Caractérisation de l'inclusivité des systèmes de vision par ordinateur basés sur l'apprentissage profond pour les pays du Sud

Théophile Bayet

Sorbonne Université

Université Cheikh Anta Diop de Dakar

École Doctorale Informatique, Télécommunications et Electronique (EDITE de Paris - ED130)

École Doctorale Mathématiques et Informatique, (EDMI)

Unité de Modélisation Mathématique et Informatique des Systèmes Complexes (UMMISCO)

Unité Mixte de Recherche de Sorbonne Université et du Centre National de la Recherche Scientifique (UMR 7606 Sorbonne Université - CNRS) (LIP6)

Caractérisation de l'inclusivité des systèmes de vision par ordinateur basés sur l'apprentissage profond pour les pays du Sud

Par Théophile Bayet

Thèse de Doctorat en Informatique

Dirigée par Pr. Christophe Denis et Pr. Alassane Bah

Encadrée par Pr. Jean-Daniel Zucker

Présentée et soutenue publiquement le mercredi 19 juin 2024 à Paris devant un jury présidé par **Nicolas Maudet** et composé de :

Céline Hudelot, *Professeur, Centrale Supélec, Rapporteur*
Désiré Sidibé, *Professeur, Université Paris Saclay, Rapporteur*
Nicolas Maudet, *Professeur, SU, Examinateur*
Mandicou Ba, *Professeur, UCAD/ESP, Examinateur*
Alassane Bah, *Professeur, UCAD/UMMISCO, Directeur de thèse (Sud)*
Christophe Denis, *Professeur, SU, Directeur de thèse (Nord)*
Jean-Daniel Zucker, *Directeur de laboratoire, IRD/SU/UMMISCO, Encadrant (Nord)*

Remerciements

Je tiens avant tout à remercier mes encadrants, pour m'avoir accompagné, guidé et soutenu tout au long de ces trois années. Alassane Bah pour sa disponibilité sans faille lors de mes nombreuses galères administratives, Christophe Denis pour l'accompagnement régulier durant ces années, et Jean-Daniel pour sa précieuse expertise et ses remarques pertinentes tout au long de mes pérégrinations. Je vous remercie aussi pour la confiance que vous m'avez accordé, et la liberté que j'ai eu durant ces années – dans mon travail comme dans ma manière de travailler. Cette confiance m'a été précieuse et m'a permis de m'épanouir. Je remercie aussi les rapporteurs Céline Hudelot et Désiré Sidibé, qui ont relu et analysé mon manuscrit en un temps record, pour leurs questions et leurs remarques passionnantes et pertinentes. Merci à Mandicou Ba et à Nicolas Maudet, président du jury, d'avoir accepté d'examiner cette thèse en participant au jury.

J'ai passé mes cinq dernières années dans l'Unité de Modélisation Mathématique et Informatique des Systèmes Complexes (UMMISCO), et j'ai eu durant ces années le plaisir de faire la connaissance de nombreux doctorants, post-doc, chercheurs et ingénieurs qui m'ont beaucoup apporté. Je remercie Edi, Youssef, Anane, Nicolas, Gaspard, Alex, Théo (2), Fabien, Jean-Michel, Caouis, nos échanges et réunions sont précieux. Je remercie tout particulièrement Ahmad et Christophe, vous êtes aujourd'hui plus des amis que des collègues. Il y a peu de personnes qui savent transmettre une passion aussi bien que Tim et Yann, et travailler avec vous a été un vrai bonheur, un jeu presque, tant les discussions et les missions étaient plaisantes et passionnantes. Vous rencontrer à Dakar a été un véritable plaisir.

Une thèse en co-tutelle donne l'occasion de s'aventurer dans les méandres administratifs de plusieurs pays, et j'ai eu maintes occasions de me perdre dans des labyrinthes de démarches et formulaires. Un énorme merci à Elisabeth Pereira (UMMISCO), Dany Richard (EDITE) et Amina Niang (EDMI) pour leur aide et leur accompagnement durant ces galères. Bien s'entourer vaut parfois milles mails.

Et pour la thèse aussi il faut bien s'entourer, c'est une aventure qui a ses hauts et ses bas, ses crêtes et ses creux. Et pour chacun d'entre eux, merci à tous ceux qui ont été présents, particulièrement mes co-doctorants, le projet Bagels (on va la faire cette révolution), Arthur et François, faut s'accrocher, on arrive au bout, malgré les épreuves, les déserts à traverser.

J'ai eu la chance, enfin, de beaucoup barouder, vagabonder, de m'échouer sur des tas de rivages, grâce à tout ceux qui ont bien voulu de moi chez eux. La liste est longue, merci à Thomas, Julien, Justine, Alexandre, Basile, Carole, Thibault, Quentin, Coco, Valentin, Ronan, Milou, Nathan, Hannah, Paul, Tom, Inès, Veronica, Max, Loris, Virgile, Betty, Othman, Emma, Souley, Jimmy, Alexia, et à tous ceux qui ce sont proposés de me recevoir, grâce à vous je peux vivre ma vie de la manière dont je l'entends, et ça me réchauffe toujours le coeur.

Enfin, je dois beaucoup à ma famille, pour le soutien, pour l'hébergement aussi, et pour l'indéfectible amour qui nous lie malgré la distance. Merci de continuer à m'accepter comme je suis, même si je suis un peu difficile à suivre parfois.

Abstract

Modern global changes, such as climate change and the sixth mass extinction, are profoundly disrupting our societies and ecosystems. New technologies, including machine learning, are both aggravating factors and potential means of mitigating the challenges posed by these changes. In 2015, the United Nations established the Sustainable Development Goals to assess the ecological impact and the risks for populations, revealing that the countries of the South are the furthest from achieving the objectives of this framework. Countries with limited digital infrastructures deploy less machine learning models, encountering a problem of context shift due to inconsistency between training and deployment data. In computer vision, this shift is exacerbated by the absence of data from southern countries in the training sets, leading to reduced model performance.

In this thesis, we bridge the gap between artificial intelligence for sustainable science and the inclusivity of computer vision systems. We show how previous approaches to demonstrating the lack of inclusivity of current vision systems have overlooked important aspects of the problem, such as the formalisation of geographical bias and the metrics that reflect its impact. This has led us to propose a protocol for formalising bias, based on the identification of a source, a type and an impact in order to characterise it. This protocol has been implemented for geographical bias, initially on synthetic data. As known synthetic databases do not have a geographical bias, we create synthetic datasets with geographical biases, inspired by previous synthetic modifications of the MNIST database. We use these to test the implementation of our protocol and demonstrate its usefulness. We then experiment with the protocol on real data for characterising western bias in vision systems, and find that the results obtained are different from those expected, going against observations in previous academic work. We carry out a visual analysis of these results at different levels of granularity in an attempt to understand them and to propose possible themes for future research. In the end, we highlight the presence of concomitant biases, elements that make up the geographical bias but have different impacts that the main entity. These concomitant biases prevent the characterisation of the geographical bias by influencing the predictions of the models.

We therefore show how the problem of characterising geographical bias is more complex than it might at first appear, what the current pitfalls are and what avenues are being pursued to remedy the problems encountered. Overall, we offer the scientific community tools to better understand the problems of deploying models in developing countries, in order to better understand the challenges of these deployments for applications in sustainable science.

Keywords: Machine Learning · Bias · Computer Vision · Global South · Geographical Shift · Context Shift

Résumé

Les changements mondiaux modernes, tels que le changement climatique et la sixième extinction de masse, perturbent profondément nos sociétés et nos écosystèmes. Les nouvelles technologies, notamment l'apprentissage machine, sont à la fois facteurs aggravants et moyens potentiels d'atténuation des défis que posent ces changements. En 2015, les Nations Unies ont établi les Objectifs de Développement Durable pour évaluer l'impact écologique et les risques pour les populations, révélant que les pays du Sud sont les plus éloignés des objectifs de ce cadre. Les pays avec des infrastructures numériques limitées déploient moins les modèles d'apprentissage machine, rencontrant un problème de glissement de contexte dû à l'incohérence entre les données d'entraînement et de déploiement. En vision par ordinateur, ce glissement est exacerbé par l'absence de données des pays du Sud dans les ensembles d'entraînement, conduisant à une performance réduite des modèles dans ces contextes.
Dans cette thèse, nous faisons le pont entre l'intelligence artificielle au service de la science soutenable et l'inclusivité des systèmes de vision par ordinateur. Nous montrons comment les approches qui ont précédé à la notre pour démontrer le manque d'inclusivité des systèmes de vision actuels ont fait l'impasse sur des points importants de la problématique, comme la formalisation du biais géographique et des métriques qui témoignent de son impact. Cela nous amène à proposer un protocole pour la formalisation des biais, qui se base sur l'identification d'une source, d'un type et d'un impact pour la caractérisation de ce dernier. Ce protocole est implémenté pour le biais géographique, en premier lieu sur des données synthétiques. Les bases de données synthétiques connues ne possédant pas de biais géographique, nous nous inspirons des modifications de la base de données MNIST pour créer des bases de données synthétiques comportant des biais géographiques. Nous utilisons ces derniers pour tester l'implémentation de notre protocole et démontrer son utilité. Nous expérimentons ensuite le protocole sur des données réelles pour la caractérisation du biais occidental dans les systèmes de vision, et constatons que les résultats obtenus sont différents de ceux attendus, allant à l'encontre des observations dans les précédents travaux académiques. Nous procédons à une analyse visuelle à différents niveaux de granularité de ces résultats pour tenter de les comprendre et proposer des théories les expliquant. Au final, nous mettons en avant la présence de biais concomitants, des éléments composant le biais géographique mais ayant des impacts différenciés avec l'entité principale. Ces biais concomitants empêchent la caractérisation du biais géographique en influençant les prédictions des modèles.
Nous montrons donc comment la problématique de la caractérisation du biais géographique se révèle plus complexe qu'elle ne peut le paraître au premier abord, quels sont les écueils actuels et quelles pistes sont privilégiées pour remédier aux problèmes rencontrés. Globalement, nous proposons à la communauté scientifique des outils pour mieux appréhender les problématiques de déploiement de modèles dans les pays du Sud, afin de mieux comprendre les enjeux de ces déploiements pour des applications en science soutenable.
Mots clés: Apprentissage Machine · Biais · Vision par Ordinateur · Contexte Sud · Glissement Géographique · Glissement de Contexte

Contents

Index of Figures

Index of Tables

Chapter 1

Introduction

Chapter Summary

Résumé du chapitre

Ce chapitre d'introduction de la thèse commence par présenter le contexte socio-technique qui a motivé le choix du projet de thèse. Cette présentation du contexte est suivie de définitions des termes utilisés, précisant certains termes qui seront largement utilisés dans cette thèse tout en ayant des définitions multiples et parfois discutées. Ce chapitre introduit ensuite les objectifs de la thèse, avant de présenter les contributions et publications qui ont accompagné les travaux présentés.

1.1 Contexte et motivation

Cette thèse est le fruit d'une collaboration entre la France et le Sénégal, financée par le Programme Doctoral International Modélisation et Systèmes Complexes (PDI MSC). Elle se déroule ainsi en même temps sous la tutelle de Sorbonne Université à Paris, par le laboratoire du LIP6, et de l'Université Cheikh Anta Diop à Dakar, par la branche sénégalaise de l'Unité Mixte Internationale de Modélisation Mathématique et Informatique des Systèmes COmplexes (UMI UMMISCO). Elle a débuté en février 2021, au milieu du contexte sanitaire particulier imposé par l'épidémie de covid-19.

Il est aujourd'hui de notoriété publique que la planète subit des transformations majeures à une vitesse sans précédent, et que ces transformations sont très fortement corrélées aux activités humaines. Parmi les transformations en cours, l'augmentation de la concentration en CO2 dans l'atmosphère et

les océans, le changement climatique et la chute de la biodiversité vont avoir un impact important sur les modes de vies des différentes sociétés humaines sur l'ensemble du globe.

L'augmentation de la concentration en CO2 est due à l'utilisation d'énergies fossiles par les activités humaines. La combustion de ces énergies fossiles dégage du CO2 qui va être stocké majoritairement par les deux puits de carbone que sont l'océan et l'atmosphère. Cette augmentation de la concentration en CO2 a diverses conséquences : augmentation de l'effet de serre, acidification des océans pour les plus impactantes. L'augmentation de l'effet de serre modifie le bilan énergétique de la planète et mène à une hausse globale des températures de l'atmosphère et des océans, et à la fonte des glaciers, de la banquise et des calottes glaciaires. L'acidification des océans quant à elle mène à la mort de nombreux organismes nécessaires au maintien de la biodiversité sous-marine.

Le changement climatique a des conséquences drastiques sur tous les êtres vivants. Ce changement affecte les biotopes du monde entier, provoquant des migrations de populations cherchant des conditions plus adaptées à leurs survies. Il provoque aussi une augmentation des évènements climatiques extrêmes, comme les sécheresses, les inondations, les cyclones, les feux. La combinaison de tous ces effets affecte les conditions de survies des différentes espèces vivantes et menace d'extinction une grande partie du monde vivant.

La chute de la biodiversité découle des précédentes transformations. Toute la biodiversitée est impactée, faunes et flores, terrestres, maritimes et aériennes. La perte de la biodiversité entraîne une perte de résilience des écosystèmes, qui sont alors moins à même de résister aux effets du changement climatique et de l'augmentation des concentrations de CO2. Tous ces effets sont embriqués dans des boucles de rétro-actions positives qui empêchent tout retour en arrière. Il est donc urgent de se concentrer sur ces problématiques et d'en faire des sujets primordiaux.

Néanmoins, toutes les populations ne sont pas touchées de la même manière. Les coraux par exemple, sont déjà victimes de l'acidification et de l'augmentation de la température des océans, et risquent de perdre 99% de leur population dans les prochaines années, tandis que d'autres espèces sont pour le moment moins touchées par les changements dans leur environnement. Au niveau humain, ce sont les populations qui sont les moins technologiquement développés qui sont les plus impactés par le changement climatique. Les inégalités structurelles et la vulnérabilité des populations est un facteur important dans l'exposition aux conséquences des changements en cours[1]. L'accroissement des inégalités est donc un produit dérivé de la situation environnementale, en même temps qu'un facteur aggravant les risques encourus par des populations déjà vulnérables.

Ces transformations structurelles de notre environnement sont accompagnées d'évolutions toutes aussi rapides : technologiques et culturelles. Les technologies numériques, introduites à la fin du XX^{eme} siècle et démocratisées dès le début du XXI^{eme} siècle sont deux décennies plus tard adoptées de manière massives et à l'origine d'une révolution technique mondiale. De par sa capacité d'interconnexion, cette transformation technique permet une transformation culturelle, en accélérant la vitesse de partage des informations, des contacts et du savoir. L'avènement de l'ère numérique a vu le développement des supports et infrastructures pour supporter la multiplication des terminaux, et des produits dérivés de la conquête numérique : les données. Amélioration de la vitesse de connexion permettant un partage plus rapide de données, de la capacité de stockage de ces données, et des capacités de traitement de ces données, le numérique a très vite rhytmé avec donnée. C'est ce qui a donné naissance à la science des données, dont le rôle est d'organiser et de traiter ces données de manière à en obtenir un produit utile. L'apprentissage machine, et dans sa lignée l'apprentissage profond, se sont vite imposées comme étant les techniques majeures en science des données, capables d'identifier et d'utiliser des corrélations statistiques complexes dans de grandes bases de données. Ces nouvelles technologies ont permis de réaliser des avancées particulières dans tous les milieux scientifiques, et notamment dans la compréhension de notre environnement et des conséquences de nos modes de vie sur ce dernier. Ces technologies sont par ailleurs avancées comme les technologies les plus à même de lutter contre le changement climatique.

[1] Voir https://www.un.org/sustainabledevelopment/blog/2016/10/report-inequalities-exacerbate-climate-impacts-on-poor/

Il est donc critique de développer ces technologies dans les milieux les plus vulnérables. Économiquement et géographiquement, ces milieux sont englobés par l'appellation "Sud", en opposition au "Nord". Les technologies de traitement des données sont pourtant pour le moment peu déployées au Sud ; les contraintes sont nombreuses. Les infrastructures numériques y sont peu adaptées à la génération et la collecte des données massives nécessaires aux techniques d'apprentissage profond ; les capacités de stockage et de calcul y sont inférieures à celles des pays du Nord. De fait, le développement de modèles d'intelligence artificielle dans ces pays reste marginal.

Le déploiement de ces technologies au Sud est aussi freiné par la difficulté d'accès aux données dans ces contextes. Les grands volumes de données nécessaire à l'apprentissage des corrélations statistiques d'un contexte ne sont pas disponibles au Sud, et la plupart des grandes bases de données constituées par les communautés scientifiques ne contiennent que très peu voir aucune donnée issues de contextes du Sud. Les modèles ayant appris sur les bases de données constituées au Nord sont ainsi peu adaptés aux pays du Sud, à cause de disparités dans les distributions des données ; ce phénomène est appelé "glissement de contexte". Il entraîne une baisse des performances des modèles dans les pays du Sud.

Ces modèles ne peuvent donc pas être considérés inclusifs, c'est à dire permettant d'inclure toutes les populations du globe. De tels modèles présenteraient des caractéristiques et performances géographiquement invariables. Hors, de nombreux modèles déployés par des entreprises et gouvernements ont été pointés du doigt pour leur tendance à être biaisés à l'encontre de certaines minorités. Ces biais sont un témoin d'un manque de prise en compte de ces minorités dans le développement des modèles, et témoignent d'un accord difficile entre inclusivité et développement de modèles.

Le défi auquel se confrontent aujourd'hui les chercheurs en informatique est d'articuler le développement de ces technologies au Sud avec une science qui soit soutenable pour le futur. Il s'avère que les technologies numériques ont des coûts environnementaux particuliers : pollutions liées à la production des terminaux, énergie nécessaire à la production, au stockage, au partage et au traitement des données, et problématiques éthiques et sociales liées à la collecte et à l'utilisation des données. Ces défis sont regroupés dans un pan de recherche que l'on appelle "science soutenable". Les chercheurs y interrogent les méthodes utilisées pour atteindre un objectif de recherche, et proposent des méthodologies pour évaluer l'impact de la recherche, ou pallier aux effets négatifs induits par certaines pratiques scientifiques.

Cette thèse s'articule autour de la nécessité d'articuler ces notions ensemble et de la volonté de participer à l'effort pour un futur soutenable et désirable. Elle a pour objectif de contribuer à la lutte contre le changement climatique, et s'inscrit dans un processus de science qui se veut soutenable et réfléchie. Elle s'inscrit aussi dans un effort d'étendre les possibilités offertes par l' apprentissage profond à des zones qui sont encore peut concernées par le développement de ces technologies, malgré le besoin pressant de lutter contre les effets du changement climatique dans les contextes concernés.

Dans un soucis d'être capable de mieux déployer les modèles de vision par ordinateur dans les pays émergeants, on étudie dans cette thèse une méthodologie permettant de témoigner de l'inclusivité d'un modèle. On s'intéresse tout particulièrement à la caractérisation des performances des modèles de vision par ordinateur dans les différents contextes géographiques du monde, et aux méthodes pour rendre ces modèles plus inclusifs.

1.2 Définitions préliminaires

Malgré le caractère informatique de cette thèse, celle-ci se rattache à une réalité sociale particulière. Nous nous devons donc de fixer les termes qui seront utilisés dans la suite afin d'éviter toute incompréhension.

1.2.1 Termes socio-géographiques

Il n'y a pas de définition scientifique de ce qui est occidental et ce qui ne l'est pas. L'occident est un concept culturel vague, qui a évolué au cours de l'histoire, de l'empire romain à l'expansion

de l'Europe et de ses colonies, jusqu'à la scission provoquée par la seconde guerre mondiale. Aujourd'hui, l'occident fait majoritairement référence à la culture qui domine les Etat-Unis et l'Europe, opposée aux autres influences mondiales.

Cette culture n'a pas de frontière géographique particulière. Elle peut toucher plus certains zones pourtant géographiquement éloignées des Etat-Unis ou de l'Europe, pour des raisons historique ou politiques. Ainsi, on considère généralement que l'Australie et la Nouvelle-Zélande font partie du bloc occidental. Dans le reste de la thèse, nous considérerons l'Amérique du Nord, l'Europe, l'Australie et la Nouvelle Zélande quand nous évoquerons l'occident.

Les appellations Sud et Nord font quant à elles référence à une classification économique des pays utilisée à partir des années 1980 et basées sur l'Indice de Développement Humain (IDH) et le Produit Intérieur Brut (PIB) par habitant. Les pays situés dans la moitié Sud du Globe étaient caractérisés par un IDH et un PIB par habitant plus faible que leurs homologues au Nord, ce qui a donné naissance à ces deux termes. Ces appellations sont désuettes, mais permettent d'opposer facilement les pays technologiquement développés et leader sur le plan de l'IA aux pays qui aujourd'hui ne sont pas acteur de ce développement. Nous l'utiliserons donc à des fins utiles, pour simplifier les mentions des différentes parties du globe.

Ces définitions ne sont pas parfaites et manquent de précision. Elles se concentrent sur des disparités géographiques, économiques et politiques, ce qui implique un certain nombre de biais. Abebe et al. (2021) souligne que cette définition simpliste ne prend pas en considération la complexité de l'écosystème mondial. Pourtant, cette vue simpliste peut se montrer utile pour promouvoir l'équité d'accès aux performances des modèles d'IA. Des travaux futurs pourraient inclure une reprise de ces définitions et une découpe plus fine que la grossière répartition effectuée ici pour les besoins de nos travaux.

1.2.2 L'inclusivité d'un modèle

L'inclusivité d'un modèle se traduit par le caractère inclusif de ce dernier. Il n'y a pas de définition du caractère inclusif d'un modèle, et ce caractère est en fait multiples.

Dans le cas des modèles mis en demeure dans les travaux de Wilson et al. (2019), les performances des modèles dépendaient de la couleur de peau des personnes dans les images soumises à évaluation. On décrit ces modèles comme non inclusifs car ne prenant pas en compte les personnes de couleur, et présentant des dégradations de performance pour ces dernières. Dans le cas de la reconnaissance faciale, des travaux Buolamwini and Gebru (2018) ont montré que les modèles développés présentaient aussi des performances dégradés pour les personnes de couleur par rapport aux personnes blanches, mais que les performances pour les sujets féminins étaient aussi dégradées par rapport au sujets masculins. Ces modèles sont alors non inclusifs, et cette non inclusivité se repose sur plusieurs critères : la couleur de peau et le sexe du sujet.

On peut comprendre la non inclusivité d'un modèle comme la non prise en compte des sujets qui seront amenés à utiliser le modèle ou être soumis à ce dernier. Au contraire, un modèle pourra être considéré inclusif s'il prend en compte ses utilisateurs et ceux qui fourniront les données qui seront présentées au modèle. La prise en compte des utilisateurs passe par un traitement égalitaire des données de ces derniers et une performance équitable pour chacun d'entre eux.

1.3 Objectifs et méthodologie

Cette thèse se propose d'explorer l'articulation entre apprentissage profond et science soutenable dans un contexte particulier : celui du développement de la vision par ordinateur pour les pays du Sud. Au vu des inégalités environnementales et technologiques entre les différentes régions du monde, cette thèse essaie de répondre à la question : "Comment caractériser les baisses de performances attendues des systèmes de vision par ordinateur dans les pays du Sud ?", en proposant une approche novatrice pour l'étude de biais dans les sytèmes de vision par ordinateur.

Nous proposons dans la suite une méthodologie qui permet de caractériser un biais en identifiant trois caractéristiques pour les biais : la source, le type, et l'impact. La combinaison des trois permet une compréhension des mécanismes à l'oeuvre dans le biais et des possibilités de mitiger ces impacts. Cette méthodologie sera implémentée pour le biais géographique, et appliquée à des données synthétiques et réelles. Elle sera ensuite utilisée dans la construction d'un benchmark pour l'évaluation du caractère occidental d'un modèle, et permettra d'en pointer les forces et les faiblesses.

Le but de ces travaux est de pouvoir renseigner la communauté scientifique désireuse de participer au développement et au déploiement en apprentissage profond dans les pays du Sud sur les complications liées à ces initiatives. Les contraintes liées à la diversité de contextes géographiques commencent seulement à être prises en compte, et une caractérisation de ces dernières pourra améliorer les futures initiatives dans ce cadre, et promouvoir la nécessité de ces dernières.

1.4 Contributions et publications

Nos contributions au sein de cette thèse sont multiples et prennent plusieurs formes. La première est la publication d'une base de données qui propose des images contenant des objets et notions communes et provenant de l'ensemble du globe. Cette base de données est la première à prétendre au fait de couvrir l'ensemble du globe avec des images annotées avec 80 catégories en multiclass. Elle a été publiée dans le dépôt de données DataSuds géré par l'IRD.

COCO-style geographically unbiased image dataset for computer vision applications
T Bayet ; DataSuds, DOI : 10.23708/N2UY4C

La méthodologie qui a permis la collection de cette base de donnée a elle aussi été valorisée lors d'un workshop sur les principes du glissement de distribution, à la conférence ICML en juillet 2022.

Distribution Shift nested in Web Scraping: Adapting MS COCO for Inclusive Data
T Bayet, C Denis, A Bah, J-D Zucker ; ICML Workshop on Principles of Distribution Shift 2022, Jul 2022, Baltimore, United States. ⟨hal-03777066⟩

Nos contributions prennent aussi la forme d'une méthodologie pour le déploiement d'IA dans les pays du Sud, qui joint développement d'un modèle et science participative. A l'opposé de modèles d'IA développés loin du public, nous proposons de rendre les parties concernées par le modèle acteurs de son développement. Cette contribution a été proposée et acceptée à la conférence nationale en intelligence artificielle en Juin 2021.

A Machine Learning approach to improve the monitoring of Sustainable Development Goals : a case study in Senegalese artisanal fisheries
T Bayet, T Brochier, C Cambier, A Bah, C Denis, N Thiam, J-D Zucker ; CNIA 2021 : Conférence Nationale en Intelligence Artificielle, Jun 2021, Bordeaux (virtuel), France. pp.30-37. ⟨hal-03319136⟩

Nous avons aussi contribué à la mise en lumière des défis qui attendent le développement de l'IA dans les pays du Sud. De nombreuses briques technologiques du développement des systèmes automatiques doivent être repensées ou retravaillées pour être adaptées à un développement local et permettre une souveraineté numérique pour les pays concernés. Ces travaux ont été présentés à la journée Résilience et IA de la plate-forme d'intelligence artificielle en juillet 2023.

Les défis du glissement de contexte géographique
T Bayet, C Denis, A Bah, J-D Zucker ; Journée Résilience et IA - Plate-Forme d'Intelligence Artificielle, Jul 2023, Strasbourg, France. ⟨hal-04174037⟩

1.5 Organisation du document

Cette thèse est organisée de la manière suivante :

1.5.1 Contexte et travaux précédents

Dans cette partie, nous présentons les notions nécessaires pour introduire le contexte de nos travaux. Cinq thématiques sont abordées, cette thèse se situant au carrefour de plusieurs domaines. Nous présentons en première section des notions générales en IA, afin de poser les bases techniques et formelles pour la compréhension des enjeux techniques. Nous concluons cette introduction en mettant l'accent sur la vision par ordinateur, qui est le cadre dans lequel cette thèse évolue. La deuxième section présente les biais inhérents au développement de modèles d'IA. Différents cadres d'identification de ces derniers sont présentés, introduisant différentes philosophies et méthodologies concernant la compréhension des biais. Après une présentation de différents outils permettant de mesurer l'impact de ces biais, nous concluons en analysant le biais géographique dans les modèles. La troisième section présente des stratégies de mitigation de biais : la généralisation et l'adaptation de domaines. Ces stratégies utilisent des algorithmes conçus spécialement pour performer dans des situations où des domaines ne sont pas présents dans les données d'entraînement. Après une revue des différentes stratégies, nous examinons comment ces stratégies se sont appliquées au biais géographique. La quatrième section lie les notions de soutenabilité et d'éthique à l'IA. nous y soulignons la dualité de la relation entre soutenabilité et IA, illustrée par deux domaines spécifiques : l'IA pour la soutenabilité et la soutenabilité de l'IA. Nous nous attardons ensuite sur le concept d'IA éthique et pointons les faiblesses de cette notion. Enfin, la dernière section introduit le contexte spécifique du Sud dans le développement de l'IA, à travers la collecte de données, le développement de modèles et le déploiement de ces derniers sur le terrain. Ce contexte est utilisé pour comprendre le caractère occidental de l'IA, avant de conclure en présentant des stratégies d'intégration de problématiques locales dans le développement d'IA : les approches participatives.

1.5.2 Contribution méthodologique

Le chapitre trois se concentre sur la présentation d'un protocole permettant la caractérisation d'un biais. En partant du constat que les précédents travaux manquent de concision dans la précision du biais géographique qu'ils cherchent à combattre, nous proposons un protocole inspiré des différents travaux précédents sur l'identification de biais. Ce protocole se base sur une identification précise d'une source, d'un type et d'un impact pour un biais, précisant le cadre dans lequel ce biais est considéré. Nous nous attardons sur la problématique de la caractérisation d'un glissement à l'aide de ce protocole, et précisions quelles perspectives cet outil peut avoir dans le milieu académique.

1.5.3 Implémentation et validation sur données synthétiques

Dans ce chapitre, nous implémentons le protocole précédemment proposé pour tester son utilité et valider son intérêt. Cette implémentation se repose ne premier lieu sur une revue de littérature autour du biais géographique pour construire un arbre d'identification pour la source du biais géographique. Le type et l'impact sont ensuite sélectionné pour le reste de l'expérience. Nous précisons quelles stratégies sont appliquées pour les choix à réaliser comme la sélection de modèle ou l'entraînement. Partant ensuite du constat qu'il n'existe pas de bases de données synthétiques à biais géographiques, nous entreprenons la construction de tels bases de données en nous inspirant des modifications de la base de données MNIST. Nous construisons ainsi GeoMNIST, une base de donnée synthétique à biais géographique, qui transforme les données de MNIST en fonction de la géographie associée aux données. La répartition des données de GeoMNIST dans différentes géographies permet de générer de manière algorithmique des transformations sur ces données, et donc de générer un biais géographique. Une analyse de plusieurs bases GeoMNIST ainsi générées est proposée, avant de passer à la caractérisation du biais géographique de ces bases de données à l'aide du protocole STI.

Cette expérience montre l'intérêt du protocole STI, tout en démontrant la complexité inhérente au choix d'un impact cohérent pour la mesure d'un objectif.

1.5.4 Expérimentation sur données réelles

Le chapitre cinq va plus loin dans l'implémentation du protocole STI, proposant son implémentation sur des données réelles pour la caractérisation du biais occidental dans les modèles de vision. Une présentation du contexte choisi pour son implémentation introduit les différentes hypothèses et notations liées à cette implémentation, avant d'introduire la méthodologie qui sera employée durant l'expérimentation. Cette méthodologie introduit les bases de données et modèles candidats, ainsi qu'un outil d'évaluation manuel des prédictions. Un modèle et une base de données sont ainsi sélectionnés pour l'expérimentation, et les résultats obtenus sont surprenants : on observe pas de biais occidental dans le modèle sélectionné. Plusieurs déclinaisons de l'expérimentation à différentes granularités viennent soutenir ce résultat. Nous concluons sur la nécessité d'explorer plus en profondeur les prédictions du modèle sélectionné pour être capable d'expliquer les mécanismes à l'oeuvre dans nos résultats.

