology and evolution: is the double-anonymized peer review an effective pathway over time? *PeerJ*, 11:e15186, 2023.

[11] Leif Engqvist and Joachim G Frommen. Double-blind peer review and gender publication bias. *Animal Behaviour*, 76(3), 2008.

[12] Eitan Frachtenberg and Kelly S McConville. Metrics and methods in the evaluation of prestige bias in peer review: A case study in computer systems conferences. *Plos one*, 17(2):e0264131, 2022.

[13] Julie R Gilbert, Elaine S Williams, and George D Lundberg. Is there gender bias in jama's peer review process? *Jama*, 272(2):139–142, 1994.

[14] Markus Helmer, Manuel Schottdorf, Andreas Neef, and Demian Battaglia. Gender bias in scholarly peer review. *Elife*, 6:e21718, 2017.

[15] Antonio J Herrera. Language bias discredits the peer-review system. *Nature*, 397(6719):467–467, 1999.

[16] Faye Holst, Kim Eggleton, and Simon Harris. Transparency versus anonymity: which is better to eliminate bias in peer review? 2022.

[17] Adina R Kern-Goldberger, Richard James, Vincenzo Berghella, and Emily S Miller. The impact of double-blind peer review on gender bias in scientific publishing: a systematic review. *American journal of obstetrics and gynecology*, 227(1):43–50, 2022.

[18] John A Lane and David J Linden. Is there gender bias in the peer review process at journal of neurophysiology?, 2009.

[19] Emaad Manzoor and Nihar B Shah. Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4767–4775, 2021.

[20] Nathalie D McKenzie, Raymond Liu, Alden V Chiu, Mariana Chavez-MacGregor, Dean Frohlich, Sarfraz Ahmad, and Carolyn B Hendricks. Exploring bias in scientific peer review: an asco initiative. *JCO oncology practice*, 18(12):791–799, 2022.

[21] Mathias Wullum Nielsen, Christine Friis Baker, Emer Brady, Michael Bang Petersen, and Jens Peter Andersen. Weak evidence of country-and institution-related status bias in the peer review of abstracts. *Elife*, 10:e64561, 2021.

[22] Tobias Opthof, Ruben Coronel, and Michiel J Janse. The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular research*, 56(3):339–346, 2002.

[23] Tobias Opthof, Ruben Coronel, and Michiel J Janse. Submissions, impact factor, reviewer's recommendations and geographical bias within the peer review system (1997–2002) focus on germany, 2002.

[24] Ginger Pinholster. Journals and funders confront implicit bias in peer review. *Science*, 352(6289):1067–1068, 2016.

[25] David B Resnik and Elise M Smith. Bias and groupthink in science's peer-review system. *Groupthink in science: Greed, pathological altruism, ideology, competition, and culture*, pages 99–113, 2020.

[26] Erin Ross. Gender bias distorts peer review across fields. *Nature*, 21685(10.1038), 2017.

[27] Helene Schiffbaenker, Peter van den Besselaar, Florian Holzinger, Charlie Mom, and Claartje Vinkenburg. Gender bias in peer review panels:– "the elephant in the room" 1. In *Inequalities and the Paradigm of Excellence in Academia*, pages 109–128. Routledge, 2022.

[28] Anna Severin, Joao Martins, Rachel Heyard, François Delavy, Anne Jorstad, and Matthias Egger. Gender and other potential biases in peer review: cross-sectional analysis of 38 250 external peer review reports. *BMJ open*, 10(8):e035058, 2020.

[29] Kao Si, Yiwei Li, Chao Ma, and Feng Guo. Affiliation bias in peer review and the gender gap. *Research Policy*, 52(7):104797, 2023.

[30] José Osvaldo De Sordi and Manuel Antonio Meireles. Halo effect in peer review: Exploring the possibility of bias associated with the feeling of belonging to a group. *Perspectivas em Ciência da Informação*, 24:96–132, 2020.

[31] Flaminio Squazzoni, Giangiacomo Bravo, Mike Farjam, Ana Marusic, Bahar Mehmani, Michael Willis, Aliaksandr Birukou, Pierpaolo Dondio, and Francisco Grimaldo. Peer review and gender bias: A study on 145 scholarly journals. *Science advances*, 7(2):eabd0299, 2021.

[32] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology*, 73(6):811–819, 2022.

[34] Mike Thelwall, Liz Allen, Eleanor-Rose Papas, Zena Nyakoojo, and Verena Weigert. Does the use of open, non-anonymous peer review in scholarly publishing introduce bias? evidence from the f1000research post-publication open peer review publishing model. *Journal of information science*, 47(6):809–820, 2021.

[35] Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[36] Samuel P Trethewey. Medical misinformation on social media: cognitive bias, pseudo-peer review, and the good intentions hypothesis. *Circulation*, 140(14):1131–1133, 2019.

[37] Andrea C Tricco, Sonia M Thomas, Jesmin Antony, Patricia Rios, Reid Robson, Reena Pattani, Marco Ghassemi, Shannon Sullivan, Inthuja Selvaratnam, Cara Tannenbaum, et al. Strategies to prevent or reduce

gender bias in peer review of research grants: a rapid scoping review. *PLoS One*, 12(1):e0169718, 2017.

[38] Luiz Victorino, Ronaldo Pilati, and Alexandre Linhares. Priming and prejudice: The bias effect of origin information on peer review, judgment and evaluation. *Avances en Psicología Latinoamericana*, 37(1):169–178, 2019.

[39] Dario von Wedel, Rico A Schmitt, Moritz Thiele, Raphael Leuner, Denys Shay, Simone Redaelli, and Maximilian S Schaefer. Affiliation bias in peer review of abstracts by a large language model. *JAMA*, 331(3):252–253, 2024.

[40] Richard Walker, Beatriz Barros, Ricardo Conejo, Konrad Neumann, and Martin Telefont. Personal attributes of authors and reviewers, social bias and the outcomes of peer review: a case study. *F1000Research*, 4, 2015.

[41] Steven W Witt. Can journals overcome bias and make the peer review process more inclusive?, 2019.