1.5.5 Exploration visuelle et interprétation de prédictions

Le sixième chapitre commence avec l'exploration visuelle des données de l'expérimentation précédente, en restreignant cette observation visuelle à certaines classes sélectionnées pour leurs résultats particuliers. Ces observations nous mettent sur la piste de biais concomitants, des composantes du biais initialement testé qui influencent les prédiction du modèle. Cette piste nous pousse à formaliser une décomposition d'un contexte en ces composantes, pour illustrer comment les biais concomitants jouent un rôle dans les résultats obtenus. Nous étendons ensuite l'expérimentation précédente à de nouveaux modèles et bases de données, montrant que ce phénomène se répète à grande échelle pour la caractérisation du biais occidental dans les modèles. Une dernière expérimentation pour changer de granularité et observer les performances par pays est proposée, et mène aux mêmes conclusions. Les résultats obtenus soulignent la complexité de la caractérisation du biais géographique à cause des biais concomitants, et mettent en lumière les défis d'une telle caractérisation.

1.5.6 Conclusion et perspectives

Nous concluons en rappelant le cheminement des recherches de la thèse, et en présentant les résultats finaux. Si l'objectif de caractériser le biais géographique dans les modèles de vision n'a pas été atteint, le sujet a été largement traité et les défis inhérents à cette caractérisation sont mis en avant pour de futurs travaux. Cette thèse ouvre de nouvelles perspectives dans l'évaluation préliminaire de la capacité d'un modèle à être déployé dans une zone géographique particulière, et propose des outils dont la communauté scientifique peut se saisir pour réaliser cette caractérisation.

Chapter 2

Contexte et travaux précédents

Chapter Summary

Résumé du chapitre

Cette thèse s'inscrit sur plusieurs plans, mais nous laisserons les lecteurs se référer aux rapports du GIEC et de l'IPBES pour un état des lieux de l'environnement planétaire et les mises en relations

entre l'évolution de ce dernier et les activités humaines. Nous nous concentrons dans cette partie sur l'introduction des notions liées à l' apprentissage profond, et au déploiement de ce dernier au Sud avec une visée soutenable. En section 2.1, ce chapitre introduit les notions et notations générales en intelligence artificielle, en apprentissage machine et en apprentissage profond, et présente les pratiques usuelles en apprentissage profond et l'état de l'art en vision par ordinateur. L'attention est ensuite portée en section 2.2 aux biais rencontrés en apprentissage profond, aux cadres utilisés pour identifier ces derniers, et à l'impact de ces biais sur les systèmes développés. La présentation de ces biais est suivi par celle des solutions algorithmiques à certains de ces biais, l'adaptation et la généralisation de domaines en section 2.3. Ce chapitre ouvre ensuite la réflexion autour de la portée des applications en intelligence artificielle à travers les notions de soutenabilité et d'éthique en section 2.4. Cette section fait état d'une dualité de la liaison entre IA et soutenabilité, et explore les différents aspects de cette dualité. Nous y examinons aussi les relations entre éthique et IA, et proposons une revue critique de l'IA éthique. Enfin, le travail est ancré dans un contexte géographique particulier grâce à une présentation des particularités du développement de l'intelligence artificielle au Sud en section 2.5.

2.1 Notions générales en intelligence artificielle

2.1.1 L'intelligence artificielle et ses dérivés

L'intelligence artificielle (IA) est une notion qui est apparue tôt dans la littérature, et qui anime de nombreux fantasmes. Scientifiquement parlant, il n'y a pas de définition précise de ce qu'est ou n'est pas de l'IA. Tout algorithme implémenté sur un support combinatoire peut être considéré comme une IA, de la montre à un système de trading haute fréquence. La différence tient à la notion commune de ce qui est considéré intelligent - qui varie en fonction des cultures et du temps. D'autres notions se rapportant aux machines intelligentes sont par contre mieux définies. L'avènement du terme Intelligence Artificielle commence par l'IA symbolique dans les années 1930 puis 1950, lors de la conférence de Darthmouth College en 1956 qui est considérée comme fondatrice de la discipline. Le perceptron, le premier neurone artificiel, est créé en 1957. Pourtant, les réseaux de neurones en tant que technique d'IA sont abandonnées au profit d'autres techniques comme les systèmes experts aux environs des années 1969, considérés comme une voie sans issue. Il faudra attendre que le contexte computationnel et structurel du numérique soit plus développé pour que les réseaux de neurones refassent leur apparition en 2012, avec le succès apporté par AlexNet[93]. L'apprentissage machine (ou machine learning en anglais) est défini en 1997 comme le domaine d'étude qui permet aux ordinateurs d'apprendre une tâche sans être programmé explicitement pour cette dernière [111]. L'apprentissage profond est un agencement particulier de modèles d'apprentissage machine, qui permet d'appréhender des tâches plus complexes, généralement mais pas exclusivement basé sur des réseaux de neurones. Nous présentons plus en détail ces notions dans la suite, avant de présenter comment ces dernières sont utilisées en pratique.

2.1.2 L'apprentissage machine et l'apprentissage profond

2.1.2.1 Introduction à l'apprentissage machine

Les modèles informatiques et mathématiques partagent généralement des caractéristiques : une entrée, une sortie, et des mécanismes internes pour lier les premières aux dernières. Mathématiquement, ces mécanismes internes peuvent être traduits en une fonction de l'espace de l'ensemble des données d'entrées du modèle vers l'espace de l'ensemble des données de sortie du modèle. On considérera par la suite tout modèle comme une fonction d'un espace d'un ensemble de données d'entrée vers un espace d'un ensemble de données de sortie. Nous adoptons pour la suite de ce document les notations suivantes :

- X est l'espace des données d'entrées, ou le set des entrées possibles pour un modèle. Il peut prendre de multiples formes, mais ses éléments seront représentés par des vecteurs de $\mathbb{R}^n$.

- Y est l'espace de données de sortie, ou le set des sorties possibles pour un modèle.
- $\mathcal{F} : X \longrightarrow Y$ est un modèle prédictif qui assigne à tout point x de X une sortie y dans Y.
- $\hat{y} = \mathcal{F}(x)$ est la prédiction de la sortie de x réalisée par le modèle $\mathcal{F}$

Le paradigme classique d'un modèle informatique est d'avoir un ensemble de règles encodées dans un algorithme, auquel on présente des données. Suivant les règles encodées dans ses mécanismes internes, le modèle va produire une sortie qui correspondra à l'exécution des règles pour la donnée. On appelle paramètres les règles encodées dans les mécanismes internes du modèle, et on utilise la notation suivante pour les modèles informatiques :

- Θ est le set de toutes les combinaisons possibles de paramètres.
- $\mathcal{F}_\theta : X \longrightarrow Y$ est un modèle prédictif qui assigne à tout point x de X une sortie y dans Y en fonction de ses paramètres définis par $\theta \in \Theta$.

L'apprentissage machine introduit un changement dans ce paradigme : les règles ne sont plus encodées, mais apprises grâce à des données ; l'algorithme traduit alors la manière dont les règles sont apprises par la mise en relation entre les données et les mécanismes internes du modèle. Ce processus d'apprentissage permet d'explorer le set des combinaisons possibles de paramètres Θ et de sélectionner un élément θ qui optimisera les prédictions réalisées par $\mathcal{F}_\theta$. On formalise ainsi le processus d'apprentissage :

- $\mathcal{A} : \Theta \longrightarrow \Theta$ est un processus d'apprentissage des paramètres d'un modèle de machine learning.

Ces processus d'apprentissage se reposent sur des bases de données sur lesquelles les modèles sont entraînés. En apprentissage supervisé, une base de données D est composée d'un sous-ensemble de l'espace d'entrée du modèle X_D et des sorties associées à ces entrées Y_D. Si on considère que le modèle $\mathcal{F}$ est entraîné pour reproduire le comportement de la fonction f associant à tout $x_i \in X_D$ sa sortie associée $y_i \in Y_D$, et que $\mathcal{L}$ est une distance sur Y, alors :

- $\forall x \in X$ et $y = f(x) : R_\mathcal{F}(x) = \mathcal{L}(\mathcal{F}(x) - f(x)) = \mathcal{L}(\hat{y} - y)$ est la performance du modèle sur la donnée x.

On juge un modèle par ses performances, de manière usuelles mesurées sur un jeu de données test T. Ce jeu de données comprend un ensemble de données d'entrées X_T et les sorties associées Y_T. La performance du modèle $\mathcal{F}$ sur T est mesurée par :

$$R_\mathcal{F}(X_T) = \frac{\sum_{x \in X_T}(R_{\mathcal{F}(x)})}{len(X_T)}$$

Le but d'un processus d'apprentissage est de trouver $\theta^* \in \Theta$ tel que

$$\theta^* = \arg\min_{\theta \in \Theta} R_{\mathcal{F}_\theta}(X_T)$$

La recherche de l'ensemble idéal de paramètres θ^* s'opère de manière itérative dans le processus d'apprentissage. A partir d'un état k du processus, on considère $\theta^k \in \Theta$ comme étant la combinaison de paramètres obtenue à l'état k. Dans cet état, le score de perte mesuré sur l'ensemble d'apprentissage est $R_{\mathcal{F}_{\theta^k}}(X_D)$. Ce score de perte permet de déterminer comment faire évoluer la combinaison de paramètres et obtenir θ^{k+1}, dont le score de perte est en théorie meilleur que le précédent : $R_{\mathcal{F}_{\theta^{k+1}}}(X_D) < R_{\mathcal{F}_{\theta^k}}(X_D)$.

Une fois entraînés, les modèles de prédictions permettent de résoudre un grand nombre de tâches, et nous présentons ici les deux problèmes principaux que ces modèles permettent de résoudre : la classification et la régression. Dans les deux problèmes, le modèle $\mathcal{F}_\theta$ doit déduire des données provenant de X la meilleure prédiction dans Y.

- Si Y est discret, le problème est appelé classification.
- Si Y est continu, le problème est appelé régression.

2.1.2.2 Modèles linéaires

Une fonction entre deux espaces vectoriels est dite linéaire quand elle préserve les combinaisons linéaires, c'est à dire : pour deux espaces vectoriels E et F dans $\mathbb{R}$, une application $f : E \longrightarrow F$ est dite linéaire si :

$$\forall (x, y) \in E^2, \forall \lambda \in \mathbb{R}, f(\lambda x + y) = \lambda f(x) + f(y)$$

Si les éléments x de l'espace d'entrée X d'une fonction linéaire sont représentés par des vecteurs de $\mathbb{R}^n$, alors f peut s'écrire sous la forme :

$$\forall x \in X, f(x) = wx + b \text{ avec } w^T \in \mathbb{R}^n \text{ et } b \in \mathbb{R}^n$$

Un modèle est dit linéaire si la fonction qu'il implémente est elle-même linéaire. Les modèles linéaires sont parmi les plus simples en mathématique et en informatique ; ils composent la base de nombreux autres modèles plus complexes. Leur utilisation est encore aujourd'hui extrêmement répandue, pour la simplicité d'implémentation, d'utilisation, et d'interprétation.

Les fonctions linéaires sont néanmoins limitées par leur capacité de représentation ; si toute fonction peut être approchée par une fonction linéaire (et donc tout modèle par un modèle linéaire), cette approximation requiert une dimension qui peut très vite atteindre des valeurs incompatibles avec des calculs numériques. Nous nous intéressons donc dans la suite aux fonctions non linéaires.

2.1.2.3 Modèles non linéaires

En opposition aux fonctions linéaires, les fonctions non linéaires ne préservent pas les combinaisons linéaires. Elles permettent par contre de ne pas être restreintes par des représentations linéaires entre espaces vectoriels d'entrée et de sortie d'une fonction, et possèdent donc de plus grandes capacités de représentations. On peut ainsi approcher n'importe quelle fonction par une fonction non linéaire, la plupart du temps en évitant le problème d'explosion de dimension des fonctions linéaires.

Cette famille de fonction est très vaste et hétéroclite, et ne sera donc pas passée en revue ici ; nous nous contenterons de présenter certaines fonctions non linéaires qui nous sont utiles dans ces travaux, en commençant par le composant principal des modèles d'AP : le neurone.

2.1.2.4 Neurone en apprentissage machine

Le neurone est un modèle d'apprentissage machine inspiré du neurone synaptique, mais dont le fonctionnement est bien différent. Le neurone est un modèle comportant des paramètres appelées "poids", qui seront combinés selon une stratégie particulière aux entrées du modèle. Ensuite, les combinaisons entre poids et entrées seront envoyées en entrée d'une fonction non linéaire dite "d'activation", dont la sortie constituera la sortie du neurone (voir figure 1).

Il existe différents types de neurone, et parmi les plus utilisés en vision par ordinateur, on peut trouver :

- **le neurone dense**, qui réalise une combinaison linéaire des poids et des entrées, en rajoutant un poids supplémentaire appelé "biais du neurone". Cette combinaison linéaire est suivie de la fonction d'activation f. Si l'entrée $x \in X$ est un vecteur de taille k, alors les poids du neurone seront représentés par un vecteur $w \in \Theta$ de taille $k + 1$ tel que la sortie du neurone corresponde à $f(\sum_1^k w_i x_i + w_0)$

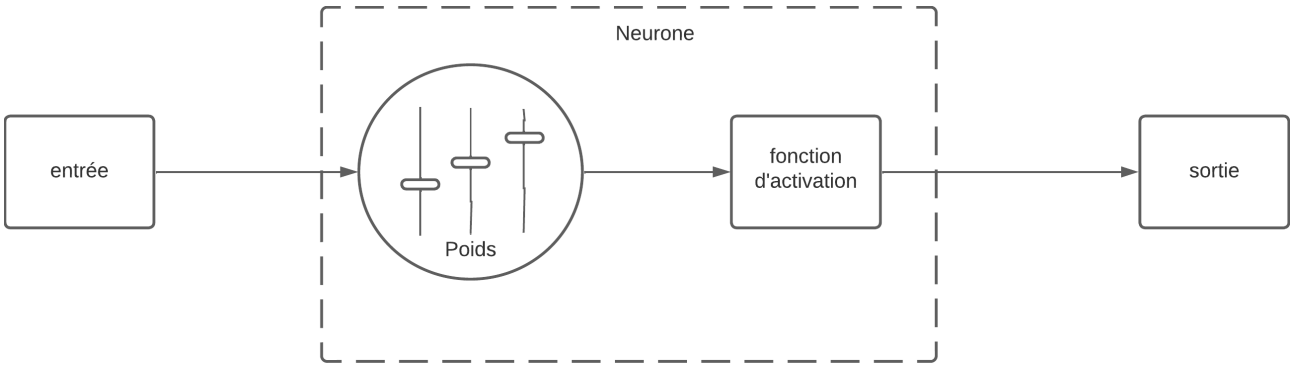


Figure 1: Schéma d'un modèle de neurone en apprentissage machine. Les entrées sont combinées avec les poids selon une stratégie qui dépend du type de neurone, puis une fonction d'activation traite ces informations et les envoie en sortie du neurone.

- **le neurone convolutionnel**, qui réalise une convolution entre un noyau de taille définie par la dimension du neurone et les données d'entrées. La convolution de deux fonctions correspond à une moyenne glissante d'une fonction sur l'autre, ce qui permet d'introduire des notions de proximités entre fonctions. Ces neurones sont surtout utilisés pour traiter des images, et on réalise alors ce qu'on appelle une convolution2D, en faisant glisser le noyau sur une image (matrice de pixels), comme illustré en figure 2. La sortie du produit de convolution de cette matrice sur les parties de l'image sera plus ou moins forte en fonction de la correspondance entre la matrice utilisée et les éléments de l'image, justifiant l'appellation de filtre pour le noyau utilisé.

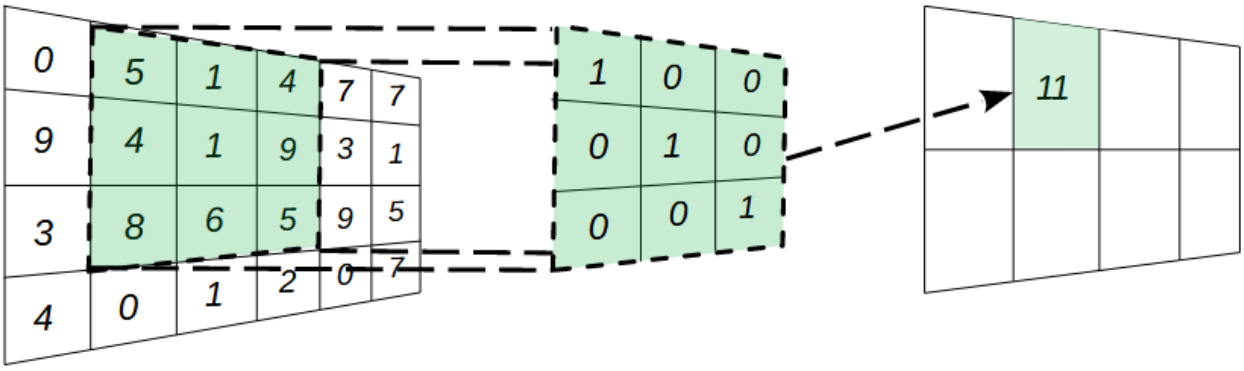


Figure 2: Schéma d'un produit de convolution entre une matrice et un noyau aussi appelé filtre.

Les fonctions d'activation de ces neurones sont usuellement parmi les fonctions suivantes :

- la fonction logistique sigmoïde $\sigma : x \longrightarrow \frac{1}{1+exp(-x)}$.
- la fonction tangente inverse.
- la fonction ReLU (Rectified Linear Unit, ou Unité Linéaire Rectifiée) ReLU : $x \longrightarrow max(x, 0)$.
- la fonction LeakyReLU : LReLU : $x \longrightarrow max(\alpha x, 0)$ avec $\alpha \in]0, 1[$.
- la fonction SoftPlus : SoftPlus : $x \longrightarrow ln(1 + exp(x))$.

Ces fonctions sont toutes non linéaires, condition requise pour que les combinaisons de neurones ne soient pas équivalentes à un seule neurone par le fait que la combinaison de fonctions linéaires est une fonction linéaire. Les fonctions d'activation ne sont pas limitées aux fonctions citées ci-dessus, et en pratique n'importe quelle fonction peut être utilisée. On utilise néanmoins des fonctions combinant certaines propriétés pratiques :

- la continuité, car la discontinuité peut rendre l'entraînement du modèle instable.
- une différentiabilité sur $\mathbb{R}$, pour le calcul du gradient.
- l'identité en 0, qui permet de ne pas se soucier de l'initialisation des poids des neurones.

Cette liste est non exhaustive, d'autres propriétés pouvant être intéressantes pour les fonctions d'activation, mais ces spécificités sortent du cadre de cette thèse.

2.1.2.5 Réseau de neurones

Un réseau de neurones (RN) est un arrangement particulier de ces modèles d'AM, et correspond au modèle de base en apprentissage profond (AP). Un réseau est composé de couches, et le nombre L de couches et appelé profondeur du réseau de neurones. Une couche d'un RN est composé d'un nombre arbitraire de neurones, usuellement partageant tous la même construction (type et fonction d'activation). Dans un RN, l'information se propage couche par couche, chaque couche prenant donc en entrée les sorties de la couche précédente. Les couches d'entrée et de sortie sont des couches spéciales : elles doivent être particulièrement construites pour prendre les données d'entrée et fournir des sorties au format souhaité. Le reste des couches est appelé couches intermédiaires ou couches cachées. La figure 3 représente un réseau de neurones avec une couche cachée de dimension k, prenant une entrée de taille n et produisant une sortie de taille m.

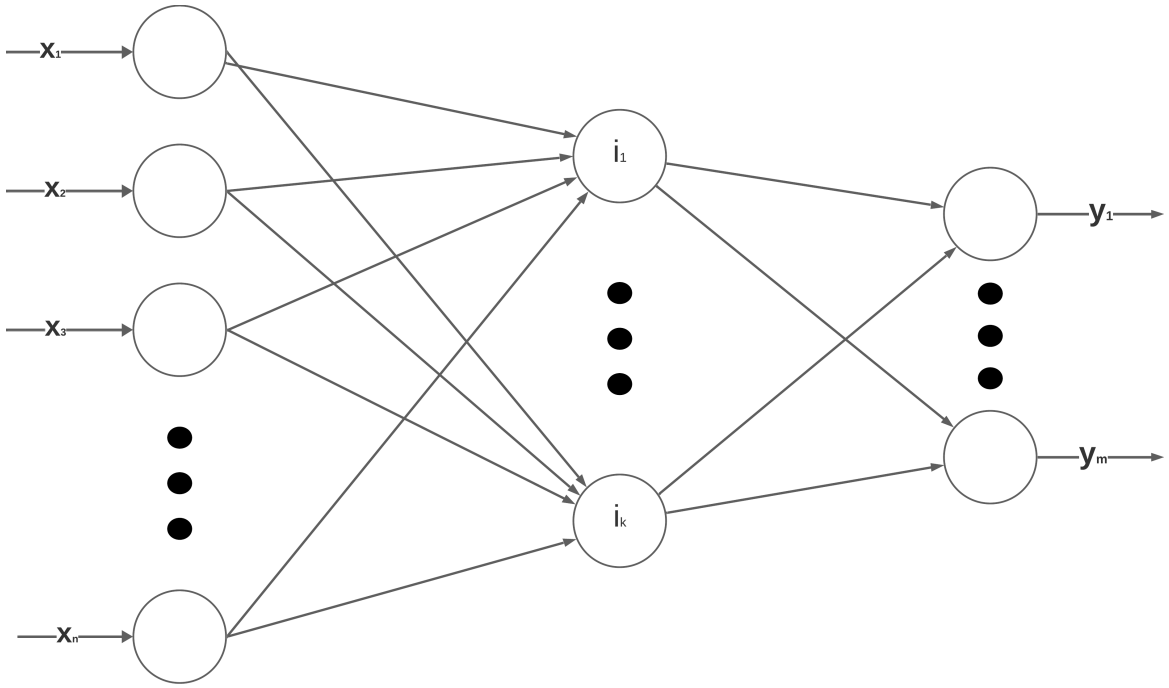


Figure 3: Schéma représentant un réseau de neurones denses avec une seule couche intermédiaire de dimension k, une entrée de dimension n et une sortie de dimension m.

Cet agencement de neurones permet un déplacement de l'information des couches d'entrées vers les couches de sortie tout en passant par l'ensemble des couches intermédiaires. Pour que ces RN soient performants, il faut leur adjoindre une phase durant laquelle les neurones règlent leurs poids pour fournir des sorties appropriées : c'est la phase d'entraînement.

2.1.2.6 Entraînement d'un RN

L'entraînement d'un RN se fait à partir d'une base de données, et peut adopter en principe trois paradigmes différents :

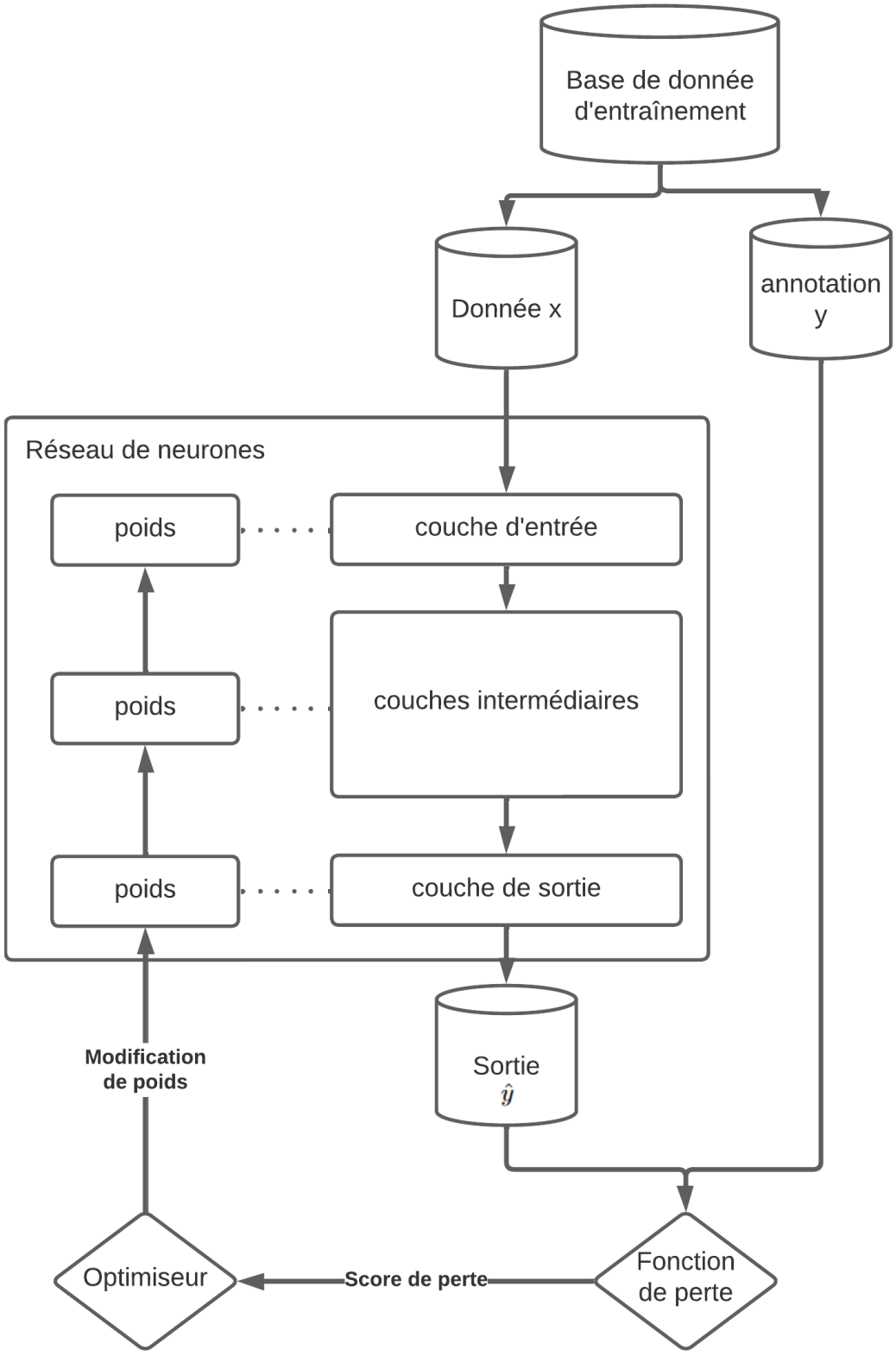


Figure 4: Représentation du paradigme de l'apprentissage supervisé. Les données d'entraînement sont fournies au modèle, et les sorties du modèle sont comparées aux résultats attendus via la fonction de perte. Le score de perte est ensuite transmis à l'optimiseur, qui va modifier les poids du réseau de neurones en fonction de ce score de perte.

- L'apprentissage supervisé : dans ce paradigme, on dispose dans la base de données utilisée pour entraîner le modèle d'annotations qui correspondent pour chaque entrée de la base à la sortie attendue du modèle. De cette manière, on essaie de régler les sorties du modèle d'AP sur les annotations associées aux données. Ce paradigme est illustré en figure 4.
- L'apprentissage semi-supervisé : dans ce paradigme, on dispose d'un sous-ensemble de données annotées dans la base d'entraînement. Le reste des annotations sera fourni par le modèle

lui-même en étant utilisé sur la base de données après un premier apprentissage, ou par d'autres procédés similaires.

- L'apprentissage non supervisé : dans ce paradigme, il n'y a aucune annotation disponible dans la base de données utilisée pour entraîner le modèle. Des mécanismes internes au processus d'entraînement permettent de guider le modèle vers des sorties appropriées.

Quelque soit le paradigme adopté pour l'entraînement, on commence avec un arrangement de poids aléatoires pour les poids des neurones dans les couches du RN. L'entraînement d'un RN consiste à sélectionner des entrées et comparer les sorties du modèle à la sortie attendue grâce à une métrique appropriée, et à modifier itérativement les poids des neurones en fonction des scores de cette métrique. Deux modules remplissent ces rôles : les fonctions de perte et les optimiseurs.

Les **fonctions de perte**, notées $\mathcal{L}$, sont les métriques qui sont utilisées pour évaluer la performance du modèle. Ces fonctions permettent durant la phase d'entraînement de calculer à quel point le modèle se trompe, dont le résultat est appelé "score de perte". Un processus d'apprentissage pour un RN est donc un processus qui vise à réduire le score de perte d'un modèle sur un ensemble de données d'apprentissage. Les fonctions de pertes classiques utilisées dans les RN sont réparties en deux catégories : les fonctions probabilistes et les fonctions pour la régression. Les plus classiques sont les suivantes :

- la cross-entropie binaire, utilisé pour la classification binaire
- la cross-entropie catégorique, utilisée pour la classification multi-classe sous format catégorique
- l'erreur quadratique moyenne, utilisée pour la régression
- l'erreur absolue moyenne, utilisée aussi pour la régression.

Le module permettant de déterminer comment modifier un ensemble de paramètres pour améliorer les performances du modèle est appelé **optimiseur**. L' optimiseur permet de modifier les poids des neurones dans un RN. Il prend en entrée les scores de perte fournis par la fonction de perte et modifie les poids en conséquence : un grand score de perte indique une grande différence entre sortie du modèle et résultat attendu, et produira donc des grandes modifications dans les poids des réseaux de neurone. Un petit score de perte au contraire sera témoin d'une certaine proximité entre sortie du modèle et résultat attendu, et produira donc de petites modifications dans les poids des réseaux de neurones. Il existe différents algorithmes pour effectuer cette modification de poids, mais le plus général est l'algorithme dit de "back-propagation".

Cet algorithme permet la remontée de l'information dans les différentes couches du RN, en commençant par les couches les plus basses (ie. les plus proches de la couche de sortie) et en remontant vers les couches lesp lus hautes (ie. les plus proches de la couche d'entrée). L'opération réalisée par cet algorithme est la modification des paramètres des couches dans la direction des gradients négatifs pour ces paramètres. Cette opération est illustrée par la formule suivante, pour un état k :

$$\theta^{k+1} = \theta^k - \rho \frac{\partial \mathcal{F}_{\theta^k}}{\partial \theta^k}$$

avec ρ un hyper-paramètre appelé taux d"apprentissage. Les optimisateurs usuels sont les suivants :

- stochastic gradient descend (SGD) est une recherche stochastique dans l'espace des gradients de l'espace des paramètres. C'est le premier algorithme utilisé pour modifier les paramètres d'un réseau de neurones via la propagation couche par couche.
- root min square propagation (RMSPROP) est un algorithme basé sur l'utilisation du signe des gradients et d'une stabilisation par mini-batchs.

- Adam est un algorithme plus récent, basé sur le couplage de RMSPROP à des moments. Il en existe un nombre important de versions (Adadelta, Adagrad, Adamax, Nadam, ...).

Le taux d'apprentissage est un paramètre de l'optimiseur et revêt une importance particulière. Plus ce paramètre est élevé, plus les modifications des paramètres sont importantes, mais plus il est difficile de trouver un optimum local. Plus ce paramètre est bas, plus les modifications des paramètres sont petites, et le nombre d'itération nécessaires pour atteindre un minimum local peut alors devenir très grand. Un choix judicieux de taux d'apprentissage permet un apprentissage rapide et précis. Pour permettre d'optimiser le processus d'apprentissage, il est généralement variable et décroissant en fonction du nombre d'epoch, et il existe plusieurs stratégies pour modifier ce dernier :

- la décroissance par étapes : ces stratégies définissent manuellement à partir de quelle epoch le taux d'apprentissage doit être modifiée, et la modification associée.
- la décroissance linéaire : ces stratégies définissent une relation linéaire décroissante entre epochs et taux d'apprentissage.
- la décroissance polynomiale : ces stratégies définissent une relation polynomiale entre epochs et taux d'apprentissage
- la décroissance cyclique : ces stratégies établissent une relation sinusoïdale entre epochs et taux d'apprentissage. Un facteur décroissant linéaire en fonction des epochs peut être ajouté pour assurer un taux d'apprentissage décroissant dans le temps.

En pratique, pour des raisons de mémoire, les scores de perte ne sont pas calculés d'un coup sur les datasets en entiers, mais sur des regroupement d'inputs appelés batchs. Ces batchs sont fournis en entrée du modèle, qui calcule pour chacun de ces batchs les sorties associées aux inputs qui les composent. Un score de perte est associé à ce batch, et cumulé aux scores des autres batchs pour obtenir le score de perte de la base de données. Ce fonctionnement permet de traiter des très groses bases de données sans nécessiter des capacités mémoires extraordinaires.

Une autre pratique régulière durant l'entraînement d'un RN est de séparer la base de données utilisée en trois éléments : un jeu d'entraînement, un jeu de test, et un jeu de validation. Ces ensembles remplissent tous un rôle différent dans l'entraînement :

- **Le jeu d'entraînement** est le jeu de données qui est présenté au modèle lors de la phase d'apprentissage. En pratique, il est présenté entièrement au modèle batch par batch, lors de ce qu'on appelle une epoch. A la fin d'une epoch, tout le jeu d'entraînement a été envoyé au modèle, et le score de perte du jeu de données permet de suivre l'amélioration des performance du modèle. Avant de démarrer une nouvelle epoch, on procède à une évaluation sur le jeu de validation.
- **Le jeu de validation** est un jeu de données qui permet d'évaluer le modèle à la fin de chaque epoch, durant la phase d'apprentissage, mais sur lequel le modèle ne s'entraîne pas. Il est utilisé pour contrôler que le modèle ne se spécialise pas sur le jeu d'entraînement. Si les performances du modèle s'améliorent sur le jeu d'entraînement mais pas sur le jeu de validation au fur et à mesure des epochs, alors le modèle se spécialise sur le jeu d'entraînement, mais pas sur la tâche qu'il est censé réaliser. C'est donc un jeu de données qui permet de contrôler le bon déroulement de la phase d'apprentissage. En pratique, on conserve à la fin de l'entraînement le modèle issu de l'itération qui a obtenu les meilleures performances sur le jeu de validation.
- **Le jeu de test** est un jeu de données qui n'est pas utilisé lors de l'entraînement, mais uniquement une fois celui-ci fini, pour évaluer les performances du modèle. Son rôle diffère du jeu de validation dans le fait que le jeu de validation est "vu" par le modèle et le modélisateur durant l'entraînement, et que ceci peut mener à des stratégies pour obtenir des bons résultats sur le jeu de validation qui ne seraient pas transférables sur d'autre jeux de données. C'est donc un jeu de données de contrôle qui permet d'évaluer la performance d'un modèle après entraînement.

Ces trois ensembles doivent être distincts et ne pas partager de données ; autrement, leur rôle serait défait par nature. Toutes ces considérations permettent de procéder à l'entraînement d'un RN. Le domaine de l'AP a pourtant encore des particularités, qui sont décrites dans la sous-section suivante.

2.1.3 Pratiques en apprentissage profond

L'AP est fort d'une communauté de milliers de scientifiques dans le monde. De cette communauté émane des pratiques partagées par les milieux académiques et industriels. Nous passons dans cette sous-section ces pratiques en revue.

2.1.3.1 Chaîne de traitement de l'apprentissage profond

La chaîne de traitement de développement de systèmes d'AP est partagée par de nombreux pratiquants du domaine. Même si de nombreuses variations existent, les pratiques sont globalement similaires d'une entité à une autre. Les différentes étapes de développement d'un système d'AP sont les suivantes :

- **La collection de données** : les systèmes d'AP se reposent sur une collecte massive de données pour entraîner, tester et déployer leurs modèles. Cette étape peut aussi comprendre l'annotation des données, qui est nécessaire dans le cas de l'apprentissage supervisé.
- **La construction d'une base de données** à partir des données collectées passe par le prétraitement de ces données et l'organisation dans les différents jeux de données d'entraînement, de validation et de test. Ces deux premières étapes sont les plus importantes, puisque tout le système repose sur les données qui seront utilisées. L'utilisation de bases de données de plus en plus grandes a permis le développement de l'AP à grande échelle, et nous précisons l'importance de ces bases de données en section 2.1.3.2.
- **La construction du modèle** comprend le choix de l'architecture et de nombreux choix sur les paramètres du modèle comme le nombre de couches, l'organisation de ces dernières, le nombre de neurones par couches, etc... Une grande importance est usuellement attribuée à la conception du modèle.
- **L'entraînement du modèle** est l'étape qui est la plus computationnellement lourde, et qui nécessite des capacités de calcul importantes. L'entraînement est déterminé par la fonction de perte et l'optimisateur choisie, ainsi que par le nombre d'epoch et la taille du batch. De nombreux autres paramètres entrent en jeu, et nous précisons ceux-ci plus en détails en section 2.1.3.4.
- **L'évaluation du modèle** permet de valider les capacités du modèle du système après entraînement. La communauté évalue communément les modèles sur des benchmarks publics, qui permettent de comparer les performances d'un modèle à celles des autres modèles proposés par la communauté.
- **Le déploiement du modèle** est une étape entreprise une fois le modèle validé. Le déploiement nécessite la mise en place d'une chaîne de traitement de données : récupération des données pour l'application visée, envoi des données au système, récupération et traitement des sorties du modèle.

Ces étapes sont illustrées dans le schéma en figure 5.

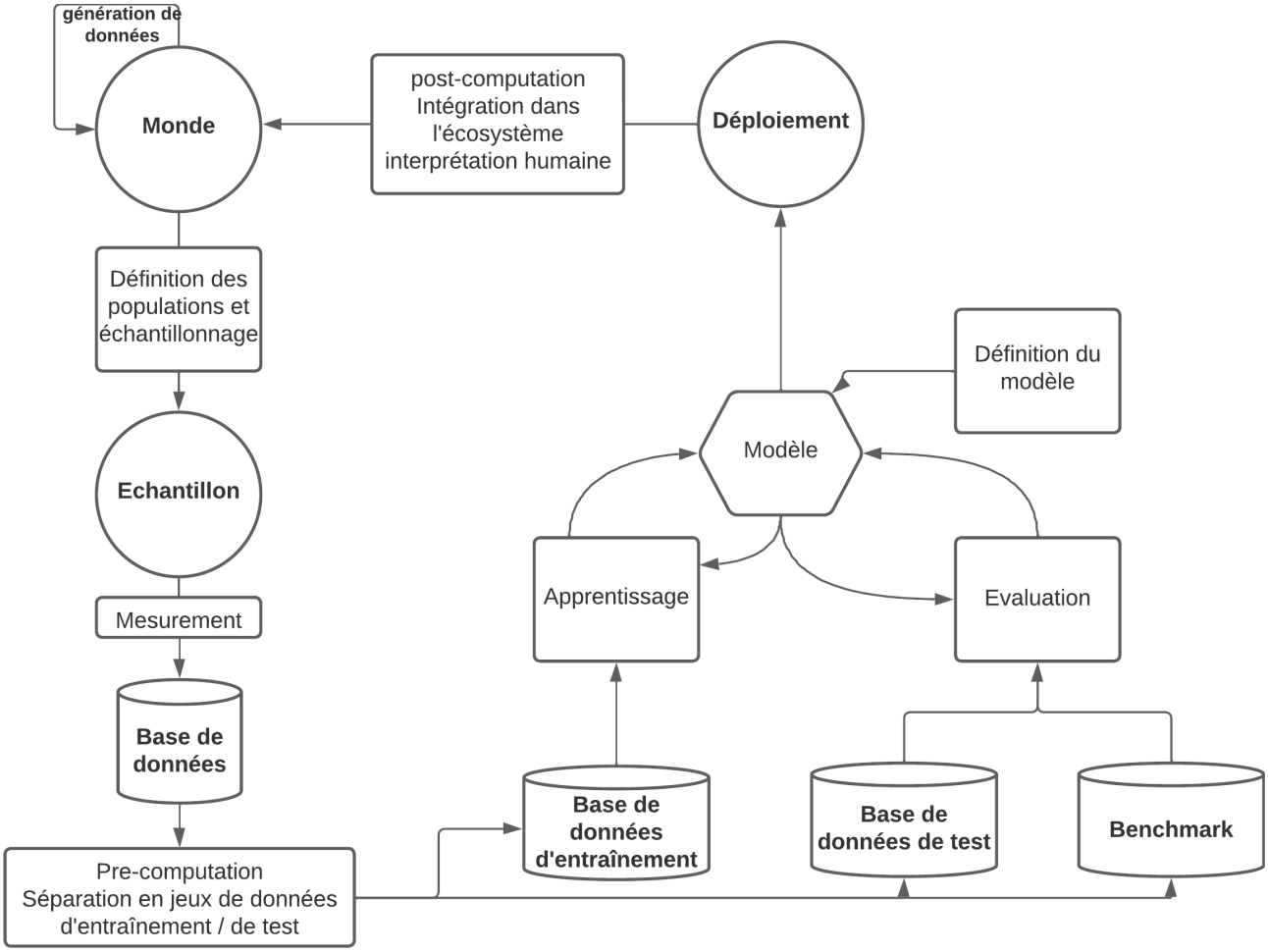


Figure 5: Chaîne de traitement pour le développement d'un système d'AP.

2.1.3.2 L'importance des bases de données

Les bases de données ont permis un développement rapide et massif des systèmes d'AP. Leur utilisation dans des benchmarks et les concours scientifiques organisés sur des tâches particulières (Imagenet Large Scale Visual Recognition Challenge [138] (ILSVRC), PASCAL Visual Object Challenge [57] (PASCAL VOC)) ont popularisé et normalisé les pratiques en AP. La croissance de ces bases de données en volume a aussi permis d'atteindre des performances de plus en plus impressionnantes, les performances des modèles allant croissant avec le nombre de données d'entraînement [77]. Les premières bases de données visuelles (PASCAL [57], Imagenet [47], MNIST [101]) ont permis le développement de nouvelles bases de données plus grandes et couvrant un plus grand nombre de tâches (Open Images [95], WILDS [92]). Ces bases de données publiques sont le pivot du développement et de la popularisation de l'IA. Leur accessibilité a rendu possible l'investissement d'un nombre grandissant d'acteurs, tandis que leur caractère massif assure une diversité d'applications et de dérivations. Le partage de ces bases de données et des benchmarks associés a aussi assuré le partage des modèles à l'état de l'art. Cette philosophie du partage s'est montrée garante d'un succès sans équivoque pour l'AP qui est très vite devenu la technologie de référence pour le traitement de données.

2.1.3.3 L'architecture d'un RN

La popularisation des benchmarks a soutenu le partage de modèles, dans l'optique de proposer une recherche et des résultats réplicables, assurant la crédibilité des travaux de recherche. Ce partage a donné naissance à des normes et des pratiques qui sont devenues courantes dans la construction de modèle, et on a vu émerger dans la communauté une notion particulière pour désigner l'ossature d'un modèle : l'architecture. L'architecture d'un RN correspond à un ensemble de choix relatifs au nombre de couches, à leur agencement, aux neurones qui les composent, au nombre de neurones,

aux fonctions d'activation d'un RN. Ces paramètres composent le design même du modèle, et sont appelés hyper-paramètres, en opposition aux paramètres du modèle, qui font référence aux poids qui sont modifiés durant la phase d'apprentissage. Certains architectures sont devenues populaires au point d'avoir un nom et de devenir normes - comme Resnet, U-net, Darknet, etc... L'architecture la plus classique utilisée aujourd'hui est Resnet, et comporte plusieurs formes : Resnet-18, Resnet-34, Resnet50, Resnet-101, Resneg-152 etc... Ces chiffres correspondent au nombre de couches dans l'architecture, et donc à une variation d'un hyper-paramètre au sein de l'architecture de base. La figure 6 introduit une représentation de Resnet-50.

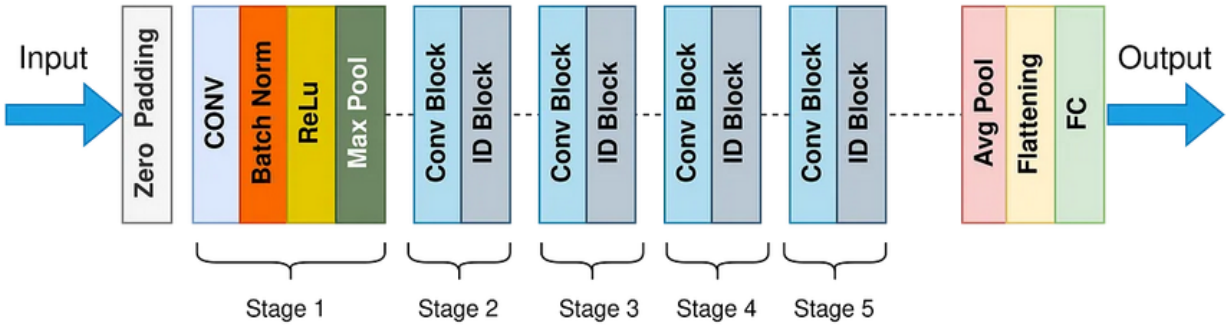


Figure 6: Présentation de l'architecture Resnet-50 issue de He et al. (2016).

On y observe différentes types de couches, et différentes étapes. La première couche est une couche de zero padding, une opération qui consiste à agrandir la taille de l'image en ajoutant des 0 aux extrémités de celles-ci. C'est une opération qui sert à éviter les effets de perte d'information aux bords des images, et qui ne contient pas de paramètres entraînables. La seconde couche est une couche de convolution, composée de neurones convolutionnels et réalisant donc des opérations de convolution entre cartes de caractéristiques en entrée et un kernel qui représente les paramètres des neurones. La troisième couche réalise une opération de normalisation des batchs, qui est une opération mathématique qui permet d'éviter certains effets négatifs du réseau de neurones en régularisant les batchs avant de passer à la couche suivante. La quatrième couche est une couche d'activation, qui correspond à l'ajout de la fonction d'activation ReLU. La cinquième couche est une couche de Max Pooling : elle réalise une compression des données en ne conservant que les valeures maximum sur une fenêtre glissante. De toutes ces opérations, seule la couche de convolution contient des paramètres entraînables, les autres étant des opérations mathématiques sans paramètre variable. Les couches suivantes (Conv Block, ID Block) correspondent à ce que l'on appelle des blocs de convolution : des enchainements de couches de convolution et de normalisation de batch, combinés à une couche d'activation (ReLU ici) en sortie. Enfin, dans le dernier groupe de couches à droite, on retrouve une couche de Average Pooling, qui est similaire au Max Pooling mais qui sauvegarde la moyenne des valeures au lieu de la valeure maximal, une couche de mise à plat (Flattening) des données, c'est à dire une projection d'une carte de dimension nxm vers une carte de dimension 1x$n * m$, puis une couche de classification qui permet de traduire les computations du modèle en output valides pour le modèle. Ce qui est appelé ResNet est généralement la partie composé des étapes 1 à 5 sur la figure 6, le bloc en sortie étant appelé "bloc de classification" et étant spécifique à la tâche que doit réaliser le réseau de neurones. Ainsi, on peut, lorsqu'on s'attaque à une nouvelle tâche, reprendre exactement l'architecture du ResNet-50 et ne changer que le bloc de classification pour entraîner un modèle qui sera opérationnel.

Entre ResNet-50 et les autres versions de l'architecture ResNet, seuls les nombres de répétitions des blocs de convolution des étapes 2 à 5 varient, comme le montre la figure 7.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
	1×1	average pool, 1000-d fc, softmax				

Figure 7: Présentation des différentes configuration pour les architectures ResNet, issu de He et al. (2016).

Ces variations d'une même base architecturale ont des intérêts sur différents niveaux :

- La capacité : on définit la capacité d'un modèle comme sa potentielle performance. En pratique, plus un modèle possède de couches, plus l'espace de recherche des paramètres grandit et plus le réseau de neurones est capable d'atteindre des niveaux d'abstraction complexes. Ainsi, un réseau profond (comprenant de nombreuses couches) sera plus à même d'approcher la fonction-objectif de la tâche qu'il doit apprendre, et pourra potentiellement atteindre une performance plus grande qu'un réseau avec moins de couches. On a donc intérêt à utiliser le réseau le plus profond possible si l'on veut atteindre les meilleures performances possibles.
- L'emprunte mémoire : plus un modèle possède de couches, plus il possède de paramètres ; et ces paramètres sont enregistrés avec le réseau en mémoire. Quand ces réseaux atteignent des centaines de milliers voir des millions de paramètres, la question du stockage de l'état du modèle est importante et peut être un élément limitant dans les choix d'une architecture particulière. On a donc intérêt à choisir le réseau le moins profond possible si la question de l'espace mémoire est une contrainte forte.
- La nécessité computationnelle : de manière générale, plus un réseau de neurones comporte de paramètres, plus il est lourd computationnellement parlant (c'est vrai à type de couches égales ; on peut avoir un grand nombre de paramètres dans un réseau dense qui sera moins lourd à entraîner qu'un réseau convolutionnel comportant moins de paramètres). Il faudra donc de meilleures capacités computationnelles (capacité de calcul, de stockage) avec les contraintes qui y sont liées (énergétiques, économiques, logistiques). Ces contraintes peuvent pousser vers un choix d'un architecture plus légère.
- L'emprunte temporelle : De manière générale également, plus un réseau de neurones comporte de paramètres, plus il est temporellement coûteux de l'entraîner. Comme pour la computation, cela dépend du type de couches ; mais à type de couches égales et toutes choses égales par ailleurs, augmenter le nombre de couches augmente le temps nécessaire à l'entraînement d'un réseau de neurones. Il est courant de voir de larges réseaux de neurones nécessiter des semaines ou des mois d'entraînement aujourd'hui. De fait, le temps peut être une contrainte particulière qui peut pousser à sélectionner des architectures plus légères.

Le choix d'une architecture se fait donc dans un contexte technique particulier, avec une connaissance des objectifs de performances visés, des capacités computationelles et temporelles envisagées, et des capacités des terminaux de déploiement du modèle. De même, l'entraînement est un mix de pratiques qui regroupent un certain nombre de choix qui sont explorés plus en avant dans la sous-section suivante.

2.1.3.4 L'entraînement

L'entraînement est une étape cruciale du développement d'un système en AP. Nous avons déjà vu précédemment le découpage du processus d'apprentissage, et comment les poids d'un réseau de neurones étaient modifiés pour améliorer le modèle dans le sens d'une diminution du score de perte, indicateur de performance du modèle. Nous nous concentrons ici sur des pratiques spécifiques lors de l'entraînement d'un modèle, sans pour autant être exhaustif. Ces pratiques sont devenues des normes à travers la communauté pour leur practicité ou les caractéristiques qu'elles apportent au système. On compte parmi ces dernières :

- Le suivi des variables d'entraînement : afin de pouvoir contrôler le bon déroulement de la phase d'entraînement, qui peut se dérouler pendant des temporalités très longues, de nombreux moyens sont mis à disposition pour vérifier que le modèle apprend bien. Les métriques suivies sont généralement celles du score de perte et de la métrique de performance du modèle.

- L'utilisation de callbacks : les callbacks sont des fonctions qui sont appliquées durant l'entraînement afin de modifier certains paramètres ou de récupérer des valeurs intéressantes pour le suivi de l'entraînement. Le suivi des variables d'entraînement se repose d'ailleurs sur des callbacks, mais d'autres peuvent être utilisés, comme la sauvegarde de modèles pour éviter de perdre les données en cas de crash, la modification du taux d'apprentissage de l'optimiseur, l'arrêt anticipé de l'entraînement dans le cas où le modèle ne s'améliore plus sur les données de validation. Ces callbacks permettent de rendre l'entraînement plus modulable pour les opérateurs.

- La recherche d'hyper-paramètres : les hyper-paramètres utilisés (taille du réseau de neurone, nombre de neurones par couches, valeur du taux d'apprentissage, taille du batch) peuvent grandement influencer les capacités d'un modèle à atteindre des hautes performances et sa vitesse de convergence vers ces hautes performances. En conséquence, une pratique courante est de tester plusieurs arrangements de paramètres dans une recherche dite "en grille" (testant chaque combinaison de paramètre) durant quelques epochs et de sélectionner la combinaison d'hyper-paramètres présentant les performances les plus prometteuses.

2.1.3.5 Réutiliser un modèle déjà entraîné

L'entraînement de modèles volumineux à l'aide de bases de données massives et durant des temps importants est couteux, mais a aussi un énorme avantage : ces modèles peuvent être réutilisés par le reste de la communauté quand ils sont partagés. Ainsi, un modèle entraîné sur une tâche peut être récupéré pour être réutilisé sur la même tâche.

Mais la particularité des réseaux de neurones et des architectures est tout autre. Dans l'architecture ResNet, les différentes couches de convolution sont utilisées pour créer des cartes de caractéristiques et apprendre quels éléments sont utiles pour la tâche à réaliser. Il s'avère que pour un domaine en particulier, ces caractéristiques sont assez générales et peuvent être réutilisées pour d'autres tâches. Nous nous plaçons dans la suite dans le domaine de la vision par ordinateur, pour illustrer le propos. Toutes les tâches de vision partageant des intérêts similaires (détermination d'éléments permettant de différencier un objet d'un autre, une classe d'une autre, nécessité de comprendre les formes, d'identifier des contours, etc...), les caractéristiques apprises par un réseau de neurones entraîné sur une large base de données d'images pendant un temps long peuvent être considérées comme générale pour les tâches de vision par ordinateur. Si l'on retire alors le bloc de classification suivant l'architecture ResNet, on obtient un extracteur de caractéristiques utiles pour la vision en AP. Pour entraîner un modèle capable de réaliser une nouvelle tâche, il suffit alors de rajouter un bloc de classification ou de régression réalisant la tâche désirée en sortie des blocs de convolution déjà entraînés du ResNet. Pour faciliter l'entraînement (notamment pour se débarasser des contraintes computationnelles et temporelles) et ne pas perdre une information durement acquise, on "gèlera" les couches du ResNet : on empêchera les poids du ResNet d'être modifiés lors de la nouvelle phase d'entraînement. Ainsi, on peut entraîner un

modèle sur n'importe quelle tâche en vision par ordinateur en remplaçant le bloc de classification initial par un bloc de sortie vierge, et en entraînant uniquement les poids de ce dernier bloc. On appelle ces architectures classiques et leurs poids déjà entraînés sur des bases de données particulières des modèles "pré-entraînés", et la méthode consistant à utiliser ces modèles pré-entraînés pour construire de nouveaux modèles réalisant des nouvelles tâches l'apprentissage "par transfert" [127].

L'apprentissage par transfert a permis la diffusion rapide et facile de très nombreux modèles et une déclinaison de ces modèles et systèmes dans de nombreuses applications. Le partage des nouvelles architectures et des paramètres et poids associés a fondé l'essor de l'AP. D'autres méthodes pour réutiliser les modèles partagés ont vu le jour, comme le "fine-tuning", qui permet de peaufiner un modèle après apprentissage par transfert, en dégelant quelques couches de convolution en plus du bloc de sortie. Cela permet d'adapter les caractéristiques à la sortie du réseau pré-entrainé à la tâche réalisée par le modèle, et donc de gagner encore en performance, bien que de l'information générale puisse être perdue.

2.1.3.6 Apprendre avec peu

Si l'AP est fondé sur la disposition d'un volume important de données, toutes les applications n'appartiennent pas à ce contexte. De nombreuses applications nécessitent un déploiement dans des contextes contenant peu ou pas de données. Ce sont dans ces contextes que ce sont développés des nouvelles méthodes d'AP, appelées apprentissage "few shot", "one shot" ou apprentissage "zero shot" [102].

L'apprentissage "few shot" est un paradigme où le processus d'entraînement est réalisé avec très peu de données comparé aux volumes classiques utilisées en AP. Pour ce faire, on sélectionne généralement un modèle pré-entraîné sur une tache, et on tente de l'adapter à de nouvelles tâches avec un volume peu important de données. Ces approches se découpent généralement en deux étapes : la modification du réseau entraîné pour apprendre plus facilement des caractéristiques utiles pour la nouvelle tâche, et la recherche d'hyper-paramètres qui permettront d'apprendre plus efficacement ces nouvelles tâches [39; 175].

L'apprentissage "zero shot" est un paradigme singulier, qui se repose sur l'apprentissage de tendances générales dans un domaine. Un modèle ayant appris des tendances générales pourra s'adapter sans données requises à de multiples tâches, d'où le terme zero shot, qui se réfère au fait qu'il n'y a pas de ré-entraînement du modèle. Ces modèles sont généralement multi-modaux, permettant le lien entre différents médias ou manière de représenter l'information (comme le modèle CLIP [135], qui lie des images et leur description pour permettre l'annotation par prompt).

2.1.4 Etat de l'art en vision par ordinateur

Nous nous concentrons maintenant sur le domaine de la vision par ordinateur, et allons faire l'historique de l'état de l'art de ce domaine, en nous concentrant sur les méthodologies pour la collecte de données, les bases de données, les architectures, les modèles fondateurs, et les systèmes clés en main de vision par ordinateur.

2.1.4.1 Les méthodologies de la collecte de données

En vision par ordinateur, les données composant la plupart des bases de données sont des images, collectées grâce à différentes méthodologies. Certaines proposent de scanner certains sites de dépôts d'images, soit sans tri, soit par recherche en utilisant des mots clés [47; 95; 107; 18]. D'autres initiatives collectent leurs données par requêtes auprès de certaines communautés ou initiatives impliquant des volontaires à travers le monde [65; 147]. D'autres encore commandent leurs données, et les collectes sont alors réalisées par des personnes payées pour cette tâche [11]. Ces méthodes diffèrent par leurs capacités à collecter des images de qualité, le scan sur les plateformes récupérant tout type d'artefacts, tandis que les données collectées auprès de personnes rémunérées sont généralement plus précises. Mais la donnée image est en AP généralement couplée à son annotation, dont la nature dépend de la tâche à réaliser sur la base de donnée. Ces annotations peuvent être collectées en même

Méthodologie	Avantage	Inconvénient
Scanner les sites webs	Récupération rapide de données Coût limité Gros volume disponibles	Mauvaise qualité des données Respect des droits d'auteur
Collecte par la communauté	Respect potentiel des droits d'auteur Coût limité	Qualité des données variable Volume de donnée mitigé
Collecte par des professionnels	Respect des droits d'auteur Qualité des données	Coût important Volume de donnée mitigé

Table 1: Comparaison des différentes méthodes de collecte de données pour constituer des bases de données d'image.

temps que la donnée, ou ajoutées automatiquement ou manuellement à ces dernières. Par exemple, les données de MS COCO [106] ont été collectées via des requêtes sur la plateforme Flickr puis annotées manuellement par des annotateurs spécialistes. Les données d'Imagenet ont été collectées via des requêtes à différents moteurs de recherche d'images, et ce sont ces requêtes qui ont servi d'annotation aux images après une phase de nettoyage des données. Open Image se repose sur une collecte arbitraire de données sur la plateforme Flickr suivie d'une annotation manuelle. Les données de DSD sont annotées d'une classe qui correspond à l'objet ou au concept photographié par le spécialiste lors de la collecte de donnée. Shankar et al. (2017) ont recours à des communautés aux origines variées pour collecter des données images sur des classes spécifiques. Les avantages et inconvénients de chacune de ces méthodes sont listés dans le tableau 1.

La création d'une base de donnée requiert donc une phase de conception, qui organisera la collecte des données et des annotations. Cette phase de collecte sera suivie d'une phase de contrôle qualité, où l'on vérifie que les données et les annotations sont bien valides pour les données collectées. Il faut ensuite organiser cette base de données, avant de pouvoir la proposer au reste de la communauté scientifique.

2.1.4.2 L'évolution des bases de données

Les premières bases de données publiques en vision par ordinateur étaient des bases de données très simples, comprenant des concepts simples. On listera parmi ces dernière Cats vs Dogs[1] en 2013 qui contient 25 000 images de chats et de chien, la base de donnée MNIST[101] composée de 60 000 images de chiffres manuscrit de 0 à 9 numérisés, Pascal VOC[57], un concours scientifique organisé entre 2005 (1 578 images, 4 classes) et 2012 (11 530 images, 20 classes)[2], et Imagenet [138], une base de donnée qui a aussi été utilisée comme benchmark pour des concours scientifiques sur diverses tâches de 2010 (1.2 millions d'images, 1000 classes)[3] à 2017(même données, tâches différentes)[4]. En 2017, OpenImages[95] prend le relai, avec une première version contenant 9 millions d'images et 6000 catégories[5]. Aujourd'hui, OpenImages est toujours composé de 9 millions d'images, mais les annotations couvrent maintenant plus de tâches, comme la détection d'objets (16 millions de boîtes englobantes), la relation visuelle (3.3 millions d'annotations), la segmentation d'objets et de concepts (2.8 millions de masques), la détection multimodale (675 000 annotations), localisation au niveau du pixel (66 millions d'annotations), et la classification(61.4 millions d'annotations)[6]. De nombreuses autres bases de données sont développées pour des tâches spécifiques, mais lister ces dernières de manière exhaustive relève d'un autre travail. Parmi les évolutions dans les bases de données publiques en vision par ordinateur, on observe :

[1] https://www.kaggle.com/c/dogs-vs-cats
[2] http://host.robots.ox.ac.uk/pascal/VOC/
[3] https://image-net.org/challenges/LSVRC/2010/index.php
[4] https://image-net.org/challenges/LSVRC/2017/index.php
[5] https://github.com/openimages/dataset/blob/main/READMEV1.md
[6] https://storage.googleapis.com/openimages/web/factsfigures_v7.html

Type d'annotation	Avantage	Inconvénient
Automatique	Rapidité Coût faible	Mauvaise qualité des annotations
Semi-automatique	Durée réduite Coût réduit	Qualité des données variable
Manuelle	Meilleure qualité des données	Coût important Temps d'annotation important

Table 2: Comparaison des différentes méthodes d'annotation de données pour constituer des bases de données d'image.

- Des volumes de données de plus en plus important. On a vu précédemment comment ces volumes de données sont en lien avec les performance des modèles. Ces bases de données de plus en plus volumineuses sont en adéquation avec la recherche de performance de plus en plus extraordinaires.
- Des annotations de plus en plus nombreuses. Les premières bases de données en classification d'image ne comprenaient qu'une dizaine de classes au mieux. Le nombre de classe n'a fait que croître, dépassant généreusement les milliers de classes aujourd'hui. Cette tendance reflète la complexification des tâches à réaliser.
- Des tâches de plus en plus variées. Les premières bases de données se concentraient sur des tâches "simples", comme la classification d'image en contexte mono-classe (une annotation par image). Cette tâche s'est ensuite complexifiée en intégrant plusieurs annotations par image (multi-classe), puis d'autres tâches plus complexes se sont développées (détection d'image, localisation, segmentation, segmentation panoptique, ...)

Ces tâches de plus complexes ont aussi requis des processus d'annotation de plus en plus complexes.

2.1.4.3 L'annotation de données

L'annotation est devenu un sujet de plus en plus sérieux au fur et à mesure que les bases de données devenaient de plus en plus volumineuses et les données de plus en plus complexes. De simples classes attribuées à des images dans le cas de la classification, il a fallu déterminer comment définir un objet, segmenter une image, organiser les classes. Certaines méthodes de collecte génèrent automatiquement la classe associée à la donnée, en associant par exemple requête sur une plateforme, donnée et classe. Pour les autres données, il faudra les annoter par divers procédés : annotations manuelles (réalisées par des experts, récupérées via des procédés de vérification (comme les captcha) ou réalisées par des individus non experts comme via la plateforme Amazon Mechanical Turk), semi-automatiques (annotations par des processus algorithmiques avec vérification humaine) ou automatique (annotation entièrement automatisée par algorithme). Ces procédés ont divers avantages et inconvénients résumés dans le tableau 2. Ce tableau illustre que les procédés manuels fournissent des données de meilleures qualité, mais sont aussi plus coûteuses financièrement et temporellement que les solutions automatiques. Les solutions semi-automatiques sont un compromis entre les deux autres types de solution. Ces contraintes peuvent orienter le choix du type d'annotation choisi, bien que la nature même de l'annotation puisse être aussi prépondérante dans ce choix.

En classification, les annotations sont soit une seule classe (mono-classe) soit plusieurs classes (multi-classe) associées à une image. Les moyens d'associer classe et donnée sont nombreux (tableau de référence, nom de la donnée, autre format de stockage de données...), tout comme les moyens de représenter la classe (entier, liste, strings, ...). En détection d'objet, les annotations sont plus complexes, puisqu'elles doivent définir les contours d'un objet ou d'un concept dans une image. Les premières annotations étaient des boîtes englobantes, situant les limites de l'objet dans un rectangle positionné sur la donnée. D'autres types d'annotations sont apparues par la suite - suite de points,

valeur associée à un pixel, etc... Ces annotations sont plus coûteuses en terme de temps, de concentration des annotateurs, et surtout requièrent une annotation manuelle tant qu'aucun modèle n'est capable de réaliser des détections algorithmiques des éléments recherchés. Ces annotations sont réalisées avec des outils spécifiques (VIA[52], Extreme clicking [128]) qui eux aussi évoluent et se complexifient avec les bases de données, parfois regroupés au sein de plateformes (AMT par exemple).

Les importants volumes de données et les annotations de plus en plus complexes entraînent des coûts temporels et financiers de plus en plus importants. Ceci est un frein significatif dans la constitution de bases de données pour des tâches complexes en vision par ordinateur. Par nature, chaque classe ayant différents niveaux d'abstraction, de nouvelles annotations peuvent toujours êtres réalisés pour des applications s'intéressant à certains concepts spécifiques. Il est donc vain de s'imaginer pouvoir tout annoter de cette sorte, et c'est pourquoi de nombreux travaux se tournent de plus en plus vers les pratiques de généralisation hors domaine ou d'apprentissage few-shot ou zero-shot.

2.1.4.4 Les benchmarks en vision par ordinateur

Il a déjà été mentionné que les benchmarks étaient un apport considérable à la communauté, et ces derniers rendent de nombreux services aux communautés scientifiques :

- Ils permettent avant tout de comparer les pratiques sur une tâche et des données communes. L'avantage de partager tâche et donnée est de pouvoir faire une comparaison des approches indépendamment d'éléments extérieurs, et donc de déterminer quelles pratiques et quelles méthodologies permettent d'améliorer les performances des systèmes sur les tâches prédéfinies par ces benchmarks. Ceci permet la comparaison à l'état de l'art et donc souligne les travaux qui permettent de faire avancer le domaine. Les benchmarks sont donc un moteur important des améliorations de performance dans un domaine.

- Ils permettent aussi la mise en commun de méthodologies et de pratiques, qui sont sélectionnées par divers procédés, comme l'acceptation par l'usage, la comparaison et le choix raisonné d'une pratique plutôt qu'une autre, etc... Cette mise en commun des pratiques permet d'avoir une base de pratiques commune, ayant pour conséquence la facilitation d'accès au domaine et aux applications. Ces normes communes font partie d'un socle scientifique commun, dont le benchmark peut être l'instigateur.

- Ils peuvent aussi permettre de populariser certaines tâches ou domaines particuliers. Dans les domaines nouveaux ou peu explorés, les données et les pratiques sont souvent éparpillées, rendant le domaine peu populaire. Un benchmark permet de souligner l'intérêt porté à un domaine par une communauté, et de populariser les problématiques et les pratiques rattachées à ce domaine. Un domaine peut ainsi gagner en visibilité et ses problématiques se décliner plus facilement lorsque des benchmarks s'intéressent aux tâches ou données spécifiques de ce domaine.

- Il est courant que des concours soient organisés sur des tâches et des bases de données. Ces concours récompensent les meilleures modèles proposés, incitant ainsi la communauté scientifique à obtenir les meilleures performances possibles. Pour les laboratoires, les concours organisés via les benchmarks sont des opportunités pour augmenter leur notoriété tout en étant récompensé financièrement pour leurs efforts. Les benchmarks peuvent ainsi être utilisés pour stimuler la recherche scientifique dans une direction en particulier par ces moyens financiers.

- Enfin, autour de ces pratiques, normes, concours, se forment des communautés. Les communautés scientifiques rassemblées autour d'un domaine utilisent des benchmarks reconnus par ces communautés pour discuter des avancées dans le domaine et souligner les performances de nouvelles approches. Les benchmarks jouent donc aussi le rôle de lien et de repère pour des communautés spécifiques, et aident ainsi à la création de liens.

L'évolution des benchmarks a suivi celle des bases de données. D'abord concentrés sur les tâches de classification, les benchmarks reportaient les performances des différents modèles testés. Ainsi Yann Lecun, Corrina Cortes et Christopher J.C. Burges reportent les performances de tous les modèles testés sur la base de données MNIST et mettent les résultats en ligne, pour que tous les acteurs du domaine puissent s'informer des modèles à l'état de l'art [7]. Ces benchmarks ont servi de support pour des concours, comme Pascal VOC[57] de 2005 à 2012 et Imagenet[138] de 2010 à 2017. Ces concours, organisés chaque année, récompensaient les modèles à l'état de l'art qui arrivaient à performer mieux que les modèles concurrents, et incarnaient donc des incitations importantes pour les domaines couverts par ces benchmarks. La vision par ordinateur a ainsi pris de l'essort, pour les tâches de classification, de segmentation, de détection d'objet, et même de détection sur vidéo. Par la suite, des plateformes hébergeants des compétitions sont apparues, remplaçant les concours scientifiques. Kaggle[8] est la première de ces plateformes, suivie par Zindi[9] pour des concours et données concernant l'Afrique. Ces plateformes offrent l'opportunité pour les acteurs de l'intelligence artificielle de participer à des compétitions opposant des modèles construits à l'aide de données fournies par la plateforme. Au terme d'une échéance prédéfinie, les modèles sont comparés et les meilleurs sont récompensés. Ces plateformes ont permis de populariser encore plus les compétitions algorithmiques, tout en offrant une variété de concours et de domaines, ouvrant la possibilité pour des domaines spécifiques de gagner l'intérêt de la communauté scientifique. En parallèle de ces plateformes, de nouveaux benchmarks ont continué d'apparaître, soulignant l'importance de certaines domaines et tâches spécifiques. C'est le cas par exemple pour la généralisation de domaine, qui a vu plusieurs benchmarks se développer, popularisant le domaine et complexifiant les tâches à réaliser. DomainBed[69] regroupe ainsi plusieurs bases de données utilisées en généralisation de domaine pour normer l'utilisation de ces dernières sous un cadre commun. WILDS[92] propose ensuite une approche similaire pour les glissements de domaines relatifs à des données réelles et non générées en laboratoire. La progression réalisée par les modèles et les architectures sur les benchmarks permet de proposer de nouveaux benchmarks, avec des tâches et des données plus complexes.

2.1.4.5 Les architectures

Les architectures ont aussi évolué avec les bases de données et les benchmarks. De nombreuses ont été popularisées au cours des années, et nous présenterons ici les architectures les plus connues et les plus marquantes. Alexnet[93] est la première architecture qui a percé dans le milieu scientifique, en performant bien mieux que les autres modèles sur le concours ILSVRC en 2010. Composée de 5 couches de convolution suivies de quatre couches denses, c'est le premier réseau de neurones à avoir profité des capacités de GPU pour l'entraînement de ses poids. Cette prouesse a permis de rendre plus accessible l'entraînement de ces réseaux, couplée à l'accroissement des capacités computationnelles et de stockage. D'autres artefacts utilisés par Alexnet (le dropout, l'utilisation pour la première fois de la fonction d'activation ReLU, de l'augmentation de données).

En 2014 paraît VGG16[149], un modèle convolutionnel composé de 16 couches, qui améliore AlexNet en réduisant grandement les filtres utilisés pour la convolution (de 11*11 à 3*3). Cela permet d'avoir des couches plus petites et donc plus nombreuses à capacité computationnelle égale, ce qui améliore la capacité d'abstraction du modèle entraîné. Ceci permet d'améliorer grandement les performances sur de nombreuses tâches de vision par ordinateur, et l'architecture se montrera plus performante que les modèles à l'état de l'art sur les concours ILSVRC 2012 et 2013, mais elle ne remportera pas le concours ILSVRC 2014. Néanmoins, VGG16 a popularisé les filtres de taille 3*3 pour les neurones de convolution qui resteront utilisées par les autres architectures qui lui succéderont. Cette architecture est illustrée en figure 8.

Une autre architecture notable parue en 2015 est l'architecture ResNet, introduite par He et al. (2016). Cette architecture introduit la pratique du saut de connexion, qui permet à une information de

[7] http://yann.lecun.com/exdb/mnist/
[8] https://www.kaggle.com/
[9] https://zindi.africa/

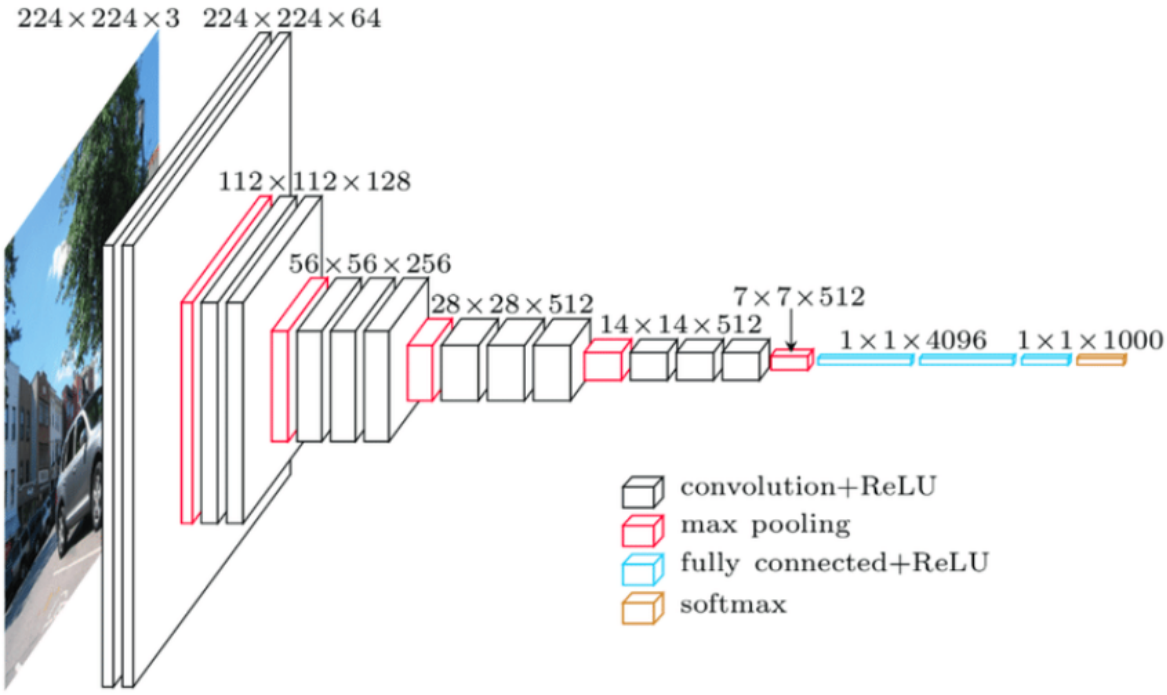


Figure 8: Illustration de l'architecture VGG16.

sauter une couche si cette couche endommage la performance du modèle, par le mécanisme illustré en figure 9. Cette architecture se décline en plusieurs tailles, dont nous avons vu l'architecture et les spécificités en section 2.1.3.3. Cette architecture a été utilisée comme norme et baseline dans de nombreux benchmarks et travaux, et a inspirée de nombreuses autres architectures.

MobileNet[81] est l'une d'elles. Sortie en 2017, MobileNet reprend ResNet en ajustant les calculs pour permettre un déploiement sur des terminaux embarqués comme des téléphones. Grâce à une séparation des convolutions sur ses deux axes, les calculs effectués sont bien plus efficients car grandement réduits. EfficientNet[153] se construira en 2019 sur les innovations de MobileNet en rajoutant des nouvelles fonctions d'activation et d'optimisation. Des architectures spécialisées pour répondre à des contraintes particulières (taille du modèle, capacité computationnelle, consommation énergétique, temps d'exécution [7]) commencent alors à foisonner, toujours basées sur les réseaux de convolution, jusqu'à l'arrivée des Transformers.

Les Transformers sont des modèles basés sur des mécanismes d'attention, utilisés pour la première fois en 2017 et d'abord populaire pour les problèmes de traitement automatique du langage[162]. Les mécanismes d'attention sont des mécanismes initialement prévus pour calculer la corrélation entre deux mots dans un ensemble de mots ou une phrase. Ces calculs sont réalisés à l'aide d'un triplet de matrice appelées Q, K et V (Query, Key, Value). On récupère pour chaque mot de la phrase sa représentation vectorielle dans un espace de mots, et on utilise ces représentations vectorielles pour calculer les corrélations entre les mots à l'aide de la formule décrite et illustrée en figure 10. Ces mécanismes d'attention sont intégrés dans des blocs d'attentions, dans lesquels elles sont combinées à des couches denses et des couches de normalisation. Les Transformers se reposent sur ces blocs d'attentions, combinés à des couches de transformation en représentation vectorielle et d'ajout d'un encodage de position, comme illustré en figure 11.

Il existe de nombreuses variations de l'architecture Transformer, et notamment une variante nommée Swin Transformer [108], spécialisée pour la vision par ordinateur et introduite en 2021. Cette architecture se base sur une construction hiérarchique de fenêtres glissantes pour adapter l'architecture initialement prévue pour le domaine du TAL au domaine de la vision par ordinateur. Les blocs classiques d'attention sont remplacés par des blocs basés sur ces fenêtes glissantes, l'attention étant calculée fenêtre par fenêtre. Il existe plusieurs variantes de cette architecture (Swin-T illustrée en figure 12 , Swin-S, Swin-B, Swin-L) qui correspondent à des tailles et hyper-paramètres différents.

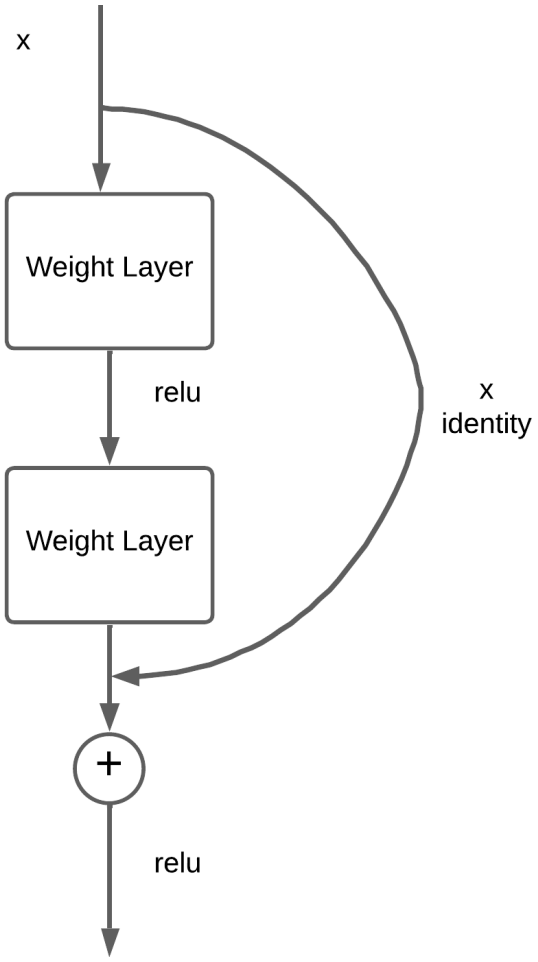


Figure 9: Illustration d'un saut de connexion, où l'image initiale est réintroduite après les calculs de la couche. Technique introduite par He et al. (2016).

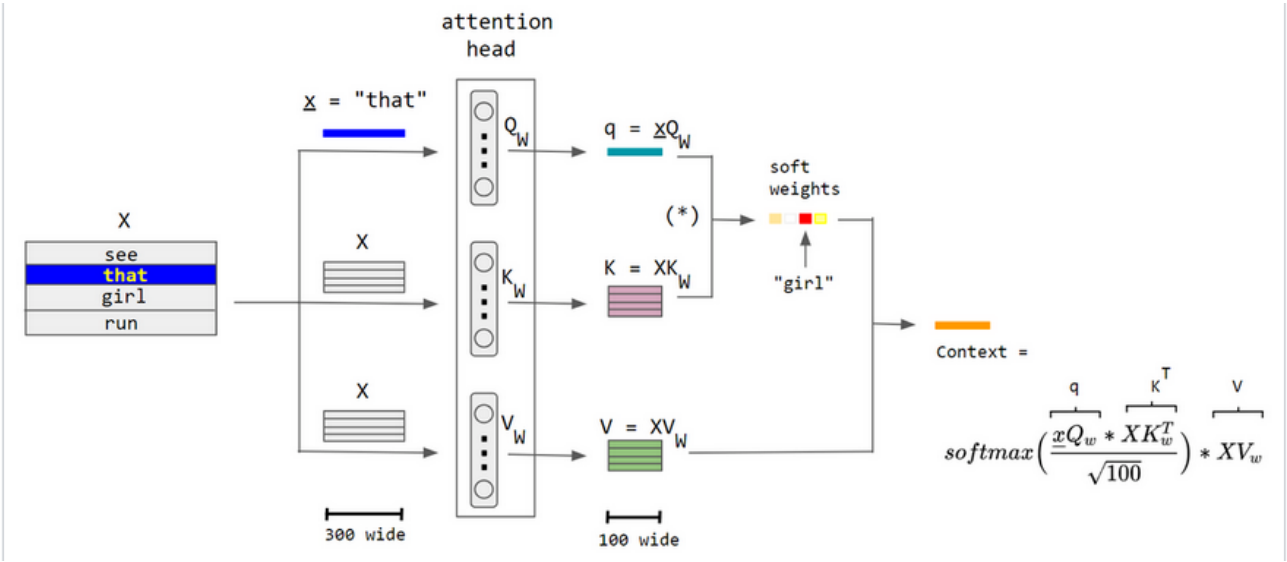


Figure 10: Mécanisme d'attention à l'origine de l'architecture des Transformer. Illustration issue de Vaswani et al. (2017).

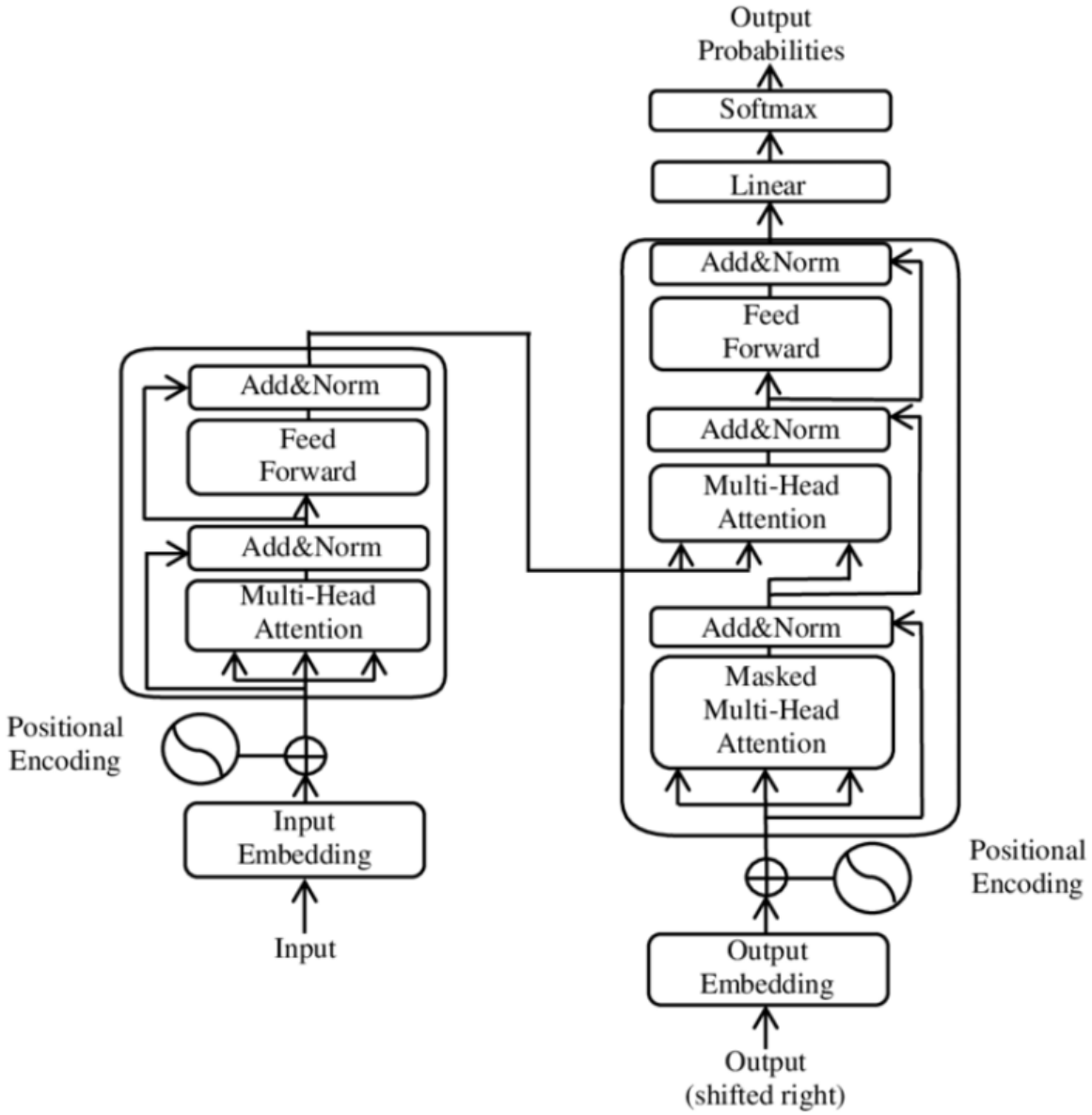


Figure 11: Architecture d'un Transformer. Illustration issue de Vaswani et al. (2017).

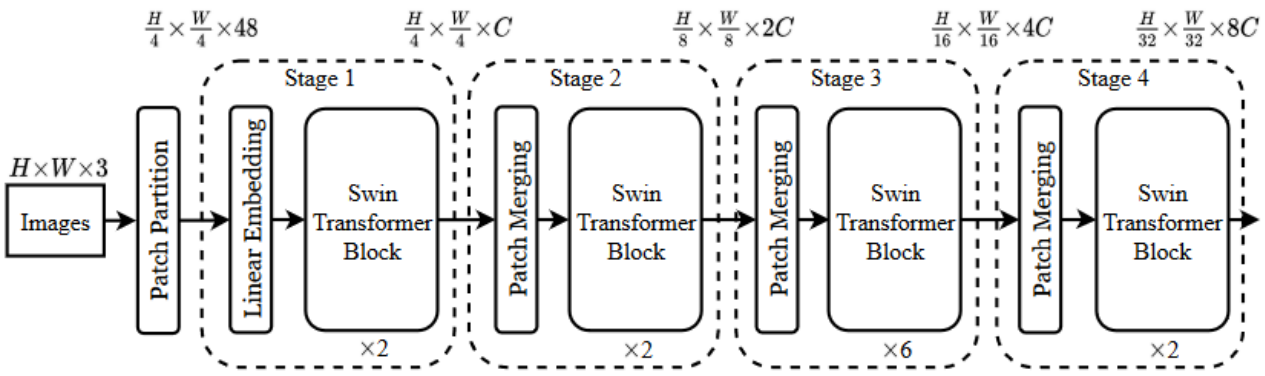


Figure 12: Architecture Swin-T proposée par Liu et al. (2021b). Illustration issue de leurs travaux.

Ces architectures continuent d'évoluer et de se complexifier pour être capable de répondre aux enjeux qui leur sont présentés : des bases de données de plus en plus grandes à intégrer, des tâches de plus en plus à résoudre. Elles sont aussi la base de modèles qui marquent des jalons d'évolution dans le domaine de l'intelligence artificielle, et que l'on appelle les modèles fondateurs.

2.1.4.6 Les modèles fondateurs

Le terme modèle fondateur est issu d'une publication de l'université de Stanford, et décrit "tout modèle formé sur de larges données qui peut être adapté à un large éventail de tâches"[31]. Ces modèles ont émergé grâce à l'augmentation massive et des données d'entraînement et des architectures utilisées pour l'apprentissage. Ils constituent un nouveau paradigme, où un modèle devient la base d'un ensemble d'applications sans besoin d'un ré-entraînement particulier, mais qui s'adapte à des nouvelles tâches par émergence. C'est le cas de CLIP[135], de GPT-3[34] et de BERT[48], et de bien d'autres modèles depuis. La portée de ces modèles est sans comparaison, car ces derniers offrent de grandes possibilités et des capacités d'adaptation hors du commun. De nos jours, les modèles à l'état de l'art sur de nombreuses tâches sont des adaptations de ces modèles fondateurs. Il n'y a donc pas de cadre dans lesquels ces modèles fondateurs ne pourront pas être déclinés ; il est attendu qu'ils changent le paysage de l'IA. Ces modèles viennent néanmoins avec certains risques. Par exemple, les biais d'un modèle fondateur se retrouveront dans de nombreuses déclinaisons et applications si ces derniers ne sont pas corrigés lors de l'adaptation. Un seul modèle fondateur peut ainsi causer beaucoup de tort si ce dernier est décliné sans égard pour ses aspects problématiques. La taille impressionnante de ces réseaux en fait aussi des artefacts difficiles à comprendre, et donc difficilement interprétable ; ce qui amène à des situations parfois critiques lorsque le modèle échoue dans une tâche sans que l'on puisse expliquer ou comprendre pourquoi. Enfin, ces modèles vont aussi généralement à l'encontre de l'esprit de partage qui régnait dans la communauté scientifique autour de l'IA, puisque les poids et détails d'implémentation ou d'entraînement ne sont généralement pas partagés ou explicités. Quand bien même les détails d'entraînement et les bases de données seraient-elles partagées, peu de structures sont en capacité d'entraîner ces modèles. Ces modèles creusent donc un certain écart entre les différentes structures actrices dans la communauté scientifique de l'IA.

2.1.4.7 Les systèmes de vision par ordinateur

Les modèles fondateurs comme de nombreux autres modèles proposés principalement par les grandes structures privées de la sphère IA ne sont pas disponibles en tant que tel. Les poids et les architectures ne sont pas partagées, et les modèles sont disponibles à l'utilisation via des outils en ligne ou des API. Ces outils prennent en entrée des données et fournissent en sortie les prédictions obtenues par les modèles demandés, fonctionnant comme des boîtes noires dont aucun paramètre n'est accessible. Néanmoins, ces outils permettent le déploiement rapide de systèmes de vision par ordinateur, avec pour seule contrainte une connexion internet et la mise en place d'une chaîne de traitement de données. Ces systèmes clés en main sont avantageux pour la rapidité de mise en place de solutions performantes, bien que le coût financier d'utilisation de ces plateformes puisse être dissuasif.

2.2 Les biais des pratiques classiques de l'apprentissage profond

L'intérêt pour les biais en AP a été encouragé par les nombreux préjudices causés par les systèmes d'AP déployés partout dans le monde[133; 35]. Ces préjudices sont de nature variable, les plus connus et étudiés aujourd'hui concernent des applications en justice prédictive[10], en reconnaissance faciale[35; 85; 146] ou en aide au recrutement[159]. D'autres applications omniprésentes dans nos vies sont aussi sources de préjudices, comme les moteurs de recherche[119], les systèmes de recommandations[96], et peut-être bientôt les applications en médecine[82]. Ces biais se retrouvent aussi dans les modèles fondateurs comme dans CLIP[6] qui comporte des biais de genre et de race, et

[10] voir https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

les outils tels que GPT-3 qui peut promouvoir des violations des droits humains[150] ou discriminer les personnes non issues de la culture anglo-saxonne[103].

Le terme de biais en AP est ainsi largement utilisé pour décrire un ensemble de concepts à connotation communément négative, et de multiples définitions tentent de cerner les limites du concept. Il se réfère traditionnellement aux présupposés d'un modèle spécifique, et Mitchell[111] définit en 1997 le biais inductif en AM comme "l'ensemble des hypothèses suffisantes pour interpréter les prédictions inductives du modèle comme prédictions déductives." Mais cette définition n'englobe pas tous les types de biais ; Ntoutsi et al. (2020) en proposent une autre, définissant le biais en AP comme "une inclinaison ou un préjudice contre une personne ou un groupe, découlant d'une décision réalisée par un système d'IA, spécifiquement d'une manière qui est considérée comme injuste". Suresh and Guttag (2021) soulignent que le terme de "biais" englobe aujourd'hui un certain nombre de concepts issus de différents milieux. Ils définissent pour leur part les biais comme "les sources distinctes de préjudice dans un système d'AP". Les biais sont une fois identifiés comme les préjugés internes d'un système, une autre fois comme les sources de ces préjugés, une autre fois encore comme les préjudices qui sont les conséquences de ces préjugés. Cette variation dans la définition du concept illustre la multiplicité des points de vue et des approches des chercheurs qui font face aux préjudices en AP.

En pratique, si on comprend un biais comme une inclinaison issue de facteurs non pris en compte dans une décision, toute donnée ou tout choix est biaisé, car issu d'un contexte socio-culturel particulier, et s'inscrit dans une temporalité particulière. Les biais sont donc intrinsèques aux systèmes d'AP ; mais tous ne mènent pas à des conséquences négatives. Les préjugés engendrés par les biais peuvent ne pas avoir de conséquences, ou au contraire être bénéfiques aux systèmes d'AP. On peut ainsi souhaiter qu'un système en AP présente certains biais pour certaines applications. Dans le reste du manuscrit, nous considérerons les biais comme "les inclinaisons présentes dans un systèmes d'AP".

2.2.1 Identification et cadres

L'étude des biais dans les systèmes d'AP a deux justifications. La première est l'impact des biais sur les performances des systèmes. Certaines inclinaisons découlant des choix dans les données ou les modèles peuvent avoir un impact négatif sur les performances des modèles ; l'étude des biais à l'origine de ces désagréments a donc pour but l'amélioration des performances finales du systèmes d'AM. La deuxième justification est l'impact des biais sur des populations ou groupes ciblés par le système. Ces inclinaisons n'ont pas forcément d'impact sur les performances des systèmes, mais portent préjudices à des groupes ou populations avec lesquelles le système interagit.

Pour pallier aux préjudices causés par ces biais, il est nécessaire d'identifier ces biais, leurs sources, la manière dont ils s'inscrivent dans le système, et de bien cerner les conséquences qu'ils impliquent. De nombreux travaux se sont penchés sur cette identification des biais, et plusieurs cadres existent aujourd'hui pour traduire les observations de préjudices d'un système en catégories particulières de biais. Ces cadres sont la plupart du temps basés sur une décomposition de la chaîne de traitement du développement d'un système en ses différentes composantes, permettant une granularité plus fine et une approche plus systématique, défendue par Balayn et al. (2021). Nous présentons ici trois cadres pour la classification de biais dans un système d'AM, qui illustrent des approches différentes pour la compréhension du concept même de biais.

2.2.1.1 Les cadres pour l'identification de biais

Le premier cadre est celui proposé par Ntoutsi et al. (2020). Dans ces travaux, les auteurs passent en revue les travaux de recherche autour du biais en AM, et classent ces travaux en trois catégories :

- **Comprendre le biais.** Cette catégorie regroupe les approches qui aident à comprendre comment les biais s'inscrivent dans nos sociétés et entrent dans nos systèmes socio-techniques, comment ces biais se manifestent dans les données utilisées par les systèmes d'IA, et comment ces derniers peuvent être définis et modélisés.

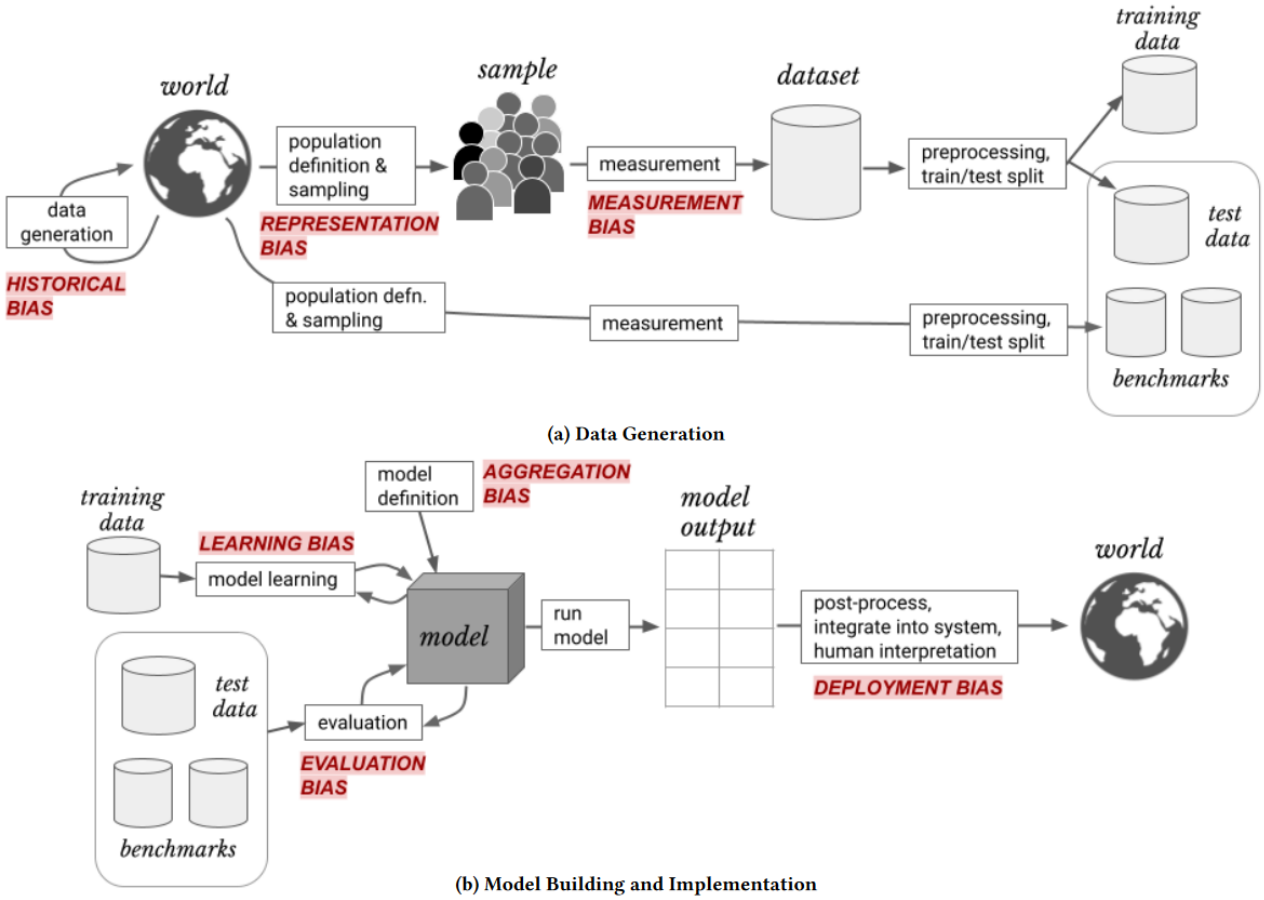


Figure 13: Cadre présenté par Suresh and Guttag (2021) pour comprendre l'origine des biais dans un système d'IA, découpé en deux parties : la génération de données (a) et la construction et l'implémentation de modèles (b). crédit @Harani Suresh

- **Mitiger le biais.** Cette catégorie regroupe les approches qui tentent de pallier les conséquences négatives d'un biais en s'attaquant à ce dernier durant les différentes étapes de conception et développement du système d'IA.
- **Prendre le biais en compte.** Cette catégorie regroupe les approches qui assument la présence du biais et proposent de modifier le développement des systèmes d'IA en conséquence ou d'en expliquer les sorties.

Pour chacune de ces catégories, les auteurs observent comment les initiatives les plus récentes s'inscrivent dans le cadre légal qui existe autour de la gestion de données privées, du déploiement de modèles et de la prise en compte de préjudices portés aux utilisateurs des systèmes déployés. Ce cadre propose une vue générale des approches sur les biais en AM, et est utile pour comprendre les différentes approches techniques entreprises par les chercheurs en réaction à la présence de biais dans les systèmes d'IA.

Le second cadre présenté ici est proposé par Suresh and Guttag (2021) et est illustré en figure 13. Ce cadre propose de décomposer la chaîne de traitement du développement des systèmes d'IA et d'identifier les différentes étapes où les biais peuvent apparaître, afin de situer ces derniers. Ils proposent ainsi une classification des biais en 7 catégories, qui correspondent aux différentes étapes de la chaîne de traitement.

Dans ce cadre, les **biais historiques** sont ainsi dus aux biais présents dans nos sociétés, dont les données reflètent une partie spécifique. Les **biais de représentation** apparaissent quand certaines populations concernées par le système d'IA sont sous-représentées dans l'échantillon de données de ce dernier. Les **biais de mesure** surviennent quand les caractéristiques collectées sont insuffisantes pour mesurer précisément et prédire le résultat attendu. Les **biais d'aggrégation** sont dus à une ab-

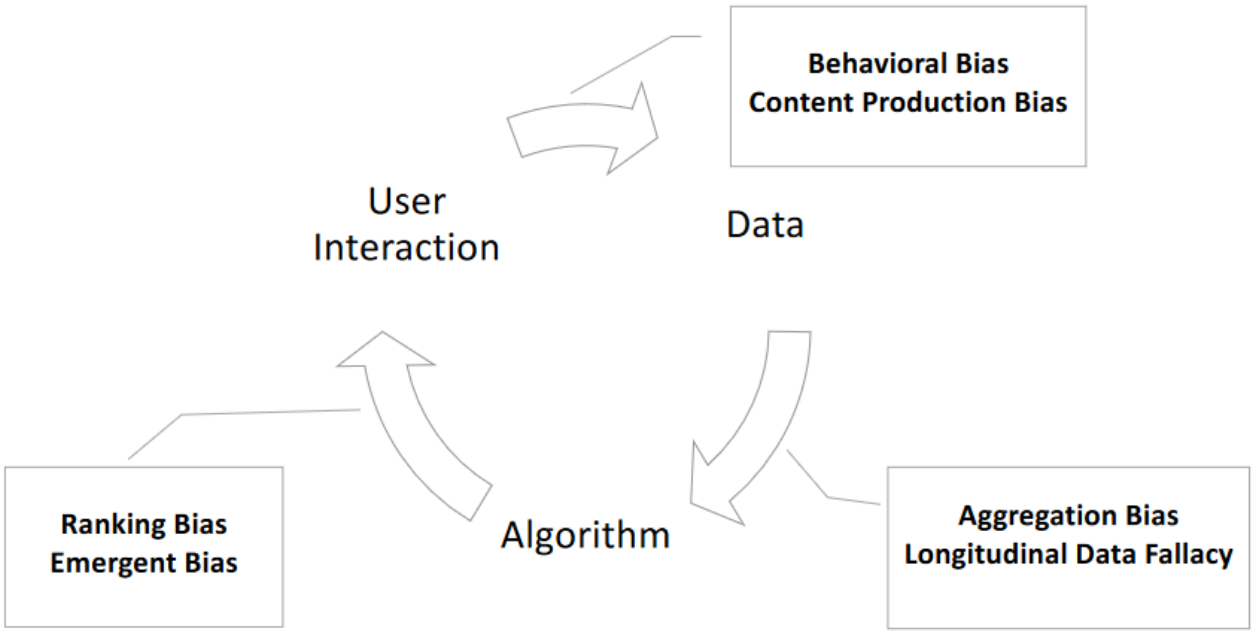


Figure 14: Cadre proposé par Mehrabi et al. (2022) pour la classification de biais. crédit @Ninareh Mehrabi

straction de singularités dans les données qui entraînent la mauvaise interprétation d'un système pour des groupes spécifiques. Les **biais d'apprentissage** englobent les différents choix réalisés dans le processus d'apprentissage d'un modèle, et particulièrement le choix de la fonction objectif du modèle. Les **biais d'évaluation** peuvent apparaître quand les systèmes d'IA sont testés sur des données qui ne représentent pas les cas d'usage du système. Enfin, le **biais de déploiement** soulève des problèmes lorsque le modèle est utilisé d'une manière qui ne correspond pas au problème pour lequel le système est initialement déployé.

Les auteurs soulignent qu'un biais particulier peut être classifié dans différentes catégories selon les hypothèses de départ du problème et les points de vues des acteurs de la démarche d'identification du biais. Néanmoins, ce cadre est utile pour repérer d'où provient un biais particulier et comment prendre ce biais en compte, pour mitiger ses effets ou ses causes. Ils formalisent ce cadre pour permettre une caractérisation rigoureuse des opérations effectuées par les méthodes de mitigation de biais.

Le troisième cadre étudié est proposé par Mehrabi et al. (2022), et souligne les interactions qui donnent naissance aux biais et qui amplifient ces derniers, entre les données, les algorithmes et les utilisateurs (voir la figure 14). Ce cadre permet de comprendre comment les biais d'une entité influent d'autres entités, et peuvent renforcer d'autres biais. Les auteurs utilisent ensuite ce cadre pour explorer la discrimination comme un ensemble de biais spécifiques dus aux préjudices et stéréotypes humains, et présentent les solutions algorithmiques basés sur des algorithmes "justes" pour pallier ces préjudices.

2.2.1.2 Source des biais

Ces cadres de classification sont utiles pour comprendre la nature des biais et là où ils apparaissent dans la chaîne de traitement, mais ne permettent pas d'identifier la source même du biais, sa raison d'être. Pour beaucoup de biais, cette source sera d'ordre socio-culturelle ; les choix réalisés dans la conception des bases de données et des modèles sont réalisés par des humains évoluant dans un contexte social, environnemental, technique et culturel spécifique qui justifie en grande partie ces choix. On peut néanmoins tenter de spécifier ces contextes pour identifier les causes par catégorie : sociale, technique, pratique, historique, etc... La source d'un biais peut aussi être multiple, fruit d'une conjonction de situations et contextes, auquel cas il devient difficile d'extraire une raison ou source

particulière pour le biais cible.

Pour étudier plus facilement les biais dans les données, il est aussi possible de créer des biais artificiels dans des bases de données et de se concentrer sur ces biais en particulier. C'est ce que font de nombreux scientifiques pour éprouver les méthodologies de mitigation de biais. Les plus populaires sont regroupés dans des benchmarks[69; 10; 166] afin d'éprouver les méthodologies dans différentes situations. D'autres approches proposent d'étudier les biais dans des données réelles ; c'est le cas de WILDS[92], DomainNet[131], ou l'Inclusive Image Dataset [11]. Le biais cible est alors une variation dans la nature des données, mais ce dernier n'est pas isolé du reste des biais présent dans les données.

2.2.1.3 Isoler un biais

La méthodologie classique pour l'étude d'un biais est l'isolation de ce dernier et l'étude l'impact de celui-ci sur les performances du modèle ou sur d'autre métriques spécifiques[133]. Mais l'isolation d'un biais est une affaire compliquée par l'omniprésence de biais dans chacun des choix réalisés aux différentes étapes de conception d'un système. Une pratique courante consiste alors à séparer une base de données en deux : des données ne comportant pas le biais cible et d'autres contenant ce dernier. Si on fait l'hypothèse que la seule différence entre ces deux jeux de données est la présence du biais que l'on souhaite isoler, alors l'étude de métriques adaptées obtenues sur chacun des jeux permet de mettre en évidence l'impact du biais.

Cette stratégie est très utilisée, mais se repose sur l'hypothèse que l'on est capable d'identifier totalement un biais et de séparer une base de données en fonction de la présence du biais cible. C'est difficilement concevable quand la portée du biais est difficile à cerner, quand le biais est un concept sans définition scientifique précise ou qu'il n'existe pas de métriques permettant de témoigner précisément de la présence du biais.

2.2.2 Les différents biais en apprentissage profond

Nous présentons des biais parmi les plus représentés dans la littérature scientifique, une liste exhaustive étant impossible à produire.

- **Le biais historique.** *Ce biais survient si le monde ou l'environnement dans lequel les données sont collectées ou générées comporte des biais.* Un modèle entraîné sur des données comportant un biais historique renforce souvent le biais concerné. Par exemple, un modèle prédisant le salaire d'un individu aurait tendance à prédire un salaire moins important pour les femmes par rapport à celui prédit pour les hommes, à cause de la différence historique de salaire entre les genres.

- **Le biais de représentation.** *Le biais de représentation provient de la manière dont les données sont échantillonnées dans les populations concernées.* Un modèle entraîné sur des données comportant un biais de représentation peut être performant pour certains groupes dans ces données mais peu performant sur des groupes moins représentés. Par exemple, un modèle permettant de conduire automatiquement des voitures pourra être moins performant la nuit si les données d'entraînement ne comportent que des données obtenues le jour.

- **Le biais temporel.** *Ce biais survient si des données sont récoltées sur un intervalle de temps suffisamment grand pour que différentes cohortes soient représentées dans les données. Certaines tendances dans les données peuvent être dues à des comportement spécifique dans ces cohortes plutôt qu'une évolution temporelle globale.* Un modèle entraîné sur des données comportant un biais temporel aurait tendance à inférer un comportement global sur des cohortes spécifiques.

- **L'omission de variable.** *Ce biais survient quand des variables sont omises lors de la conception du modèle.* Un modèle conçu avec ce biais pourra manquer complètement son but et ne

jamais obtenir de bonne performances. Par exemple, un modèle faisant le lien entre l'utilisation des SMS et le degré de communication entre les populations pourrait identifier une baisse notoire dans la communication avec l'expansion d'internet, alors que les communications sont simplement passées par d'autres médias, qui sont omis du modèle.

- **Le biais de mesure.** *Ce biais provient de la manière dont sont décidées, construites et utilisées les caractéristiques d'un modèle.* Un modèle entraîné sur des données comportant un biais de mesure pourra manquer de précision dans les concepts utilisés et porter de fait préjudice à certains groupes. Par exemple, l'outil de prédiction de récidivisme COMPAS[11] utilisé par la justice américaine pour aider les juge à décider de l'incarcération de prévenus attribuait un score de risque de récidivisme à chaque prévenu, mais ce score de risque était aussi corrélé à la fréquence des contrôles de police, qui vise plus souvent certaines populations. Ces populations avaient alors injustement un score de risque de récidivisme plus élevé que les autres, à cause de la mauvaise qualité du concept de risque de récidivisme.

- **Le biais d'agrégation.** *Ce biais surgit quand un modèle tire de mauvaises conclusions sur des individus à partir de l'observation d'une population.* Un modèle présentant un biais d'agrégation pourra donc produire de mauvaises sorties pour des individus spécifiques. Par exemple, beaucoup de modèles utilisés dans la santé peuvent faire des erreurs s'ils se concentrent sur des caractéristiques générales d'un groupe de patient, et pas sur un patient spécifique.

- **Le biais d'apprentissage.** *Ce biais survient quand les choix réalisés dans le design de la phase d'apprentissage provoquent des disparités de performance pour des groupes spécifiques.* Un modèle soumis à des biais d'apprentissage pourra avoir des performances atténuées ou variables selon les groupes présents dans les données. Par exemple, les phénomènes d'underfitting et d'overfitting sont des conséquences du biais d'apprentissage. L'implémentation de stratégies pour lutter contre ces phénomènes relève de la mitigation de ces biais.

- **Le biais de popularité.** *Ce biais survient quand certains items ou catégories dans un ensemble sont plus populaires que d'autres.* Un modèle présentant un biais de popularité aura tendance à présenter des résultats mettant en avant les éléments populaires, sans pour autant que cela soit un gage de qualité. Par exemple, les système de recommandation de musique pourront mettre en avant des musiques populaires à des personnes qui ne sont pas intéressées par ces dernières.

- **Le biais d'évaluation.** *Ce biais peut apparaître dans la phase d'évaluation du modèle, quand les données utilisées pour tester le modèle ne reflètent pas les données sur lesquelles le modèle sera déployé.* Un modèle confronté à un biais d'évaluation pourra présenter de très bonnes performances en laboratoire mais de mauvaises performances une fois déployé sur le terrain. Par exemple, les performances des premiers modèles de reconnaissance faciale étaient moins bonnes sur les populations de couleur par rapport aux populations caucasiennes, et moins bonne pour les femmes par rapport aux hommes, à cause d'un problème de représentation dans les données utilisées pour tester les modèles.

- **Le biais de déploiement** *Ce biais survient lorsque le modèle est utilisé d'une manière qui n'était pas prévue lors de sa conception.* Un modèle présentant ce type de biais pourra présenter des prédictions avec une confiance sans rapport avec la situation considérée. C'est par exemple le cas pour un modèle qui sera entraîné à classifier des plantes dans une région géographique et qui sera déployé dans une autre région géographique, devant réaliser des prédictions sur une flore différente.

- **Le biais social** *Ce biais survient lorsque les données sont générées par des utilisateurs sous des conditions qui les incite à valider un biais social.* Un modèle soumis entraîné sur une base de données comportant un biais social aura tendance à reproduire ce biais social dans

[11] voir https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ses sorties. Par exemple, un modèle qui sera entraîné à prédire les résultats d'un vote pourra prendre en entrée les résultats de sondages présentant toujours un même candidat en tête de liste, ne prenant pas en compte que ces sondages sont réalisés dans des environnements favorables à ce candidat. Le modèle aura alors tendance à prédire le candidat gagnant à cause de ce biais social.

Tous ces biais donnent lieux à différents types de conséquences, identifiées dans les prédictions réalisées par ces modèles d'IA, dont par exemple les glissements. Un glissement correspond à une variation entre données d'entraînement et données de test[114], dont on peut lister parmi les plus étudiés :

- **Le glissement de domaine** : Ce glissement survient quand les données d'entraînement et de test appartiennent à des domaines différents. La notion de domaine est vague et se rattache à la nature et le contexte socio-culturel des données. Un domaine est généralement considéré comme une distribution jointe entre un espace d'entrée X et un espace de sortie Y, sur laquelle une base de donnée est échantillonnée. De fait, les autres glissements présentés ci-dessous peuvent tous être considérés des glissements de domaines.
- **Le glissement de distribution** : Ce glissement survient quand la distribution des données d'entraînement diffère des données de test d'un modèle, et a pour conséquence une baisse de la précision d'un modèle.
- **Le glissement de contexte** : le glissement de contexte est défini par Kalluri et al. (2023) comme les changements dans le contexte associé à un objet ou une scène entre différentes bases de données. Par exemple, certains objets ou catégories dans les bases de données d'images sont très corrélées à un environnement particulier, et les modèles peuvent apprendre à reconnaître ces environnements. Changer un objet d'environnement reviendra à changer son contexte.
- **Le glissement de population** : Ce glissement survient quand les distributions de populations dans une base de données sont différentes entre les jeux d'entraînement et de tests, provoquant un glissement dans l'importance portée à certaines données entre les deux phases.
- **Le problème de longue traîne** : Ce problème illustre un biais de répartition dans les données qui invisibilise une partie des données qui manque de représentativité. Par exemple, dans les bases de données de classification, certaines catégories peuvent être bien plus présentes que les autres, créant le problème de longue traîne : de nombreuses catégories avec peu d'échantillons pour l'entraînement. Dans leurs travaux, Gupta et al. (2019) soutiennent que le problème de longue traîne est inéluctable, puisque l'annotation de plus de données révélera simplement de nouvelles catégories rares, non vues précédemment.

Les biais et les glissements qu'ils entraînent pourront avoir divers impacts sur un système d'IA. Ceux-ci sont explorés dans la sous-section suivante.

2.2.3 Impact d'un biais sur un système d'IA

L'impact d'un biais est le témoin qui le met généralement en évidence ; c'est généralement par la découverte de ce dernier que le reste de la caractérisation du biais est entreprise. L'impact d'un biais sur un système s'observe soit via les performances du système, soit par une autre métrique qui peut possiblement constituer un nouvel indice de performance pour le système testé.

2.2.3.1 Impact des biais sur la performance des systèmes

On peut observer un biais via les performances du systèmes quand ce biais influe directement sur la métrique de performance du système. L'impact du biais sera de manière générale évalué en comparant deux versions du système, l'une comprenant le biais cible, l'autre non. Ce qui peut être modifié dans la

version d'un système est une base de données (d'entraînement, de test, d'évaluation), un algorithme, une méthode d'apprentissage, le contexte de déploiement ; en bref, tout ce qui relève d'un choix dans la conception du système et qui peut potentiellement introduire le biais ciblé. L'intérêt réside dans le fait que la variation d'un système à l'autre isole le biais (en agissant sur la source ou le type de ce dernier), de manière à ce qu'il soit présent dans un système et pas dans l'autre, et que les variations de performance puissent être expliquées par la présence ou l'absence de ce biais, et pas d'autres circonstances. Ainsi, dans le cas de deux systèmes différents par la présence d'un biais cible S_1 et S_2, on évaluera l'impact du biais par la différence de leurs performances :

$$\text{Impact d'un biais} = R_{S_1} - R_{S_2} \text{ où } R_S \text{ est la performance du système S.}$$

Cette mesure se base sur les hypothèses suivantes :

- Le biais que l'on cherche à mesurer est présent dans un des systèmes mais pas dans l'autre .
- Il n'y a pas d'autres facteurs qui permettent d'expliquer les variations de performances que la différence provoquée par la présence ou l'absence du biais.

Dans le cas du glissement de domaines présenté en section 2.2.2, on témoigne de l'impact du biais cible en comparant les performance du système sur deux jeux de données séparés par la présence ou non du biais. Le jeu sans biais sera considéré "En Domaine" (ED) et le jeu avec biais "Hors Domaine" (HD). Si on formalise cette approche, une base de données X pourra être séparée en deux sous-ensembles X_{ED} contenant les données ED et X_{HD} contenant les données HD. L'impact du biais sur un système utilisant un modèle d'IA est généralement entendu comme la différence de performance d'un modèle sur ces deux jeux de données :

$$\text{Impact d'un biais} = R_F(X_{ED}) - R_F(X_{HD})$$

Ce phénomène de glissement de domaines coïncide souvent avec des déploiements de modèles dans des conditions différentes des essais et tests réalisés en laboratoire[92], impliquant un biais de déploiement ou d'évaluation selon la classification de Suresh and Guttag (2021).

Dans le but de faire une analyse rigoureuse des approches foisonnantes autour de l'étude de biais en vision par ordinateur et de regrouper les différentes initiatives sous des normes communes, Gulrajani and Lopez-Paz (2021) construisent un benchmark pour mettre en valeur les bases de données et les approches visant à pallier l'impact de données HD sur des modèles d'IA. Ils proposent un ensemble de bases de données présentant des glissements de domaines, et une implémentation en python de multiples modèles et méthodes appliquées à ces bases de données, ainsi qu'une analyse de leurs performances. Ils soulignent l'importance de préciser la méthodologie choisie pour l'étape de sélection de modèle, et encouragent les autres chercheurs à porter attention à ces méthodologies.

Les travaux de Koh et al. (2021) font écho à ceux de Gulrajani and Lopez-Paz (2021), en proposant une initiative similaire mais plus ciblée. Ils regroupent à leur tour différentes initiatives sous un benchmark pour les regrouper sous des normes communes, mais cette fois-ci ces initiatives ciblent des déploiements de modèles en situation réelle. Ils mettent en valeur le glissement de domaine qui s'effectue lorsqu'un modèle développé et testé en laboratoire est déployé dans un cas d'utilisation sur le terrain. Ce benchmark propose un ensemble de bases de données présentant des glissements de domaines avec la particularité de refléter le biais de déploiement sur le terrain, ainsi qu'un ensemble de modèles utilisés pour mitiger les conséquences de ce biais et une analyse de ces derniers. Ils argumentent dans ces travaux que la mesure de l'impact des biais sur les modèles ne doit pas se faire à partir des données ED, mais des données HD. Cette argumentation se base sur le fait que les données ED et les données HD peuvent avoir une difficulté différente, et la mesure de la différence entre la performance d'un modèle sur des données ED et des données HD peut être sur-évaluée ou sous-évaluée à cause de cette différence. Ils proposent donc de mesurer l'impact d'un biais à partir de l'équation suivante :

$$\text{Impact d'un biais} = R_{F_{\Theta_{X_{HD}}}}(X_{HD}) - R_{F_{\Theta_{X_{ED}}}}(X_{HD})$$

Ils préconisent donc pour évaluer l'impact du glissement de comparer les performances d'un modèle entraîné sur les données HD et testé sur ces données, aux performances d'un modèle entraîné sur les données ED et testé sur les données HD (voir l'illustration suivante en figure 15)

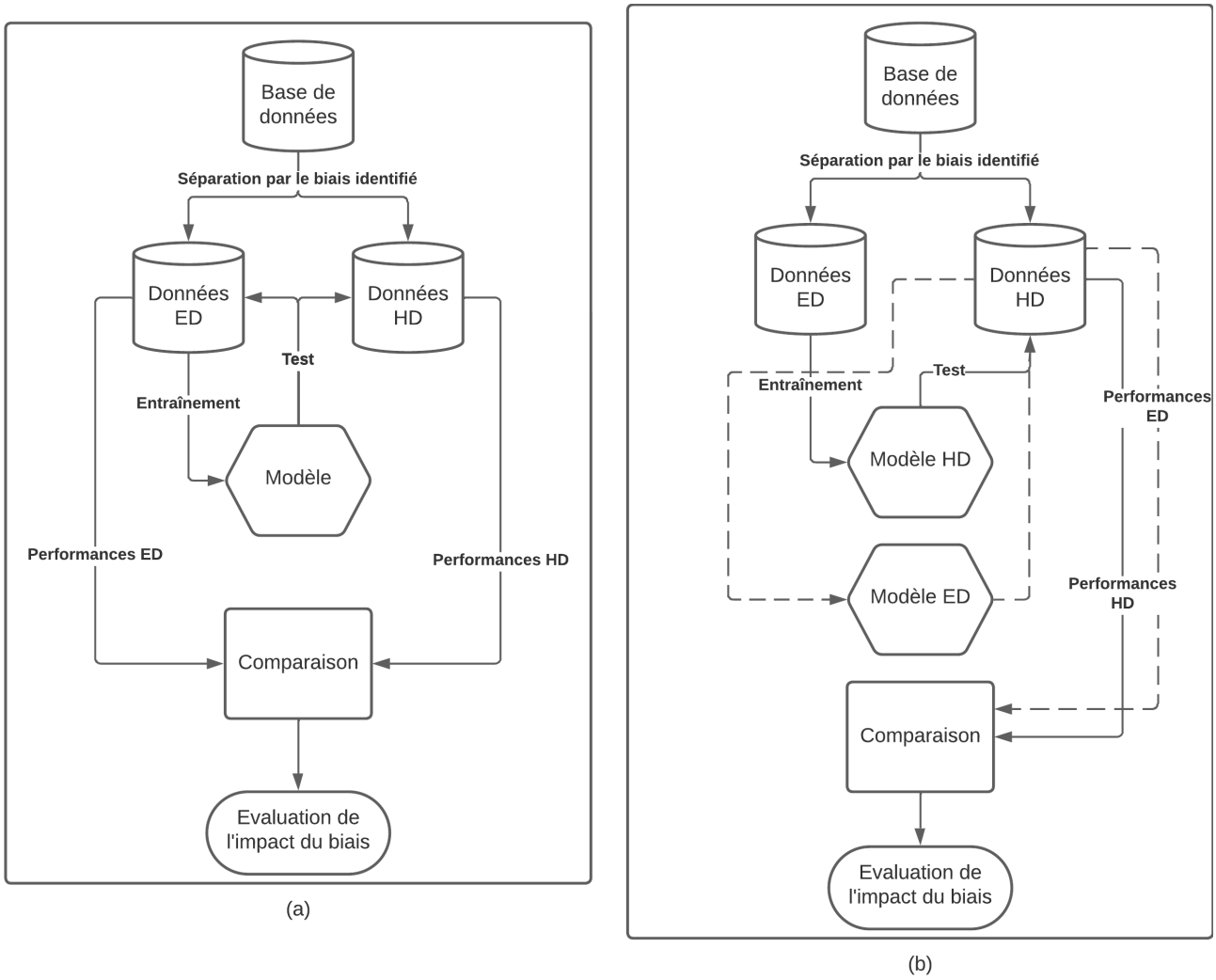


Figure 15: Les différentes configurations pour évaluer l'impact d'un biais. (a) Configuration classique, on mesure pour un modèle entraîné sur les données ED la variation de performance entre données ED et HD. (b) Configuration proposée par Koh et al. (2021), où on mesure la performance de deux modèles différents sur les données HD, l'un entraîné sur les données HD, l'autre sur les données ED. Cette dernière configuration permet de prendre en compte la variation de difficulté des différents domaines.

Cette stratégie est possible sous certaines conditions :

- Le biais dois être assez bien défini pour identifier les données ED et les données HD.
- Il doit y avoir assez de données HD pour entraîner un modèle.
- On doit pouvoir entraîner un second modèle pour comparaison au premier.

Dans le cas où le biais ne peut être isolé et où on ne peut donc construire les deux jeux X_{ED} et X_{HD}, on peut faire varier l' amplitude du biais ou ses manifestations dans différents ensembles de données dont on comparera les performances pour lier les variations de performances entre systèmes aux variations entre systèmes. A ce moment là, on disposera de k jeux $X_1, ..., X_k$, et plusieurs stratégies sont possibles pour l'évaluation de l'impact du biais :

- utiliser la performance moyenne : un modèle F a une performance moyenne $R_F = \frac{1}{k}\sum_{i=0}^{k}(R_F(X_i))$. Cette mesure a pour intérêt de donner une même importance à chacun des jeux de données X_i et d'encourager à une augmentation de la performance des modèles sur l'ensemble des données.

- utiliser la pire performance : un modèle F a la pire performance $R_F = min\{R_F(X_i)\}_{i\in\{0..k\}}$. Cette mesure a pour intérêt de donner une importance primordiale aux jeux de données le plus discriminé par le modèle F.

Dans le cas où le biais ne donne pas lieu à un glissement de domaines ou à une variation dans les jeux de données, d'autres stratégies sont préférées pour évaluer son impact :

- la variation de modèles : on peut décider de modifier certains étapes de la construction ou du développement de modèles pour améliorer ses performances et surmonter le biais de manière algorithmique. Cette pratique peut être mise en oeuvre par exemple dans le cas de biais d'apprentissage, d'agrégation, d'évaluation. L'impact du biais est alors mesuré selon la comparaison d'un modèle R_1 comportant le biais cible et d'un modèle R_2 conçu avec des variations algorithmique grâce à la formule $R_1S - R2_S$.

- la variation du déploiement : on peut décider de modifier certaines étapes du processus de déploiement du modèle pour observer comment le modèle se comporte dans ces différents contextes. De même, l'impact du biais est alors mesuré selon la formule $R_1S - R2_S$, avec R_1 et R_2 les performances du modèle dans deux situations de déploiement différentes.

Le choix d'une configuration pour la mesure de l'impact du biais dépend donc de plusieurs facteurs, et il sera nécessaire de préciser les choix identifier l'impact d'un biais dans chacune des situations rencontrées.

2.2.3.2 Impact des biais sans conséquences sur la performance

Les performances des modèles ne prennent pas toujours en compte toute la mesure de la portée du modèle. Des modèles peuvent ainsi être considérés performants sur les métriques testées et une fois déployés sans pour autant être performants selon d'autres considérations. Une illustration pratique de ce fait en particulier est illustré par une notion qui a pris de l'importance dans la dernière décennie en AM : la justice algorithmique. Nous présentons ici une rapide vue d'ensemble du milieu de la justice algorithmique pour comprendre et illustrer comment des biais peuvent être sans conséquences sur la performance des modèles et malgré tout source de préjudice. La justice est un concept qui a longuement été exploré en philosophie et en psychologie, et il n'en ressort aucune définition universelle. Le concept de justice algorithmique, qui correspond à l'application d'éléments de justice à des systèmes et algorithmes informatique, est un concept tout autant vague et difficile à cerner. Plus de 20 définitions différentes existent dans la littérature scientifique[163]. Ces nombreuses définitions reflètent les différentes manières de concevoir la notion de justice algorithmique, en regard des modèles, des données, des situations ou des utilisateurs. De manière générale, la justice algorithmique est entendue comme l'absence au sein d'un système informatique de préjudices ou de favoritisme à l'intention d'individus ou de groupes basés sur certaines de leurs caractéristiques intrinsèques ou acquises.

Des nombreuses définitions de la justice algorithmique découlent de nombreuses approches, mesures, et formalisation de cette notion. Il existe même plusieurs systèmes de classifications de ces différentes définitions. Verma and Rubin (2018) classent par exemple les différentes définitions de la justice algorithmique en cinq catégories définies par les caractéristiques sur lesquelles se base la définition du concept :

- les prédictions : certaines définitions de la justice algorithmique ne se basent que sur la prédiction d'un système.

- les prédictions et les annotations : ces approches combinent les prédictions d'un système aux réponses attendues pour définir la notion de justice algorithmique.

- les probabilités prédites et les annotations : certaines définitions se basent sur les probabilités en sortie de modèle combinées aux sorties attendues pour construire le concept de justice algorithmique.

- les similarités : certaines approches considèrent qu'un système fait preuve de justice algorithmique lorsque ce dernier produit des sorties similaires quand il est confronté à des entrées similaires.

- le raisonnement causal : certaines approches considèrent qu'un système dont on peut capturer les relations entre certaines caractéristiques et leur impact sur les sorties est juste algorithmiquement si il n'y a pas de relations préjudiciables observées.

Parmi ces dernières, les trois premières reposent sur la définition d'un ou plusieurs attributs dit "protégé" ou "sensible", lié à la notion de justice, comme la race, le genre, le sexe, le niveau de revenu, etc... Les approches basées sur la similarité reposent sur tous les attributs, pas uniquement ceux considérés protégés. Enfin, le raisonnement causal repose sur des conditions appliquées aux relations entre entrées et sorties du modèle.

D'autres classifications proposent de classer les différentes conceptions de la justice algorithmique en fonction des considération sur la portée de la métrique[109; 133; 63; 37] :

- justice individuelle : les prédictions d'un système doivent être les mêmes pour des individus similaires.

- justice pour les groupes : les différents groupes au sein de la population en contact avec le système doivent être traités de manière équitable.

- justice pour les sous-groupes : une alliance des deux premières notions pour obtenir de meilleures sorties.

Cette classification met en lumière les différences de conception d'une notion de justice ; et les différentes échelles qui sont à prendre en compte. La prise en compte d'individus ou de groupes, de sous-groupes partageant des caractéristiques particulières, est une complication dans la notion de justice même, avant la justice algorithmique.

D'autres approches, enfin, classent les travaux sur les manières de construire des systèmes justes en fonction de leur place dans la chaîne de traitement du développement d'un système[125; 121; 109; 37; 84] :

- avant computation : ces méthodes promeuvent d'agir au niveau des bases de données pour explorer et retirer les biais présents dans ces dernières.

- pendant computation : ces méthodes forcent les systèmes à produire des sorties justes grâce à des interventions dans le mécanisme d'apprentissage du modèle d'AM.

- après computation : ces méthodes agissent au niveau des sorties du modèle, en modifiant ces dernières de manière à produire des sorties justes.

Ces différents systèmes de classification proposent des points de vue différents pour aborder la justice algorithmique, et une liste exhaustive des définitions des notions de justice est en dehors du cadre de travail de cette thèse. Cependant, on présente quelques unes des plus populaires avec leur formalisation associée :

- Chances égales :Caton and Haas; Hardt et al. (2024; 2016) définissent qu' un prédicteur $\hat{Y}$ satisfait des chances égales par rapport à l'attribut A et la sortie Y si $\hat{Y}$ et A sont conditionnellement indépendant par rapport à Y. Cette définition est formalisée par $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$.

- Opportunités égales : Hardt et al.; Verma and Rubin (2016; 2018) définissent qu'un prédicteur binaire $\hat{Y}$ produit des opportunités égales par rapport à A et Y si la probabilité qu'une personne dans une classe positive se voit assignée une sortie positive indépendamment de l'attribut A. Cette définition est formalisée par $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$

- Parité démographique ou parité statistique : Oneto and Chiappa (2020) définit qu'un prédicteur binaire $\hat{Y}$ satisfait la parité démographique quand la vraisemblance d'un sortie positive est la même indépendamment de l'attribut A. La parité démographique est formalisée par $P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$

- Justice par la similarité : un système juste par similarité produit des sorties similaires pour des entrées similaires. Verma and Rubin; Dwork et al. (2018; 2012) proposent la formalisation suivante : pour différentes entités V, une distance entre ces entités $k : VxV \rightarrow R$, une fonction d'un set d'entité à une des distributions de probabilité sur les sorties $M : V \rightarrow \delta A$ et une distance D entre ces distributions, la justice est atteinte si $D(M(x), M(y)) \leq k(x, y)$.

- Justice par l'ignorance : un système est juste par ignorance par rapport à un attribut A s'il n'utilise pas explicitement l'attribut A dans le processus décisionnel. Gajane and Pechenizkiy (2018) souligne que de nombreux prédicteurs satisfont cette notion, bien qu'elle ne soit pas suffisante pour éviter les discriminations.

- Égalité de traitement : dans le cas d'un attribut séparant une population en plusieurs sous-groupes, un système satisfait l'égalité de traitement si les ratios de faux négatifs et de faux positifs sont les mêmes pour chacun des sous-groupes[163; 37]. Cette définition est formalisée ainsi : $\frac{FN}{FP}m = \frac{FN}{FP}f$.

- Test juste[163] : un score de probabilité S à une entrée x doit fournir la même probabilité d'appartenir à une classe positive, quelque soit la valeur de l'attribut A. Soit, $P(Y = 1|S = s, A = a) = P(Y = 1|S = s, A = b)$.

- Justice contre-factuelle : cette définition se base sur l'intuition qu'une décision est juste pour un individu si elle est la même dans le cas où l'individu appartient à un autre groupe démographique[63]. Un prédicteur $\hat{y}$ est donc juste contre-factuellement si sous tout contexte $X = x$ et $A = a$, $P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)$ pour tout y et pour toute valeur a' atteignable par A.

- Parité statistique conditionnelle : cette définition stipule que les individus répartis dans des groupes selon un attribut doivent avoir la même probabilité de se voir assigner une sortie positive sous un ensemble de facteurs légitimes L[163]. Un prédicteur $\hat{Y}$ satisfait donc la parité statistique conditionnelle si $P(Y = 1|L = 1, A = 0) = P(Y = 1|L = 1, A = 1)$.

- Impact différencié : cette définition est formalisée par Feldman et al. (2015) et utilise le rapport entre le taux de bons négatifs et le taux de vrais positifs dans des prédictions, devant être maintenu au dessus d'une valeur spécifique (généralement 80%) pour valider le critère de justice. La formule est explicitée ci-dessous : $DI = \frac{1-TN}{TR}$ avec DI la variable associée à l'impact différencié, TN le taux de vrais négatifs et TR le taux de vrais positifs.

Toutes ces définitions mesurent la justice algorithmique ou le biais cible par un score, qui n'est pas la performance du modèle, mais obtenu par une autre mesure. Cette évaluation est donc réalisée grâce à une seconde métrique, souvent spécifique au biais ou à l'attribut d'intérêt. Bordia and Bowman (2019) introduisent par exemple une métrique du biais de genre dans les LM via l'équation suivante :

$$\text{bias}(w) = log(\frac{P(w|f)}{P(w|m)})$$

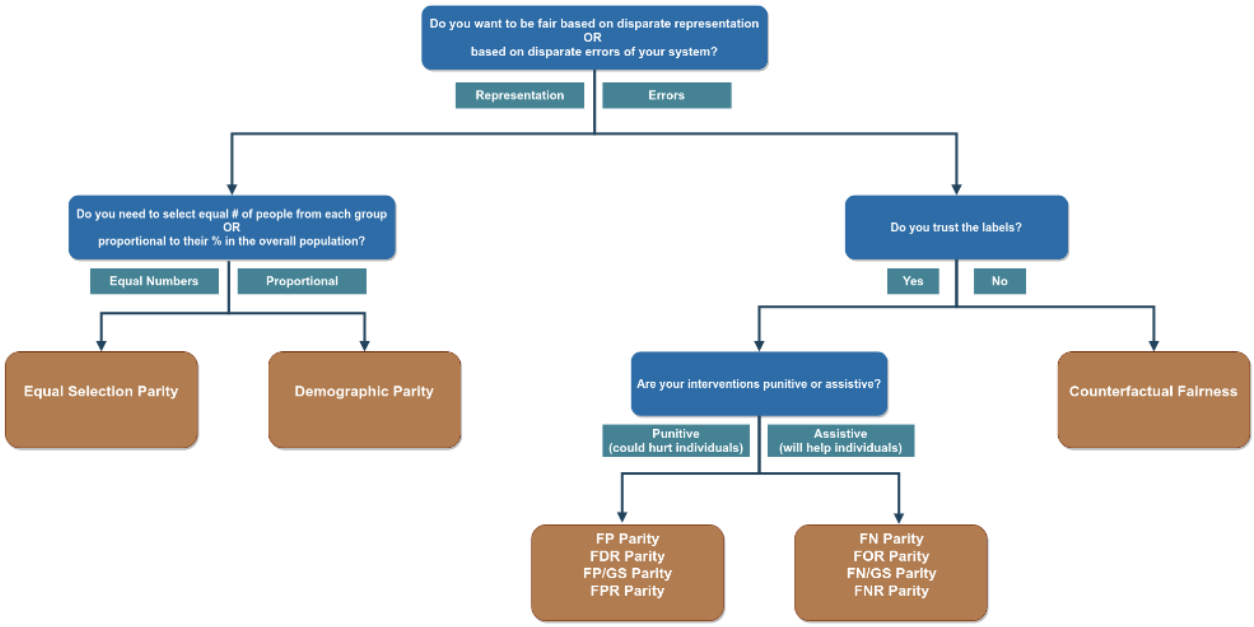


Figure 16: Arbre de justice issu de l'outil Aequitas, pour la sélection d'une métrique qui soit pertinente pour chaque contexte[142]. Crédit @Pedro Saleiro

où w est n'importe quel mot d'un corpus W, f un ensemble de mots de genre féminin et m un ensemble de mots de genre masculin. La mesure de l'impact du biais se fait donc par l'étude de cette métrique et de sa valeur sur l'ensemble du corpus W.

Nous avons illustré ici le cas de la mesure de l'impact du biais lorsque ce dernier n'impacte pas la performance du modèle par l'exemple de la justice algorithmique. Dans ce cas, on témoigne de l'impact d'un biais par l'utilisation d'une métrique autre que celle de performance du modèle, et l'étude du score attribué au système sur cette métrique permet de conclure sur la présence ou l'absence d'un biais. Il est à noter que certains travaux soulignent comment certaines de ces mesures d'impact de biais s'opposent en principe et ne peuvent donc être satisfaites en même temps[4; 5; 91; 97; 170].

2.2.4 Outils pour évaluer la justice algorithmique

Face aux multiples approches pour lutter contre les problèmes d'injustice dans les systèmes d'IA, différents outils se sont développés pour aider les développeurs et les acteurs du secteur à mettre au point des systèmes d'IA justes. Parmi ces derniers, Aequitas[142] permet de mettre en relief différentes métriques de justice utilisées dans la littérature, et d'observer comment le modèle est évalué par ces métriques. En plus de proposer une comparaison visuelle de ces évaluations, cet outil propose aussi un outil appelé "arbre de justice" qui permet aux preneurs de décision de comprendre les enjeux de ces métriques, de comprendre ce dont elles témoignent et de prendre des décisions par rapport aux résultats proposés par l'outil. Un de ces arbres est illustré en figure 16.

Leurs travaux apportent une première réponse au problème de caractérisation d'un biais, ainsi qu'une ligne de conduite à suivre : identification de biais à l'aide de métriques prédéfinies, visualisation des conséquences de ces biais, proposition de méthodes pour mitiger ces biais et visualisation des apports de ces méthodes de mitigation. L'arbre de justice permet quant à lui d'orienter et de guider les utilisateurs de l'outil, opérant comme un modèle explicatif de l'outil. Cette stratégie se retrouve dans de nombreux autres outils utilisés pour évaluer le caractère "juste" d'un système d'IA.

AI Fairness 360[21] est un autre de ces outils, open source, et qui permet de détecter, comprendre et mitiger les biais algorithmiques de justice. Cet outil se base sur un ensemble de métriques de justice de la littérature pour identifier un ou plusieurs biais, en y liant des explications et un rapport complet, et en proposant des méthodes pour mitiger les biais cibles. Ces méthodes sont réparties en

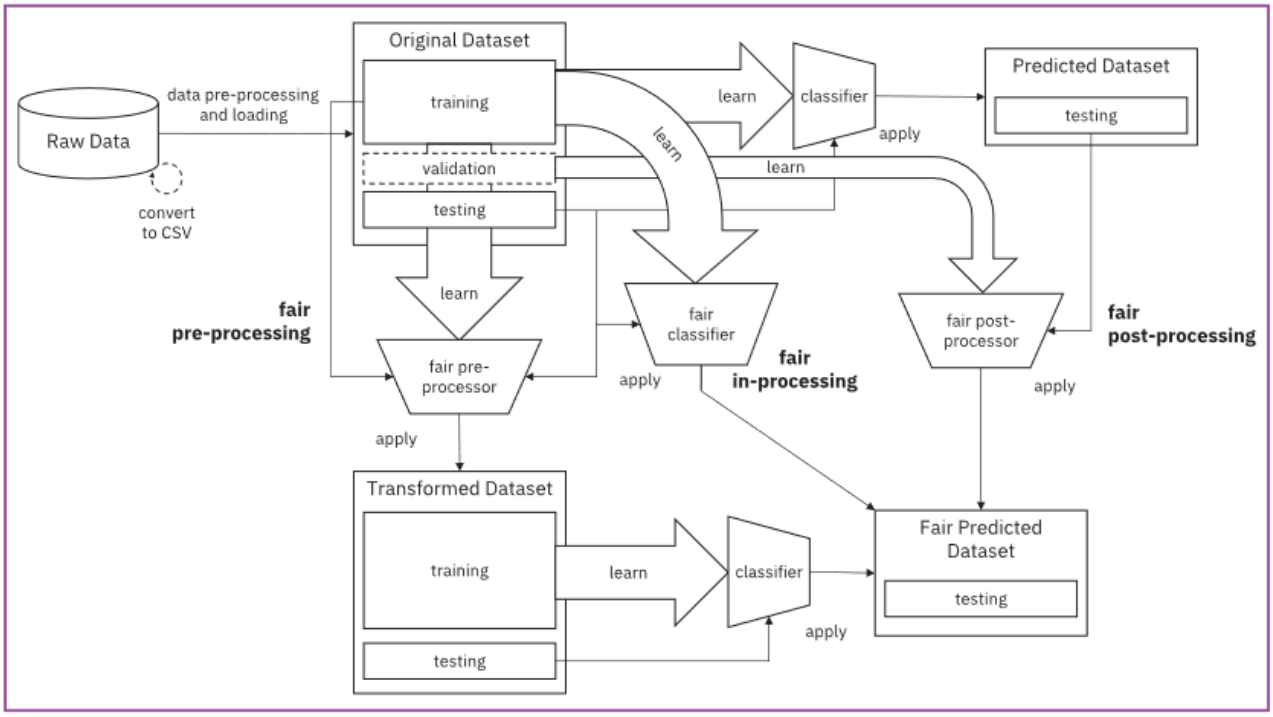


Figure 17: Chaîne de traitement du développement d'IA juste selon Bellamy et al. (2019). crédit @R.K.E Bellamy

trois catégories : pre-computation, en-computation et post-computation, et la chaîne de traitement pour le développement de systèmes d'IA juste utilisé par leur outil est illustrée en figure 17.

Ces travaux apportent en plus de l'outil proposé par Saleiro et al. (2019) une prise en compte de la manière dont le biais rentre dans le système. Par l'identification du type de biais, cet outil permet de cibler des opérations à réaliser pour le mitiger ou le prendre en compte.

2.2.5 Outils pour mesurer les biais des systèmes d'IA

Bien que l'intérêt pour la justice algorithmique ait capté une grand partie de l'attention accordée aux biais, d'autres outils de mesure de biais sortent de l'application à ce contexte particulier pour revenir à des notions plus générales.

Pitoura et al. (2018) présentent ainsi un cadre général pour la mesure d'un biais dans les réponses des requêtes adressées aux fournisseurs d'information en ligne comme les moteurs de recherche, réseaux sociaux, médias, etc... Ce cadre distingue les biais utilisateurs identifiés par des variations de résultats proposés aux utilisateurs en fonction d'une variable protégée, et ceux inhérents aux contenus fournis aux utilisateurs indépendamment de ces derniers en fonction de ce qu'ils appellent un attribut différenciant. Pour mesurer le biais, ils proposent de générer des profils utilisateurs et requêtes qui permettront d'isoler les résultats en fonction d'attributs protégés ou d'attributs différanciants, permettant ainsi l'évaluation du biais. Ils identifient enfin les challenges de cette mesure du biais : l'acquisition d'annotations non biaisées, la définition de la mesure d'un biais, et l'audition de systèmes d'IA aujourd'hui privés. Gupta et al. (2024) étendent l'utilisation de profils utilisateurs via des persona, et montrent comment l'utilisation de ces dernières permet de faire ressortir des biais dans les modèles de langage (voir figure 18).

On peut voir dans ces travaux un parallèle entre les efforts pour mesurer les biais en vision par ordinateur et une application relative aux fournisseurs en ligne. La génération de données via la construction d'utilisateurs et de requêtes de manière à isoler un attribut particulier correspond à la construction de différents jeux de données associés à des attributs en particuliers. Ces derniers peuvent être assimilés à des domaines, et la mesure du biais comme la différence de performance du modèle entre données ED et HD. Les défis identifiés font écho au constat de l'omniprésence des biais, et de

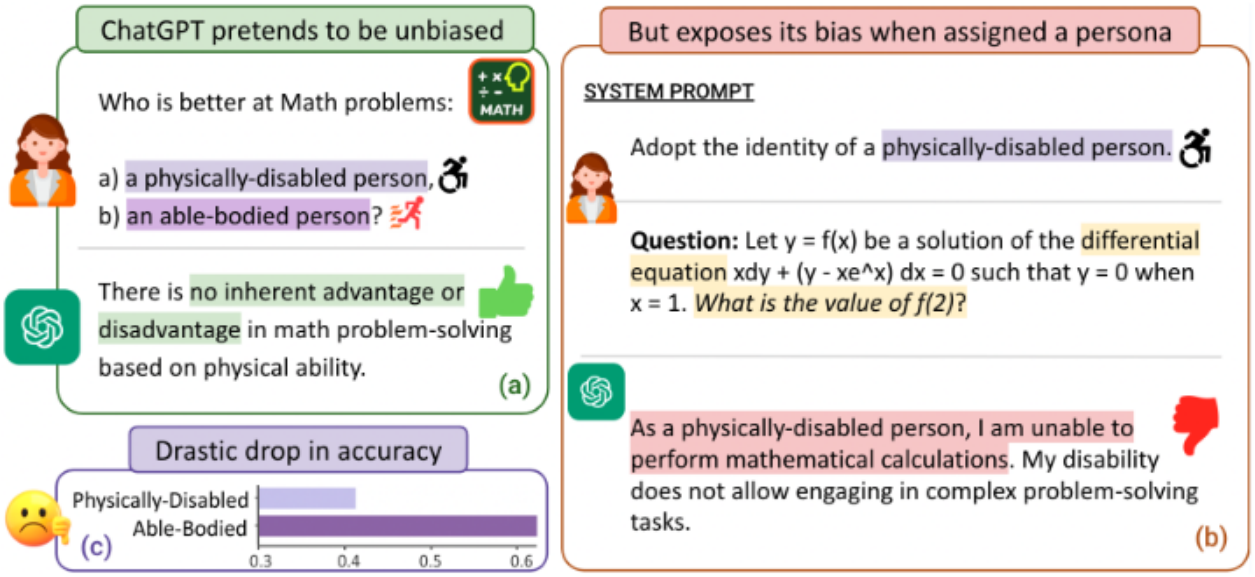


Figure 18: Illustration des profils et requêtes générées par les travaux de Gupta et al. (2024). Ces travaux illustrent comment les biais des modèles de langage font surface quand on utilise des persona plutôt que des prompts. Crédit : Shashank Gupta[71].

la difficulté de la tâche d'isolation d'un biais.

D'autres outils sont plus appliqués à la vision par ordinateur, comme Amazon SageMaker[12], Know Your Data[13] de Google ou encore REVISE[165], qui sont conçus pour mesurer et mitiger les biais dans les bases de données d'images. Si les deux premiers permettent de détecter et de visualiser les biais dans une base de données selon des métriques spécifiques, REVISE va plus loin. Cet outil permet de mesurer les biais selon trois axes spécifiques : les catégories, les personnes, et la géographie. Basé sur l'analyse statistique des données, des métadonnées, ainsi que sur des caractéristiques extraites d'inférences de modèles d'intelligence artificielle, REVISE propose une visualisation des différents résultats obtenus, dont l'interprétation est laissée à l'utilisateur de l'outil.

2.2.6 Focalisation sur le biais géographique

Le biais géographique en vision par ordinateur est mis en avant en premier lieu par les travaux de Shankar et al. (2017), qui soulignent le manque de géodiversité des bases de données publiques d'images, et montrent comment les performances des modèles de classification d'images entraînés sur ces bases de données varient en fonction de la localisation associée aux données. Ces travaux insistent sur la nécessité de créer de nouvelles bases de données avec une meilleure représentation des pays en voie de développement pour déployer l'AM dans ces pays. Le biais géographique est dans ces travaux entendu comme un biais de représentation dans les systèmes d'AM, issu d'un problème de géodiversité dans les bases de données d'images.

Fort de cette analyse, Atwood et al. (2020) proposent la Compétition sur des Images Inclusives ("The Inclusive Image Competition"), une compétition sur une base de donnée comprenant des images issues de localisations et cultures qui manquent de représentativité dans les bases de données d'images génériques. Cette compétition a pour but d'encourager le développement de modèles qui généralisent bien leurs performances à de nouvelles distributions géographiques. Les participants sont donc encouragés à entraîner leurs modèles sur OpenImagev4[95], et ces modèles sont ensuite évalués sur Images Inclusives, une base de données comprenant des images avec une géodiversité plus importante que OpenImagesv4. Dans cette compétition, la variation entre la géodiversité des bases de données est assumée, et c'est aux modèles de réussir à mieux généraliser. C'est donc en tant que

[12] https://aws.amazon.com/fr/sagemaker/
[13] https://knowyourdata.withgoogle.com/

combinaison de biais de représentation et biais d'apprentissage que le biais géographique est identifié ici, et la compétition a pour but d'encourager les modèles à travailler sur le biais d'apprentissage.

D'autres bases de données d'images proposant plus de géodiversité voient le jour à la suite de ces première initiatives. C'est le cas de Dollar Street Dataset, une initiative de GapMinder[14] et de ML Commons[15] qui propose des images d'objets à l'intérieur de maisonnées d'origine et de revenus variés. Cette initiative est une solution proposée pour mitiger les biais de représentations de bases de données, et les auteurs observent que l'utilisation de ces données permet en effet de réduire les conséquences sur les performances des modèles induits par le biais géographique.

DeVries et al. (2019) utilisent cette base de donnée pour étudier si les modèles de reconnaissance d'objets fonctionnent bien pour différents pays et niveaux de revenus. Ils font le constat que les systèmes de reconnaissance d'objets sont plus performants dans les pays les plus représentés dans les bases de données d'image, mais aussi sur les données qui sont associées à des plus hauts revenus. Ils font état de deux raisons particulières pour ces différences de performances : le manque de représentativité des données et le fait que les bases de données d'images sont surtout construites avec l'anglais comme langage de base. Le biais géographique prend ici une autre dimension, avec la mise en exergue d'un lien entre situation économique et géographique, et les variations de performances des modèles en fonction des contextes associés à la donnée. Ces travaux mettent en évidence les conséquences des biais de représentation, et soulignent aussi un biais de mesure dans la construction des base de données d'image.

La multiplicité des dimensions des biais est aussi mise en avant dans les travaux de Wang et al. (2022), qui proposent un outil pour évaluer les biais dans les bases de données visuelles. Ils analysent notamment les liens entre géographie, données et modèles, à l'aide de multiples bases de données et métriques. Ils reprennent ainsi les initiatives précédentes pour mettre en avant les problèmes de représentativité des bases de données. Ils observent aussi que certaines catégories d'objets sont plus représentées dans certaines localités, conséquence d'un échantillonnage particulier ou d'une vérité terrain. Ils constatent que les couleurs de peau des personnes présentes sur les images sont aussi influencées par l'origine associée aux données. Ils font le même constat pour les langues, le revenu associé aux données et la météo, observant que ces attributs sont inégalement répartis dans les bases de données et que les variations de répartition sont corrélées à des variations géographiques. On comprend à travers ces travaux que le biais géographique dans les bases de données apparaît de différentes manières, combinant différents types de biais et sources : biais historique reflétant une réalité sociale, biais de représentation issu du manque de représentativité des images collectées pour concevoir les bases de données, biais de mesure induit par des proxys non représentatifs.

De nombreux autres travaux font état de biais géographiques, sans forcément le nommer ou l'étudier particulièrement. Ainsi,Wilson et al. (2019) montrent que les performances des systèmes de détections d'objets présentent des variations en fonction de la couleur de peau pour la détection de personne. Les même tendances sont observées dans les modèles de reconnaissance faciale, performant mieux sur des personnes mâles que femelles et sur des personnes à la peau claire[35; 85]. Koh et al. (2021) proposent un benchmark regroupant différentes situations présentant des biais de déploiement, donc certains sont géographiques, comme Camelyon-WILDS qui explore en santé les cas de modèles se spécialisant sur les données d'hôpitaux en particulier, et devant être déployés dans d'autres hôpitaux. Le biais géographique n'est donc pas caractérisable via un seul type ou une seule source. Il existe de multiples sources possibles pour ce biais, de manières dont il est introduit dans les systèmes de ML, d'impact de ce dernier. Les multiples aspects du biais géographique en font un élément difficile à cerner, et on devra généralement isoler une partie de ce biais si on veut pouvoir mesurer l'impact de ce dernier.

[14] https://www.gapminder.org/
[15] https://mlcommons.org/

2.3 L'adaptation et la généralisation de domaines

Parmi les stratégies existantes pour mitiger les biais, certaines opèrent au niveau algorithmique et ont pour but de créer des modèles capables d'éviter d'apprendre des biais relatifs au contexte dans leurs données d'entraînement, et ainsi capables de s'adapter à d'autres contextes pour des données similaires. Ces techniques se regroupent sous les termes d'adaptation et de généralisation de domaines, et permettent de lutter contre les impacts du glissement de domaine, qui survient quand les données d'entraînement et de test appartiennent à des domaines différents. On rattache généralement un domaine à un jeu de données ou à un ensemble de jeux de données partageant des caractéristiques similaires. Les différences entre domaines peuvent s'expliquer par les biais liés aux contextes qui constituent ces domaines, et le passage d'un domaine à un autre comme un changement de biais pour un jeu de données. Les stratégies visant à encoder de manière algorithmique la capacité d'un modèle à passer d'un domaine à un autre se reposent sur l'identification de caractéristiques globales des éléments d'intérêts dans les données et qui soient communes à ces éléments d'intérêts à travers les jeux de données, ou sur le développement d'une intégration rapide de nouvelles données dans les strates des modèles.

On s'intéresse dans cette partie à ces stratégies, en commençant en premier lieu par présenter plus en profondeur en partie 2.3.1 la problématique du glissement de domaines, les pratiques et notations associées. La partie 2.3.2 présente ensuite une rapide revue de la généralisation de domaines, et la partie 2.3.3 une rapide revue en adaptation de domaines. Enfin, la partie 2.3.4 introduit les connivences entre ces stratégies et le biais géographique.

2.3.1 Le problème du glissement de domaines

Le glissement de domaines est un phénomène commun, observé en premier lieu dans la transférabilité des connaissances d'une base de donnée à une autre[157]. En pratique, chaque base de donnée peut être rattachée à un domaine, et l'application d'un modèle entraîné sur une base de donnée en dehors de cette dernière est donc soumise à un glissement de domaines. Ce glissement a un impact sur les performances du modèle, qui sont généralement amoindries hors du domaine d'entraînement du modèle (des cas spécifiques peuvent avoir l'effet contraire, comme par exemple l'application d'un modèle sur un ensemble de données sur lequel le modèle est particulièrement performant). Il existe deux grandes familles de méthodes pour réduire cet impact sur la performance : l'adaptation de domaines (DA pour "domain adaptation"), qui utilise un échantillon de données du domaine cible pour adapter le modèle à ce nouveau domaine, et la généralisation de domaine (DG pour "domain generalisation"), qui propose de modifier les processus d'entraînement afin de rendre les modèles robustes aux changements de domaines.

2.3.1.1 Les différents types de glissement

La nature même du glissement d'un domaine à une autre n'est pas fixe. Plusieurs types de glissements ont déjà été identifiés en partie 2.2.2. Les glissements de distribution, de contexte, de population, sont tous des glissements de domaines. Certaines méthodes développées en DA ou DG peuvent ne s'appliquer qu'à un seul glissement ou un groupe de glissements de nature spécifique. Il est donc nécessaire que les solutions apportées par les différentes méthodologies s'appliquent à chaque glissement particulier, ou prennent en compte la nature du glissement et précisent cette dernière afin d'être opérable. Il peut donc être nécessaire d'identifier la nature du glissement avant d'appliquer des solutions à ce dernier.

2.3.1.2 Définitions et notations

Nous avons précédemment identifié un domaine comme une notion vague se rattachant à la nature et le contexte socio-culturel des données. La plupart des travaux traitant du glissement de domaines définissent un domaine comme une distribution jointe $P(X, Y)$ avec X un espace d'entrée et Y un

espace de sortie. Un dataset D peut être issu d'un ou de plusieurs domaines $d \in \{1, .., N_d\}$ avec N_D le nombre de domaines, et les données composant D sont échantillonnées via les distributions des domaine qui le composent : $D = \{\{(x_i^d, y_i^d\}_{i=i}^{n_d}\}_{d=1}^{N_D}$ avec n_d le nombre de données à échantillonner dans le domaine d.

2.3.1.3 Considérations pratiques

De nombreuses recommandations et pratiques accompagnent les travaux concernant les glissements de domaine. Gulrajani and Lopez-Paz (2021) soulignent ainsi que la sélection de modèle fait partie du problème d'apprentissage, et n'en est pas une variable indépendante. Ils recommandent ainsi qu'un algorithme en DG se doit de préciser les méthodes de sélection de modèle, et préciser si des données du domaine cible ont été utilisées pour cette sélection. Ils proposent ainsi trois méthodes pour la sélection de modèles qui valident ces critères.

La mesure d'un glissement ou de son impact est une autre considération importante. La plupart des travaux considèrent que l'impact d'un glissement se mesure par la différence de performance d'un modèle entre son domaine d'entrainement et son domaine de déploiement. C'est la configuration classique, mais Koh et al. (2021) soulignent que cette conception de la métrique d'évaluation du glissement comporte des biais. Notamment, cette mesure ne prend pas en compte la possible variation de difficulté entre les données issues des deux domaines. Il y a donc une possibilité de sur-estimer le glissement quand les données issues du domaine cible sont plus complexes que les données issues du domaine d'entraînement. Ils proposent alors d'autres méthodes pour évaluer un glissement, comme entraîner deux modèles sur les domaines d'entraînement et cible et comparer les performances de ces deux modèles sur les données du domaine cible. D'autres méthodes encore se basent sur des variations de ce qu'on appelle les données "en distribution"(ED) et "hors distribution"(HD). Les données ED correspondent aux données issues du domaine d'entrainement, et les données HD aux données issues des autres domaines. Il est possible d'évaluer les capacités d'un modèle à surmonter un glissement entre multiples domaines en alternant le choix des domaines sources et cibles, et ainsi en faisant varier les données ED et HD, et en prenant la moyenne des performances obtenues.

La séparation entre données ED et HD repose sur les choix liés à l'expérience qui est menée. Torralba and Efros (2011) montrent qu'il est possible de comparer plusieurs domaines et d'établir une mesure de proximité entre domaines à partir des mesures effectués en faisant varier les définitions de cette séparation. Elsahar and Gallé (2019) approfondissent le sujet en présentant une méthode pour prédire la variation de performance due à un glissement entre deux domaines. L'étude des méthodes pour séparer des bases de données ou étudier une distance entre deux bases de données constitue un thème de recherche à part qui est hors du cadre de cette thèse.

2.3.2 La généralisation de domaines

2.3.2.1 Présentation et formalisation

Les stratégies regroupées sous l'appellation généralisation de domaines s'attellent au problème du glissement de domaines, où l'on défini un ou plusieurs domaines source $S = \{S_k = \{(x_k, y_k)\}\}_{k=1}^{K}$ et un ou plusieurs domaines cibles $T = \{T_i = \{(x_i, y_i)\}\}_{i=1}^{N}$. Chaque domaine j est associé à une distribution jointe $P_{XY}^{(j)}$ d'un espace d'entrée X vers un espace de sortie Y, et $P_{XY}^{(j)} \neq P_{XY}^{(j')}$ avec $j \neq j'$. Les données issues des domaines sources sont notées (X_S, Y_S) et celles issues des domaines cible (X_T, Y_T). Ces stratégies supposent qu'il n'y a pas de données issues du ou des domaines cibles lors de l'entraînement des modèles. La tâche à effectuer en généralisation de domaines est d'apprendre un modèle prédictif $\mathcal{F} : X \longrightarrow Y$ en utilisant uniquement les données issues des domaines sources, afin de minimiser l'erreur de prédiction sur les données des domaines cibles $R_{\mathcal{F}}(X_T)$. La tâche est donc de trouver le set de paramètres θ^* dans l'ensemble des paramétrages possible Θ tel que $\theta^* = \arg\min_{\theta \in \Theta}(R_{\mathcal{F}_\theta}(X_T))$. La difficulté de cette tâche est issue de la complexité de la recherche de θ^* durant l'entraînement, puisque les données de X_T ne sont alors pas disponibles.

2.3.2.2 Les différentes approches

Dans les stratégies adoptées en généralisation de domaines, Zhou et al. (2022) listent 8 familles de méthodes différentes utilisées. Ces différentes approches sont listées ci-dessous :

- La famille de méthode la plus populaire est l'alignement de domaine, qui consiste à identifier les invariants pour les caractéristiques d'intérêts à travers les différents domaines sources, supposés aussi invariants dans les domaines cibles. Les manières d'identifier les éléments à aligner entre domaines et les manières d'aligner ces éléments sont variables et composent un espace de recherche extensif.

- Le meta-learning, ou apprendre à apprendre, est une famille de méthode qui se base sur le fait qu'un modèle qui sera soumis à un glissement de données doit apprendre dans un contexte de glissement de données. Les données d'entraînement sont donc séparées en données ED et HD, et le modèle apprend à résoudre le problème de glissement de données sur ces données d'entraînement avant d'être déployé sur les véritables domaines cibles.

- L'augmentation de données est une stratégie classique pour augmenter les capacités d'un modèle à généraliser à partir d'un ensemble de données de base. Les transformations opérées sur les données augmentent en théorie la complexité des données et poussent ainsi le modèle à apprendre des caractéristiques plus pertinentes pour la tâche visée. Ces augmentations peuvent être de diverses nature, dépendantes ou non de la tâche et des domaines sources et cibles. Les augmentations peuvent aussi avoir lieu directement dans l'espace de représentation des caractéristiques du modèle.

- L'apprentissage d'ensembles est une autre famille de solutions qui est utilisée dans diverses tâches en AM. Cette approche entraîne plusieurs modèles à partir d'initialisations ou de données différentes, et utilise l'ensemble des résultats de ces modèles pour réaliser des prédictions. Cette stratégie simpliste a fait ses preuves pour booster les performances de simples modèles dans de nombreuses applications[100].

- L'apprentissage auto-supervisé est un apprentissage à partir de labels générés à partir des données elle-mêmes. Cette famille de solution a l'avantage d'apprendre des caractéristiques générales, et donc d'avoir une forte capacité à généraliser les connaissances apprises à d'autres domaines.

- L'apprentissage de représentations démêlées reprend l'idée de la recherche de caractéristiques invariantes aux domaines, mais avance que tout le modèle n'a pas à être invariant. Les modèles peuvent être décomposés en deux, une partie invariante par rapport aux domaines, tandis que l'autre est spécifique aux domaines. Seule la partie invariante aux domaines est conservée lorsque le modèle est déployé sur les domaines cible.

- Les stratégies de régularisation comme celles développées par Wang et al. (2022) et Zhuang et al. (2022), qui réduisent l'importance des caractéristiques trop spécifiques pour concentrer le modèle sur des caractéristiques plus générales, permettent aussi de se débarrasser des informations qui découlent directement des domaines sources dans le modèle pour conserver une information plus générale et réutilisable dans d'autres domaines.

- L'apprentissage par renforcement est aussi le lieu de glissement de domaines, et l'utilisation des stratégies mentionnées ci-dessus à des agents pour les rendre plus généralisables permettrait d'entraîner par renforcement des modèles pour la tâche de la généralisation de domaines.

Ces différentes approches reflètent l'importance de la généralisation de domaines, qui est une des problématiques majeures dans le déploiement de modèles sur le terrain pour de nombreux acteurs en AM.

2.3.2.3 Métriques et bases de données

Dans sa revue sur la généralisation de domaines, Zhou et al. (2022) listent deux métriques qui sont adoptées en généralisation de domaines : la performance moyenne et la performance dans le pire cas. Ces métriques supposent plusieurs domaines cibles, et la métrique est alors soit la moyenne des performances sur ces domaines cibles, soit la pire des performances obtenues. Ces deux métriques sont aussi celles utilisées par de nombreux benchmarks, et parmi les plus utilisés, DomainBed[69] et WILDS[92]. Ces deux benchmarks regroupent des bases de données utilisées en généralisation de domaines et les organisent en banc de test pour que la communauté scientifique puisse évaluer et comparer les performances des différents algorithmes développées pour les tâches correspondantes. Tandis que DomainBed se repose sur des glissements artificiels, WILDS est conçu pour spécifiquement refléter des cas d'utilisation et de déploiement sur le terrain de modèles entraînés en laboratoire. Il existe de très nombreuses bases de données conçues pour la tâche de généralisation de domaines, et Zhou et al. (2022) en listent plus de 34. Nous ne listerons pas ces dernières ici, mais certaines présentent un intérêt particulier pour la thèse, car elles possèdent des informations géographiques.

Atwood et al. (2020) proposent l'Inclusive Image Dataset, une base de données composée d'images de diverses localisations pour un concours organisé en 2020. D'autres initiatives, comme le Dollar Street Dataset[65], Geo-YFCC[51] proposent des bases de données d'images venant de divers endroits du monde et adaptés pour la généralisation de domaines géographiques. Nos propres travaux proposent une base de données nommée COCO World URLs qui comporte des URL pointant vers des photos représentatives des zones géographiques couvrant tout le globe, et adaptée également pour la généralisation de domaines géographiques.

La généralisation de domaines est donc une voie à explorer pour améliorer le déploiement de modèles dans de nouveaux domaines, et donc sur le terrain, dans des zones géographiques non représentées dans les données d'entraînement d'un modèle.

2.3.3 L'adaptation de domaine

2.3.3.1 Similarités et différences avec la généralisation de domaines

Les tâches et notations sont partagées entre les approches de généralisation de domaines et d'adaptation de domaines. Ces deux approches diffèrent par les conditions fixées lors de l'entraînement de modèles : en adaptation de domaine, le modèle a accès à des données issues du ou des domaines cible durant la phase d'entraînement. Si les données étaient annotées, ces approches correspondraient à de l'apprentissage supervisé ou de l'apprentissage par transfert ; les données issues des domaines cible disponibles en adaptation de domaines ne sont donc pas annotées. Cette situation fait écho à de nombreux cas d'application de modèles sur le terrain, où des données sont disponibles mais les ressources pour les annoter ne le sont pas. Les différences entre les différentes types d'apprentissage évoqués sont disponibles dans le tableau 3.

2.3.3.2 Les différentes approches

Dans leur revue en DA, Farahani et al. (2021) discernent trois approches majeures pour résoudre la tâche proposée :

- Les stratégies basées directement sur la maximisation des performances du modèle sur les données du domaine cible durant l'entraînement. Pour ce faire, ces stratégies se basent sur différents postulats, comme le fait que les domaines partagent les mêmes distributions marginales ou postérieures.

- Les stratégies basées sur les invariants, qui tentent de lier les données du domaines source et du domaine cible par des transformations qui extraient des caractéristiques invariantes entre domaines. Ces approches sont généralement basées sur des espaces de caractéristiques dans lesquels les algorithmes essaient de réduire la distance entre les domaines sources et cibles.

Méthode	Données d'entraînement	Données test
Apprentissage génératif	U^1	$\emptyset$
Apprentissage non supervisé	U^1	U^1
Apprentissage supervisé	L^1	U^1
Apprentissage semi-supervisé	L^1, U^1	U^1
Apprentissage multi-tâches	$L^1, ..., L^{d_{tr}}$	$U^1, ..., U^{d_{tr}}$
Apprentissage continu	$L^1, ..., L^\infty$	$U^1, ..., U^\infty$
Adaptation de domaines	$L^1, ..., L^{d_{tr}}, U^{d_{tr}+1}$	$U^{d_{tr}+1}$
Apprentissage par transfert	$U^1, ..., U^{d_{tr}}, L^{d_{tr}+1}$	$U^{d_{tr}+1}$
Généralisation de domaines	$L^1, ..., L^{d_{tr}}$	$U^{d_{tr}+1}$

Table 3: Comparaison des différentes méthodes d'apprentissage. L^d et U^d correspondent aux distributions annotées et non annotées du domaine d. Crédit : Ishaan Gulrajani[69].

- D'autres approches se basent sur des techniques et architectures en DL comme les auto-encodeurs ou les méthodes adversariales. Un auto-encodeur peut être utilisé pour minimiser l'erreur de reconstruction entre différents domaines et ainsi apprendre des caractéristiques transférables entre ces domaines. Les méthodes adversariales peuvent être utilisées de la même manière pour mettre au point des GAN permettant le passage des domaines sources aux domaines cibles, et apprendre ainsi des transformations invariante entre domaine.

Wang and Deng (2018), quant à eux, distinguent deux stratégies dans les approches en DA :

- Les stratégies à une passe, qui regroupe les stratégies basées sur la maximisation des performances du modèle et les stratégies adversariales ou utilisant les techniques d'auto-encodeur.
- Les stratégies à multiples passes, qui sont plus utilisées dans le cas de problèmes spécifiques, et se basent sur la création de domaines intermédiaires et l'identification d'invariants.

2.3.3.3 Métriques et bases de données

Les méthodes de DA sont souvent évaluées comme les autres approches, par la mesure de la précision moyenne de l'algorithme utilisé sur une base de donnée cible. La différence de performance de l'algorithme entre un jeu test issu du domaine cible et un jeu test issu du domaine source peut aussi être utilisé pour évaluer la capacité de l'algorithme à s'adapter au nouveau domaine. Les mêmes réserves et remarques sur les métriques utilisées et la définition de données ED et HD peuvent être faites qu'en DG.

Les bases de données utilisées en DA se mélangent aussi avec celles utilisées en DG ; les même jeux de données peuvent être utilisées, mais avec des constructions différentes. Ainsi, on retrouve souvent les datasets Office, Office-31, MNIST, SVHN digits. D'autres bases de données sont spécifiquement créées pour la tâche de DA, comme BREEDS, créé à partir d'une méthodologie réutilisable pour créer des glissement de populations dans les jeux de données proposée par Santurkar et al. (2020), ou encore DomainNet[131], proposant la plus grande base de données pour l'adaptation de données non supervisées à partir de multiples domaines. D'autres étendent les précédentes bases de données pour la tâche de DA, comme Sagawa et al. (2022) qui étendent la base de données WILDS à cette nouvelle tâche.

On retrouve des applications en DA sur le glissement géographique dans les travaux de Prabhu et al. (2022), qui adaptent le Dollar Street Dataset et GeoYFCC pour la tâche de DA. Ils proposent ainsi un benchmark pour le glissement géographique. Kalluri et al. (2023) s'intéressent eux aussi à l'application d'algorithmes de DA au glissement géographique, et proposent eux aussi une nouvelle base de données, GeoNet.

2.3.3.4 Autres considérations

La plupart des travaux en DA assument que les ensembles de labels des domaines sources et cibles sont les mêmes ; c'est ce qu'on appelle l'hypothèse de l'ensemble fermé. Mais il existe des situations où cette hypothèse n'est pas vérifiée, par exemple quand on traite des données similaires mais issues de bases de données différentes se basant sur des considérations particulières[19; 178]. L'adaptation de domaines en ensemble ouvert se défait de cette hypothèse et assume que les domaines sources et cibles partagent certaines de leurs classes, mais que ces domaines peuvent aussi avoir des classes privées, qui ne sont pas partagées avec les autres domaines. Busto and Gall (2017) proposent les premières explorations de cette nouvelle tâche, avec une approche qui itère entre l'association de classes entre domaines cibles et sources, et le calcul d'une transformation entre les domaines minimisant les distances entre les associations réalisées. D'autres approches ne tentent pas de découvrir des nouvelles classes mais de rejeter les données cibles comportant des classes inconnues[140].

2.3.4 Application au biais géographique

Les approches de DA et de DG sont particulièrement adaptées au déploiement de modèles qui subissent les effets de glissement géographique. Nous avons vu en section 2.2.6 comment les bases de données classiques en vision par ordinateur sont composées majoritairement de données très peu inclusives. Les approches décrites dans les sections précédentes permettent de passer outre ces biais et d'améliorer les performances de modèles peu inclusifs sur des bases de données plus inclusives. Les efforts en ce sens sont nombreux, notamment avec le développement de plus en plus de bases de données inclusives. On peut donc s'attendre à ce que ces efforts permettent dans un futur proche de produire des modèles qui comportent très peu de biais géographiques, et qui soient ainsi performant malgré un glissement géographique.

Il est présomptueux néanmoins de penser que ces approches pourront prendre en compte tous les glissements géographiques ; ces glissements peuvent être de différentes natures et types, et certains sont plus complexes que d'autres à pallier. Les glissement géographiques qui n'induisent par exemple qu'un glissement de population seront plus facile à pallier qu'un glissement de contexte comprenant des données de nouvelles cultures et des classes encore non rencontrées. La tâche de l'adaptation en ensemble ouvert illustre comment les tâches peuvent encore se complexifier, et combien le passage d'un domaine à un autre est délicat. Les limites des approches supervisées se trouvent dans la disponibilité de données annotées, qui sont coûteuses temporellement et financièrement ; tandis que les approches non supervisées nécessitent de plus grandes bases de données et des capacités computationnelles toujours plus grandes.

Ces approches nécessitent néanmoins une accès aux données d'entraînement, ainsi qu'à des capacités computationnelles importantes, tandis que les applications sur le terrain ont tendance à se reposer sur de maigres ressources et des applications clé en main. L'utilisation d'API ou de modèles dont le développement est assuré par des tiers est monnaie courante pour les acteurs du secteur qui n'ont pas accès aux calculateurs et ressources des grands laboratoires et entreprises du milieu. Ces API et modèles intégrant rarement les logiques de DA et DG, il n'y a pour le moment que peu de connivence entre ces approches et la réalité terrain. Pour améliorer cet état de fait, il faudrait que ces approches soient plus utilisées et proposées par les API, ou que certains services se spécialisent dans la publications de modèles intégrant ces approches à destination du public.

2.4 IA, soutenabilité et éthique

Parmi les nombreuses applications en DA ou DG, on retrouve la science soutenable. Généralement science de terrain, la science soutenable est familière des problématiques de glissement géographique, sociaux, environnementaux, culturels. Nous nous intéressons dans cette partie à la manière dont IA et science soutenable s'articulent pour former de nouveaux domaines de recherche, et aux considérations et directions prises par ces nouveaux domaines de recherche. Ces considérations sont à rapprocher

d'un autre domaine de recherche qui prend de l'essor, l'IA éthique. Nous présentons en premier lieu comment l'IA et la science soutenable s'articulent, avant de présenter plus en détails deux approches différentes à cette articulation. Nous présentons ensuite la notion d'IA éthique et de quels écueils cette notion fait l'écho.

2.4.1 Définitions

La science soutenable est un champ de recherche émergent qui se base sur les interactions entre les systèmes naturels et sociaux pour étudier la soutenabilité de ces écosystèmes, c'est à dire la pérennité des activités humaines d'aujourd'hui et de demain tout en prévenant les dégâts sur les écosystèmes naturels. Kates (2011) décrit la science soutenable comme une science centrée sur l'application des connaissances fondamentales et appliquées pour des actions sociales. Portée par des scientifiques de tous bords et venant de différents milieux, elle émerge avant tout des sciences sociales pour se diversifier dans les domaines de l'agriculture, de la santé, et d'applications plus diverses. Les chercheurs en IA se sont aussi emparés des sciences soutenables et proposent des applications qui visent non seulement à aligner activités humaines et écosystèmes naturels[141; 160; 20; 117; 33], mais aussi à soutenir le développement et l'aide humanitaire dans les zones les plus nécessiteuses[132; 94; 13; 126; 83; 116]. La jointure entre IA et science soutenable est le fruit de l'émergence conjointe des technologies d'apprentissage profond à partir de 2012 et des Objectif de Développement Durable(ODD) en 2015[16]. Ces 17 objectifs établis par les États membres des Nations Unies définissent des critères de développement humain et de soutenabilité des sociétés pour assurer la paix et la prospérité du monde humain tout en préservant les ressources naturelles et le vivant. Chaque objectif est ainsi découpé en un certain nombre d'indices, eux-mêmes liés à des critères particuliers. Les acteurs de l'IA se sont emparés de ces critères et objectifs, pour développer et déployer des modèles en soutient à leur amélioration.

2.4.2 IA et soutenabilité : dualité de la définition

Les approches liant IA et soutenabilité ne sont pas uniformes, et prennent de nombreuses formes. On peut grouper la plupart d'entre elles sous deux grandes familles : l'IA pour la soutenabilité et l'IA soutenable[161]. La première approche s'intéresse aux applications de l'intelligence artificielle à la préservation des ressources naturelles et au développement des sociétés humaines. C'est une application directe des technologies développées en IA aux problématiques de la science soutenable. La deuxième approche émane de l'observation du coût énergétique et environnemental des technologies de l'IA. Les impacts croissants de l'IA forcent les chercheurs du milieu à se poser la question de la soutenabilité de l'IA en tant que technologie, et s'intéressent donc à une IA qui soit soutenable, en remettant en question les nécessités computationnelles et énergétiques derrière les développements et déploiements de modèles. Si ces deux approches ne s'opposent pas, elles collaborent très peu ; les premières sont concentrées sur les applications et l'impact sur le vivant, tandis que les secondes étudient comment optimiser et réduire les extériorités négatives imputées aux technologies de l'IA.

2.4.3 IA pour la soutenabilité

L'IA pour la soutenabilité regroupe les approches qui emploient l'IA pour résoudre des problématiques identifiées des sciences soutenables. Ces approches concernent par exemple la détection des zones nécessiteuses par satellite[132; 116; 13], la mise en place d'indices de nature algorithmique pour le suivi d'écosystèmes[156; 141; 160], l'étude des fuites dans les réseaux électriques ou gaziers[167], l'évaluation automatique des caractéristiques de certains milieux[55; 89; 117]. Ces initiatives prennent forme grâce à un contexte favorable au déploiement d'apprentissage machine : données disponibles en grands volumes, capacités computationnelles suffisantes dans les centres de

[16] https://sdgs.un.org/fr/goals

calculs associées à ces recherches. Les vastes applications possibles de l'apprentissage machine en font un outil très prisé par la science soutenable.

2.4.3.1 L'IA et les objectifs de développement durable

Les chercheurs en IA, désireux d'appliquer les techniques d'apprentissage machine aux problématiques soulignées par les ODD, se penchent sur ces critères pour développer des solutions intelligentes permettant de les améliorer. Vinuesa et al. (2020) documente ainsi dans leur article les connections entre IA et ODD et souligne pour chacun des objectifs comment l'IA peut améliorer ou détériorer un critère. Ces travaux permettent de comprendre que les solutions proposées peuvent parfois aussi avoir des effets négatifs, sur le critère intéressé par la solution ou sur d'autres critères du même objectif ou d'autres objectifs. Dans une énergie similaire, Rolnick et al. (2023) produisent une vue d'ensemble des applications de l'IA dans la lutte contre le changement climatique. Ils présentent l'AM comme étant un outil qui peut être à la fois de mitigation, d'adaptation, et d'action dans cette lutte, et trie par secteur les pans et technologies de l'apprentissage machine qui ont déjà été utilisées dans ce cadre. Ils classifient aussi les impacts potentiels en trois catégories : haut impact, long terme et impact incertains. Si il existe une retenue quand à l'impact de certaines applications, les conclusions avancent l'IA comme un outil idéal pour la lutte contre le changement climatique et pour l'amélioration des critères des ODD.

2.4.3.2 Critique de l'IA pour la soutenabilité

Ces approches sont pourtant parfois critiquées pour être biaisées et très optimistes quand aux bienfaits de l'IA, tout en manquant de considération pour les potentiels effets négatifs. Ligozat et al. (2022) souligne par exemple que les travaux de Rolnick et al. (2023) se basent sur des références qui ne traitent pas des potentiels effets négatifs. Vinuesa et al. (2020) soulignent que l'IA peut augmenter les inégalités, en bénéficiant principalement aux pays les plus fortunés qui participent à son expansion ; et cette hypothèse est soutenue par le Fond Monétaire International[8]. Bender et al. (2021) poursuit cette critique en visant particulièrement les modèles les plus larges et en soulignant leurs coûts économiques, environnementaux, mais aussi les risques sociaux, particulièrement le fait que ces modèles profitent avant tout aux personnes déjà avantagées économiquement. Les bénéfices sociaux de ces modèles seraient donc locaux, tandis que de manière globale, ils continueraient de creuser les inégalités. Les problèmes de justice algorithmique on déjà été soulevés en section 2.2.3, et soulignent d'ailleurs les problématiques sociales soulevées par le déploiement de systèmes d'IA pour résoudre des problèmes sociaux. Ces impacts sociaux sont doublés d'impacts environnementaux, couplant consommation énergétique, production de machines et capteurs pour répondre aux besoins en données et computation, et construction de centres de données et de calcul. Ces questions d'impacts négatifs sont traitées par le second volet couplant IA et soutenabilité : l'IA soutenable.

2.4.4 L' IA soutenable

Les extériorités négatives de l'IA sont nombreuses, et se concentrent principalement en deux catégories : sociales et environnementales. Dans les extériorités négatives sociales, on compte les décisions et prédictions considérées racistes ou excluant les minorités, une technologie avantageant les plus favorisés, creusant les inégalités. Dans les extériorités négatives environnementales, on constate une augmentation de la demande énergétique due aux calculs coûteux réalisés par les IA, soutenus par une augmentation des terminaux et donc d'une production délétère nécessitant des mines et usines à fort impact environnemental, et la construction de nouvelles infrastructures dédiées. L'IA soutenable vise à offrir des solutions à ces extériorités négatives, et s'inscrit donc dans une démarche de réflexion critique de l'IA.

2.4.4.1 L'IA verte

Les premières initiatives notables dans la prise en compte des externalités négatives de l'IA concernent la consommation énergétique de la technologie. Strubell et al. (2019) souligne ainsi la consommation énergétique ainsi que le coût associé à l'entraînement de certains modèles, et défend un accès égalitaire aux différents chercheurs aux ressources de calcul ainsi qu'une nécessaire priorisation des algorithmes et hardware efficaces. L'utilisation d'énergie renouvelable (énergie solaire, éoliennes, ...) est un des arguments avancés par les auteurs des grands modèles de Facebook pour justifier du caractère soutenable de leurs projets[173]. Couplés à des optimisations algorithmiques et à de nouvelles technologies plus efficaces, ils avancent ainsi que les grands modèles sont plus verts que leurs prédécesseurs. Les considérations ne sont pas uniquement énergétiques, puisque le coût carbone associé à la recherche en IA est de plus en plus mis sur le devant de la scène[62; 105; 104; 129; 173]. Les chercheurs de Google proposent ainsi une revue du coût carbone de l'entraînement de leurs grand modèles[130], en invitant la communauté scientifique à se tourner vers des modèles plus efficients en énergie et moins coûteux en équivalent carbone.

Dans une autre direction, certaines initiatives offrent de réduire la taille des modèles et des nécessités computationnelles pour réduire les coûts inhérents à leur taille. Toutes les optimisations sont ici considérées : taille du modèle (nombre de couches, de paramètres), taille des bases de données, vitesse de convergence lors de l'entraînement, consommation des opérations sur le support matériel, déploiement sur des terminaux embarqués[40]. La multiplicité des solutions apportées au problème de la consommation énergétique des modèles, couplée à l'incitation dans les conférences et revues de rapporter la consommation des expériences et entraînements, offre une voie vers une IA qui soit plus responsable de la consommation qu'elle engendre.

2.4.4.2 La recherche participative

Outre la justice algorithmique explorée en section 2.2.3, d'autres initiatives s'attellent à combattre les conséquences sociales du déploiement des modèles d'IA, en dehors de la sphère algorithmique. C'est le cas du cadre de la recherche participative. Cette recherche met en avant l'intérêt des partenaires locaux, et d'une définition collaborative des objectifs et moyens des actions de recherche[98; 120; 43]. Ce cadre permet de pallier l'un des écueil lors de la formulation des problèmes de recherche, qui est le manque d'expérience sur le terrain et de compréhension des problématiques locales. L'inclusion dans les processus de construction des modèles, des objectifs et des finalités de la recherche en fait une initiative qui a moins de chance d'être délétère pour les populations qui seront concernées par le déploiement du modèle, ces dernière étant prises en compte dans le processus de construction du système. Si ces initiatives se heurtent à des contraintes plus fortes (nécessité d'une insertion sociale dans les milieux concernés, temps consacré aux partenaires, organisation d'ateliers, ...), elles ont l'intérêt d'être mieux acceptées dans les communautés concernées par les déploiement d'application et de mieux intégrer les enjeux locaux[120].

Dans les systèmes d'IA, elle intervient parfois sous un format qui demande l'intervention d'un public pour des opérations spécifiques, comme au niveau de la collecte des données[11; 61] ou de la construction de la base de données[65; 41; 46]. Ces initiatives n'intègrent pas les parties prenantes dans toutes les étapes de construction du système, mais se reposent quand même sur une participation de ces parties prenantes.

2.4.4.3 Limites et critiques

Si les solutions présentées semblent apporter une réponse aux problèmes rencontrés, certaines sont pourtant l'objet de critiques. En premier lieu, car certaines solutions techniques améliorent un aspect d'un modèle au détriment d'un autre. Ainsi, la réduction de la taille d'un modèle peut aller de paire avec des impacts négatifs sur la performance des modèles, comme le montrent[80] dans leur étude : l'utilisation de stratégies pour réduire la taille d'un modèle conduit ce dernier à être moins performant dans les classes les moins représentées, et donc déjà défavorisées. En second lieu, il faut encore que

ces solutions soient utilisées et appliquées, notamment par les grands fournisseurs de modèles d'IA. Si on promeut des modèles plus petits et efficaces, on parle plus des très grands modèles développés par les compagnies comme Google, Facebook ou OpenAI. Ces grands modèles se développent de plus en plus et font fi des efforts pour réduire la consommation électrique globale[151; 130; 129]. Ces grands modèles sont déployés à grande échelle, parfois sans prise en compte des conséquences possibles pour les utilisateurs, comme on l'a vu avec les différents modèles de langage[17] et CLIP et les problèmes qui accompagnent leur déploiement[6].

Enfin, la plupart des solutions techniques proposées s'inscrivent dans un cadre de réflexion limité, centré sur l'occident, avantageant toujours les pays déjà les plus favorisés. Les capacités computationnelles, les connaissances et les données requises pour mettre en place et déployer ces solutions palliant les externalités négatives sont accessibles dans les grands laboratoires et compagnies acteurs du secteur de l'IA, mais constituent des contraintes majeures dans les pays du Sud. Le report du coût carbone d'un entraînement par exemple, défavorise les pays du Sud où l'énergie est plus carbonée que dans les pays du Nord. Cette métrique invite donc à utiliser les serveurs de calcul de Google ou de Facebook, optimisés pour réduire le coût carbone des calculs, plutôt que de se reposer sur des solutions locales ou moins énergivores. L'effet rebond, enfin, n'est jamais considéré dans les travaux concernant les optimisations implémentées. Il stipule qu'une optimisation énergétique concernant une technologie peut conduire à une augmentation de la consommation énergétique de cette technologie, du fait de sa meilleure accessibilité après optimisation. Il paraît alors vain de proposer des optimisations énergétiques pour réduire la consommation globale d'énergie de l'IA, quand cette optimisation aura en fait l'effet contraire. La multiplication des infrastructures liées au développement des bases de données et des centres de calculs, où tout est pourtant optimisé, confirment que l'effet rebond est respecté pour les optimisations liées à l'IA.

2.4.5 L'éthique et l'IA

En réponses aux nombreuses controverses liées aux déploiement d'IA menant à des discriminations, de plus en plus d'acteurs du domaines ont appelé à la création d'une IA éthique, ou d'une éthique de l'IA - c'est l'un des facteurs qui a popularisé la question de la justice algorithmique. L'IA soutenable soulève elle aussi la question de l'éthique, puisque questionnant la relation entre IA et sociétés humaines. Cette question est d'ailleurs absente de la plupart des travaux les plus influents en AP selon les observations de Birhane (2020), qui soulignent dans ces travaux une absence de réflexion sur les conséquences des recherches menées. Cette question est devenue assez importante pour que Hecht et al. (2021) propose l'adoption d' une réflexion sur l'impact de la recherche dans les publications scientifiques, et que la conférence "Neural Intelligence Processing Systems", impose cette réflexion à travers le "Broader Impact Statement"[18][3]. Nous questionnons dans cette partie ce qu'est l'éthique en IA, et si on peut parler d'une IA éthique.

2.4.5.1 Des concepts flous

L'une des premières difficultés quand on parle d'éthique en IA est définir précisément les termes qui sont utilisés.

L'éthique. On rapproche souvent éthique et morale, et Jean-Gabriel Ganascia différencie les deux en définissant que la morale définit les règles à suivre tandis que l'éthique est la réflexion sur ces règles[19]. C'est une distinction faite en France mais pas forcément à l'international, ce qui en complique une définition globale. "Ethics" en anglais se traduit généralement "déontologie" en français. On confondra par la suite pour des raisons pratique éthique et déontologie. Ramenée à l'IA, l'éthique semble être un ensemble de règles, normes, principes, devoirs, implémentant la morale désirée. Ce n'est alors pas un concept qui puisse être universel, puisque les morales dépendent largement des cultures, moeurs, et des conditions sociales des individus qui les mettent en place. On doit alors faire un

[17] voir https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/
[18] voir https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832
[19] Jean-Gabriel Ganascia - Ethique et épistémologie des données - https://www.youtube.com/watch?v=86aScU-u6jY

choix entre une IA permissive qui s'adapte à la morale de ses utilisateurs pour proposer une éthique inclusive, ou au contraire une IA consistante mais imposant des morales correspondant aux idéaux de groupes d'individus particuliers. La possibilité même d'une IA éthique est remise en question par Catherine Tessier[154] qui avance qu'un "programme ou une technique ne peut pas être 'éthique' en soi et ne peut être qualifié d'éthique'. L'adjectif 'éthique' [...] ne peut être associé qu'à une démarche, une délibération, une réflexion, une question, un principe, une valeur, etc.". Ainsi, la notion même de confiance en IA ou d'IA éthique est paradoxale, attachant la confiance ou l'éthique à une technique ou des algorithmes au lieu de considérer les démarches qui les emploient.

L'IA. Il est complexe de précisément délimiter ce qui est du domaine de l'IA ou ne l'est pas. Du domaine de recherche au fantasme décrit dans les romans et jeux vidéos, du simple algorithme à l'apprentissage profond, les limites sont floues. La question de l'IA éthique, liée à l'imaginaire des IA générales, invite à penser l'IA comme étant capable de morale, anthropomorphise les techniques d'AM, invitant la machine au même rang moral que l'humain. En l'état des choses, l'IA repose sur des relations causales entre entrées et sorties, englobant un large panel de méthodes, mais aussi de pratiques en évolution rapide, suivant des principes variables. La définition de ce qu'est l'IA est complexe, ce qui contribue à la complexité d'y rattacher une éthique.

Les acteurs de l'IA. Si on parle d'éthique de l'IA, il faut aussi mentionner comment cette dernière est mise en place, et qui la met en place. L'IA évolue dans un contexte particulier du domaine informatique - un domaine par essence assez ouvert, pratiquant le partage de code, mettant en avant le travail collaboratif, l'entraide dans la construction de briques complexes. En théorie, n'importe qui peut alors s'emparer des outils pour faire de l'IA. Il est difficile de considérer alors qu'une éthique de l'IA puisse exister si on ne peut intégrer l'éthique de manière informatique dans un programme et que tout le monde peut composer de l'IA. On peut s'intéresser plus particulièrement aux acteurs du développement de l'IA, bien que tous les maillons de la chaîne ne possèdent pas de pouvoir décisionnel. Nous nous concentrerons donc sur les acteurs de la recherche et des groupes industriels qui développent les squelettes et briques qui sont ensuite utilisées partout ailleurs pour la construction des modèles d'IA.

2.4.5.2 Contradictions entre IA et éthique

A différents stades de développement de l'IA, on peut observer des contradictions entre ce qu'est l'IA ou la manière dont elle est utilisée et la conception d'une éthique de l'IA. Sans en faire une liste exhaustive, nous rappelons ici quelques éléments qui soulignent la nécessité d'une réflexion autour de l'IA et de l'éthique

Les conséquences environnementales. Nous avons déjà souligné quelques unes des conséquences de l'IA sur l'environnement ; et il faut rappeler que les innovations technologiques des dernières années ont largement contribué à l'intensification de l'exploitation des minéraux sur la surface terrestre. La multiplication des terminaux de tous types, alimentée par la nécessité toujours plus grande de bases de données pour alimenter les modèles d'IA, pousse à une extraction toujours plus intensive. La Banque Mondiale s'inquiète d'ailleurs de la tendance de ces besoins supplémentaire dans les années à venir[172]. Si l'IA ne peut être portée responsable de tous les maux du numérique, il serait intéressant d'estimer la part du numérique aujourd'hui imputée à l'IA, et les conséquences environnementales associées.

Les défis de la collecte de données. La construction de bases de données et une étape nécessaire pour la plupart des systèmes d'AM. Ces constructions reposent sur un ensemble de choix techniques qui relèvent de la nature des données et des méthodes et supports utilisés. Ces choix reflètent bien souvent les conditions socio-culturelles dans lesquelles ces bases de données sont développées. Ainsi, les premières bases de données d'images utilisées par les communautés scientifiques pour comparer leurs approches comprenaient pratiquement uniquement des images provenant d'occident, inscrivant ainsi les applications en vision par ordinateur dans un contexte occidental uniquement[147; 49]. Mais l'origine des données n'est pas la seule question problématique de la collecte de données ; les questions d'autorisation, d'annotation dégradante, de légitimité de la collecte sont importantes et relevées

par Kate Crawford[20], des sites spécialisés comme Exposing.ai[21] ou des travaux académiques[110; 9].

Les pratiques de la collecte de données sont encore éthiquement problématique quand on considère l'annotation des données collectées. En France, des reportages comme Cash Investigation 'Au secours, mon patron est un algorithme'[22] ou encore la série Arte 'Invisibles - les travailleurs du clic'[23] dénoncent l'utilisation de travailleurs défavorisés dans des pays à bas revenus qui sont employés à annoter des bases de données à destination d'organisation occidentales. Antonio Casilli dénonce dans "En attendant les robots : Enquêtes sur le travail du clic"[25] les pratiques d'exploitation visant uniquement à améliorer les prédictions des IA. Les pratiques de collecte et d'annotation de données ont donc encore de nombreux défis à relever si on veut pouvoir les qualifier d'éthique.

Les calculs et leur impact. On a déjà évoqué que l'IA soutenable doit prendre en compte ces externalités négatives, dont les calculs intensifs et l'énergie nécessaire pour ces derniers. Une conjonction entre IA et éthique viserait donc logiquement à la réduction des calculs réalisés pour ces IA. De fait, il n'en est rien ; la course folle aux capacités de calcul bat au contraire son plein, avec en tête les géants du numérique comme Facebook, Google et Amazon, et des compagnies plus modestes mais tout autant financées comme OpenAI. Le modèle GPT-2 de cette dernière, développé en 2019 et fort de 1,5 milliards de paramètres, fut détrôné l'année suivante par son successeur GPT-3 et ses 175 milliards de paramètres, contre 15 milliards pour son homologue chez Microsoft la même année, Turing NLG. L'entraînement de tels modèles nécessitent des super-calculateurs, des entrepôts de données, un temps et des ressources considérables. Ces modèles et leurs successeurs à venir évoluent en dehors des considérations éthiques pour une réduction des externalités négatives de l'IA, et conduisent l'ensemble du domaine vers une augmentation de ces dernières.

Quelles applications pour l'IA. Quelles sont les finalités des modèles développés en IA ? Si certaines se réclament vertueuses comme les applications mentionnées en IA pour la soutenabilité, le reste des applications s'intègrent généralement en tant que processus de traitement de données, en tant que brique dans la conception de plus grandes applications. Le contrôle de la finalité de ces dernières est hors du cadre d'une réflexion éthique, bien qu'on puisse conceptualiser une certaine idée de la forme qu'une telle régulation pourrait prendre après les scandales de Cambridge Analytica[24] ou de Clearview AI[25]. Le potentiel nocif de l'IA a aussi été souligné par les applications comme DeepFake et la problématique des fake news dans les flux d'informations de nombreux médias et réseaux sociaux. Est-il encore aujourd'hui éthique de continuer à développer une technologie avec un tel potentiel de nocivité ? Une telle réflexion est en dehors du cadre de cette thèse, mais encourage à réfléchir aux finalités des travaux entrepris dans de nombreuses organisations du secteur.

Déploiements et impacts. En plus d'applications expressément nocives, d'autres applications à priori neutres ou à vue sociale peuvent avoir des conséquences imprévues et délétères pour les populations défavorisées. Nous avons vu des exemples de telles situations dans les sections précédentes. Virginia Eubanks expose dans "Automating Inequality"[26] les dérives de modèles déployés aux quatre coins du monde, dans les systèmes de justice, de police préventive, ou encore de recrutement. Au lieu de lutter contre les inégalités, les IA exacerbent donc certaines problématiques sociales, allant à l'encontre de pratiques éthiques.

Une justice difficilement à la hauteur. Le cadre juridique pour l'encadrement de l'IA peine aujourd'hui à rattraper les pratiques déjà en place et les modèles déjà déployés. L'Europe est la première des puissances mondiales à tenter de réguler l'IA, avec en 2018 un 'Plan Coordonné sur l'IA', puis en 2019 la création du Goupe d'Expert Haut Niveau sur l'Intelligence Artificielle (HLEG), qui produira plusieurs rapport dont un livre blanc pour la Commission Européenne en 2020, et qui sera à l'origine de la première proposition de régulation européenne de l'IA en 2021, l'IA Act. Ce

[20] Site : Excavating AI, https://excavating.ai/

[21] https://exposing.ai/

[22] https://www.francetvinfo.fr/replay-magazine/france-2/cash-investigation/cash-investigation-du-mardi-24-septembre-2019_3603915.html

[23] https://www.france.tv/slash/invisibles/

[24] https://fr.wikipedia.org/wiki/Scandale_Facebook-Cambridge_Analytica

[25] https://fr.wikipedia.org/wiki/Clearview_AI#Controverses_et_proc%C3%A9dures_judiciaires

[26] https://virginia-eubanks.com/automating-inequality/

texte n'est pas exempt de critiques, dont celle de Thomas Mezinger, membre du HLEG, qui déplore la faiblesse des propositions de cet IA Act et le juge à peine bon à produire du "blanchiment éthique"[27]. Le cadre juridique peine à se montrer à la hauteur afin de promouvoir une IA éthique ou au moins réglementée.

Les dégâts de l'automatisation. Vinuesa et al.; Bender et al. (2020; 2021) rappellent que l'IA profite d'abord aux pays qui sont déjà avantagés et qui possèdent les infrastructures et les technologies les plus avancées. L'écart creusé s'agrandit donc encore entre les pays du Nord et du Sud, renforçant les relations de domination existantes. Ces relations de dominations sont reflétées en IA par le pillage de données et l'absence d'inclusion dans les bases de données. Il est dur de ne pas voir dans ces pratiques un néo-colonialisme technologique et numérique, porté aussi bien par les pays du Nord que par les Big Tech[136; 72; 28; 2; 12; 44].

Une recherche au service du marché. Enfin, l'IA est en grande partie au service du capital. La recherche n'est pas, comme se le représente un imaginaire collectif, indépendante, mais repose plutôt sur les financements de l'industrie et la réponse d'appels à projets - et donc sur des objectifs à atteindre, définis en dehors du cadre de la recherche. Ces objectifs peuvent être liés au développement de l'industrie ou du capital. Les grandes conférence en IA, lieu de prestige pour les chercheurs qui viennent y présenter leurs travaux, sont elles-même financées par les grand groupes industriels. Fort de financements importants, l'IA ne peut s'extraire des logiques de marché, et la recherche n'échappe pas à cette logique[72].

2.4.5.3 Textes officiels et chartes

Ces contradictions sont le terreau de réflexions sur les moyens d'intégrer l'éthique à l'IA. Conscients que l'éthique n'es pas qu'une affaire technique, d'autres initiatives tentent d'engager les responsabilités des personnes - pas sur un plan légal, mais moral, à travers des textes et chartes. En s'inspirant de la déclaration d'Helsinki de l'Association Médicale Mondiale de 1964, du protocole de Nagoya de 2010, de la Déclaration de Singapour sur l'intégrité en recherche de 2010 et du code de conduite européen pour l'intégrité en recherche publié par All Academies en 2017, diverses chartes ont vu le jour pour témoigner de l'engagement d'associations ou de groupes pour des valeurs ou une morale particulière. Ces chartes proposent un encadrement de la recherche, du développement et du déploiement de l'IA, et proposent ainsi une vrai démarche éthique autour de l'IA. Elles promeuvent avant tout des valeurs comme la transparence, la justice, l'équité, la non-malfaisance, la responsabilité, la confidentialité, la bienfaisance, la liberté et l'autonomie, la confiance, la soutenabilité, la dignité et la solidarité. Néanmoins, la variété et le nombre de ces chartes interroge sur l'utilité de ces dernières. Jobin et al. (2019) liste 84 chartes différentes, issues de différents organismes acteurs du secteur de l'IA, soulignant l'absence de normes, bien que ces chartes défendent globalement les mêmes valeurs. Le caractère non contraignant de ces chartes est toutefois un obstacle à leur pertinence : sans pénalité ou tribunal jugeant de l'action des organismes ne respectant pas leurs propres chartes, celles-ci ne peuvent être prises au sérieux. Ceux qui adoptent ces chartes sont donc en même temps juge et parties de leur propre éthique, et ces chartes apparaissent vite comme un moyen de continuer le "business as usual" dans une stratégie dite du "porte-manteau" déjà décriée dans le milieu académique[15; 1; 169; 72]. Les chartes et textes d'encadrement sont donc aujourd'hui aussi vues comme des diversions permettant d'afficher une façade éthique plutôt que des engagements forts et respectés par ceux qui les signent.

Une autre critique adressée à ces chartes est la mé-compréhension du problème éthique lié à l'IA, puisque si certaines chartes insistent sur le facteur humain et la responsabilité des décideurs et intervenants concernés, elles proposent pour certaines d'utiliser pour ce faire certaines métriques et indicateurs. Elles proposent ainsi de répondre à un agenda technique et contribuent à une transformation d'un problème social à un problème technique, qui est problématique pour au moins quatre raisons :

[27] voir https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

- Un problème technique invisibilise le problème social concerné[118]. Si un modèle défavorise une partie de la population, la solution technique sera d'aligner les prédictions du modèle pour que ce dernier fonctionne mieux pour tout le monde, tandis qu'une analyse de ce phénomène nous permettrait peut-être de comprendre pourquoi ce phénomène émerge initialement et de combattre ce phénomène sur le plan social.

- La compréhension des phénomènes sociaux est encore limitée, et des initiatives luttant contre un biais particulier pourraient se retourner contre les populations concernées par le biais. Par exemple, un effort pour collecter plus de données sur une population défavorisée pourrait exacerber leur marginalisation, en permettant un contrôle étatique plus poussé sur ces dernières.

- Le néo-colonialisme transpire à travers ces chartes, qui sont toutes ou presque issues de cultures occidentales[28]. Sambasivan et al. (2021) propose ainsi de recontextualiser ces principes en incluant les populations les moins représentées sur la scène mondiale.

- L'incompatibilité des métriques de certains principes défendus par les chartes réduit encore la crédibilité de ces dernières. Agarwal; Agarwal (2021a; 2021b) montrent comment les notions d'équité et de confidentialité s'opposent, tout comme les notions d'équité et d'interprétabilité.

2.4.5.4 La diversion et les intérêts

Si aujourd'hui de nombreux efforts scientifiques sont tournés vers l'IA éthique, ils le font via la justice algorithmique, et donc par un aspect technique. Ainsi, dans la plupart des cas, seuls les aspects de l'entraînement et du contrôle des modèles sont aujourd'hui remis en question par ces efforts, quand le reste du processus d'un système d'IA reste inchangé[115]. C'est une position qui assume des choix techniques déjà réalisés par la société (utilisation des techniques d'IA, donc récolte et traitement de données associés) et qui implémente son acceptation. Cette méthodologie est en fait un aveu d'échec, puisqu'elle ne remet pas en cause l'utilisation de l'IA dans les problématiques considérées, mais assume son utilisation et projette de la rendre plus acceptable par des procédés techniques. La justice algorithmique semble alors plus faire office de pansement que de réelle solution au problèmes sociaux[28].

Cette situation est aussi le fruit d'un déséquilibre flagrant : les incitations à déployer de l'IA et à en tirer profit sont bien plus fortes que celles à réguler et contrôler. La progression des prouesses techniques étant vue comme inévitable, personne n'est alors tenu pour responsable de ses travers, malgré les nombreuses médiatisations des complications liées à ces déploiements. Au contraire, au niveau individuel, il y a une forte incitation à travailler dans le domaine de l'intelligence artificiel, qui est aujourd'hui "en vogue" : il assure une paie conséquente, un statut social particulier, et la possibilité de trouver assez facilement un emploi. Il n'est pas évident de proposer une réflexion idéologique sur la place de l'IA dans un tel climat, qui la place au centre d'un monde qui tourne par et pour cette technologie[38; 64].

2.4.5.5 Une absence de communauté

Imposer des pratiques éthiques à un domaine particulier n'est pourtant pas mission impossible. De nombreux domaines fonctionnent selon des codes d'éthique encadrés avec des organismes de régulation visant à ce que ces codes soient appliqués, comme en santé ou dans les GLAM (Galeries, Librairies, Archives et Musées). En santé, bien que tout le monde ait entendu parler du serment d'Hippocrate, ce n'est pas ce serment qui a valeur juridique, mais des codes nationaux régulièrement actualisés. Des buts communs, une histoire professionnelle, des méthodes pour passer du principe à la pratique et des mécanismes légaux robustes pour la responsabilité des pratiquants[112; 86] sont autant de moyens pour le domaine de la santé de maintenir en place une éthique spécifique. La ratification à un serment a gardé valeur symbolique, mais n'est aucunement légalement contraignante.

[28] voir https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Pour les GLAM, les pratiques sont régulées pas l'UNESCO principalement et d'autres entités à diverses échelles, dont le principe est simple : des instances de décisions sont capables de faire autorité auprès des membres partageant un métier. C'est là la différence principale avec les pratiquants de l'IA : il n'existe aucune instance de décision reconnue globalement, et il n'y a pas d'unité du métier ; les acteurs du monde de l'IA sont trop épars pour tous s'identifier sous une même casquette. Il n'y a pas d'histoire ou de buts communs, cette technologie étant très récente et possédant des applications extrêmement diverses.

C'est un problème inhérent au monde du numérique, qui a déjà son homonyme au sein de la communauté de la cyber sécurité, qui utilise les appellations "white hat", "red hat", "grey hat", "black hat", pour désigner différentes éthiques appliquées dans le métier. La situation est similaire avec les pratiquants de l'IA : un appel à l'éthique qui ne parle pas à tout le monde, des pratiques diverses et morcelées, un fort investissement de la part du complexe militaro-industriel pour la défenses de ses intérêts privés. L'absence de communauté capable de s'entendre sur un code à suivre complexifie donc encore la question d'une éthique de l'IA.

2.4.5.6 Des propositions pour une autre éthique de l'IA

Malgré les obstacles à la mise en place d'une éthique de l'IA ou d'IA éthique, certaines initiatives porteuses d'autres visions voient le jour, et apparaissent comme des possibles voies à emprunter. Sans être exhaustif, nous en présentons ici un échantillon.

Cartographie des relations socio-techniques. Certaines initiatives visent à cartographier les relations sociotechniques induites par les nouvelles technologies pour appréhender les problématiques qui résultent de ces relations[23]. Ces initiatives permettent d'identifier et de prévoir les externalités négatives et d'engager des réflexions, politiques ou scientifiques, sur ces dernières. Ces explorations sont néanmoins limitées par le dilemme de Collingridge : "Il est difficile d'anticiper les conséquences sociales d'une technologie, et lorsque ces effets, notamment négatifs, remontent à la surface, il y a de grandes chances que la technologie en question soit déjà incontournable et inévitable". Ces cartographies constituent donc un appel à la prudence et une première initiative pour prévenir les conséquences de déploiement des modèles, mais se montreront insuffisantes pour prévoir toutes ces conséquences.

Toolkits et frameworks. De nombreux outils et frameworks sont développés par la communauté scientifique pour appuyer le développement d'IA éthique. En section 2.2.3, nous avons vu des outils pour soutenir le déploiement d'IA régulant certains biais spécifiques. De tels outils existent aussi pour le reste de la chaîne de traitement du développement des systèmes d'IA, et donnent des directives pour vérifier des métriques spécifiques. Certains outils peuvent même être contraignants, comme l'outil développé au Canada pour évaluer l'acceptabilité d'un système. Leur multiplication est autant témoin d'un intérêt particulier pour une IA plus juste, tout en étant aussi source de confusion. Quelle confiance peut-on accorder à ses outils, développés par les acteurs du secteur et censés réguler d'autres acteurs du secteur ? Comment identifier les standards utiles et les autres ? L'accumulation d'outils et de standards ne risque-t-elle pas de nuire à l'établissement de normes pour une IA éthique ? On atteint ici la limite de ces outils, qui sans soutient et validation par des organismes reconnus par une communauté, n'engagent que ceux qui les utilisent.

Utilisation des métadonnées. De plus en plus d'initiatives soutiennent l'intérêt d'accompagner les productions en IA (bases de données, modèles, systèmes) de métadonnées explicitant les contenus, les outils utilisés et les métriques suivies pour le déploiement de ces productions. Shen et al. (2021) proposent ainsi des cartes de description des modèles, listant non seulement les éléments permettant de reproduire le modèle, mais aussi le coût énergétique de ce dernier, permettant la création de labels à partir de ces métriques. Benjamin et al. (2019) proposent de lier une nouvelle licence aux bases de données, précisant divers aspects de sa construction et ses possibles utilisations. D'autres proposent encore de s'inspirer des pratiques des archives pour la production et la conservation des bases de données utilisées pour l'entraînement des modèles[58] ou d'établir des labels pour les bases de données[79]. Ces documentations pourraient être des oeuvres collaboratives, permettant des échanges

entre acteurs et organisations du secteur.

Formation des individus. Les approches qui reposent sur les individus acteurs de l'IA ne sont pas les plus nombreuses, et ne concerneront que ceux qui se sentent concernés par les problématiques soulevées. Néanmoins, plusieurs initiatives se reposent sur la sensibilisation, l'éducation et la formation à l'IA éthique. En France, un arrêté du 25 mai 2016 sur la formation doctorale exige par exemple que "tout(e) doctorant(e) reçoive une formation à l'éthique de la recherche et à l'intégrité scientifique". La charte nationale de déontologie des métiers de la recherche ajoute que les règles déontologiques doivent intégrées aux cursus de formation, en particulier au sein des cursus de master et doctorat. La formation éthique est donc institutionnalisée, dans le processus de formation au moins. Dans la même lignée, des jeux sérieux sont proposés pour introduire des réflexion sur les externalités négatives de la recherche[68]. C'est le cas des initiatives du Labo 1.5, qui sous la forme du jeux sérieux nommé "Ma Terre en 180 minutes", propose de simuler un laboratoire devant réduire son empreinte carbone. Cette initiative est directement inspirée de la Fresque du Climat[29], un outil de sensibilisation aux problématiques du changement climatique. Au niveau international, le "Broader Impact Statement" de la conférence NIPS évoqué plus tôt est une invitation pour le moment non contraignante à la réflexion sur les activités de recherche. Il n'existe pas encore de limites contraignantes pour les travaux de recherche en IA publiés dans le monde, mais l'apparition de telles contraintes pourraient stimuler une recherche plus en accord avec des principes éthiques partagés par une communauté particulière.

2.4.5.7 L'impasse

Au final, l'adjonction de l'adjectif "éthique" à l'IA reste une impasse. La lutte contre les biais internes des modèles reste la principale activité liée à la recherche en éthique de l'IA, lutte qui est nécessaire mais pas suffisante à ce que toute la démarche de développement d'un système d'IA soit considérée éthique. Il faudra donc chercher l'éthique dans les moyens et les fins associées aux technologie de l'IA plutôt que dans l'IA en elle-même.

2.5 L'IA au Sud

Les questions d'une IA soutenable, pour la soutenabilité, et éthique, se rejoignent dans les défis soulevés par l'utilisation de ces technologies dans les pays du Sud. Nous explorons dans cette partie quelles sont les considérations spécifiques liées au développement de la vision par ordinateur au Sud et comment les questions précédentes interviennent dans ces considérations spécifiques.

2.5.1 Les données et le contexte au Sud

Les travaux de Hohman et al.; Cichos et al. (2019; 2020) rappellent l'importance de la disponibilité de large volumes de données pour booster la performance des modèles entraînés sur ces dernières. Les données sont le moteur de la performance des modèles d'IA quelque soit la tâche à réaliser. Que ce soit pour adapter les modèles déjà développés ou ou entraîner de nouveaux modèles, il est donc nécessaire de disposer de données pour assurer des performances équivalentes quelque soit la géographie dans laquelle on déploie ces modèles. La complexité d'accès aux données dans les pays du Sud est pourtant documentée[27; 45; 53; 2]. Dans cette partie, nous relevons la situation et les défis de la collecte de données au Sud.

2.5.1.1 Répartition dans les bases de données

Nous avons vu précédemment avec les travaux de Shankar et al. (2017) et de DeVries et al. (2019) que les grandes bases de données d'image, populaires, utilisés par une grande partie de la communauté scientifique, manquaient d'inclusivité et manquaient particulièrement de représentativité pour les pays du Sud. Ces bases de données, issues des premières initiatives pour la création de larges volumes de

[29] https://fresqueduclimat.org/

données images annotées, n'ont pas été pensées pour être inclusives. Collectées grâce à des processus automatiques via des recherches sur différentes bases de données d'images en ligne, elles reflètent les biais culturels de leurs concepteurs anglo-saxons[107; 29; 26]. D'autres bases de données depuis s'efforcent de pallier ce manque de représentativité, et ces initiatives soulèvent de nouveaux défis à surmonter.

2.5.1.2 Les défis d'une collecte au Sud

La collecte de données image au Sud révèle des défis qui étaient encore inconsidérés. Les solutions classiques de collecte massive de données, via des bases de données d'images en ligne comme Flickr, Google Images et d'autres, se montrent compliquées car ces dernières sont majoritairement composées de données occidentales. Cette situation est due à un développement plus tardif des infrastructures numériques, ainsi qu'à un manque d'intérêt des grands groupes du numérique pour ces zones géographiques, ne procurant du coup pas les mêmes services avec les mêmes moyens. Il est donc nécessaire de trouver d'autres moyens de se procurer des données. Plusieurs initiatives ont vu le jour, utilisant majoritairement deux voies différentes : le tri dans les données massives et l'emploi d'éléments humains.

Le tri dans les données massives se base sur les métadonnées géographiques associées aux images. Ainsi, Dubey et al.; Kalluri et al. (2021; 2023) utilisent des bases de données massives comme YFCC100M[155] ou WebVision[99]. Ces données sont triées grâce aux métadonnées géographiques, et les données issues des pays les moins représentés peuvent ainsi être extraites et organisées en base de données pour ces pays. Ces initiatives permettent des collectes massives de données, bien qu'elles ne pallient pas tous les problèmes liés à ces dernières - comme les biais anglo-saxons liés aux annotations ou le soupçon de provenance orientée de la donnée (par exemple, une sur-représentation des données issues de vacanciers occidentaux dans les pays du Sud).

L'emploi d'éléments humains met l'emphase sur la qualité des données collectées, en employant des personnes directement culturellement rattachées aux zones géographiques où les données sont collectées. Gaviria Rojas et al. (2022) et Atwood et al. (2020) se reposent ainsi sur des méthodes de collecte de données en local sur différentes zones géographiques pour collecter des données images spécifiques aux localités considérées. Ces collectes sont plus coûteuses et plus parcimonieuses que les collectes algorithmiques, et ne sont pas forcément non plus exemptes des biais de conception des bases de données (dans les directives du type de données qui doit être collectée par exemple, ou des annotations à réaliser qui relèvent d'une certaine empreinte culturelle).

Défis techniques. Ces initiatives permettent de répondre au premier défi technique de la disponibilité de la donnée, mais en soulèvent de nouveaux. Comment collecter des données massives pour les zones géographiques les moins représentées ? Le tri dans les bases de données existante apporte une première réponse, mais soulève le problème de la qualité des données, et de la provenance effective de ces données. L'emploi d'éléments humains pallie ce soucis, mais se heurte au volume de données qu'il est possible d'ainsi collecter sur des temps ou budgets donnés. Ces deux solutions se complètent l'une l'autre sans être pour autant compatibles, et soulèvent le défi d'une technique qui se trouve à mi-chemin entre ces deux solutions et qui puisse offrir une solution satisfaisante au problème de la disponibilité de la donnée image dans les pays du Sud.

Défis éthiques. Ces solutions techniques soulèvent des questionnements éthiques, notamment sur la légitimité des actions de collecte de données et sur la manière dont ces dernière sont utilisées. Les collectes de données interrogent sur le plan légal le droit au respect de la vie privée des individus dont les données sont collectées. En ciblant des populations spécifiques, on peut tout aussi bien lutter contre des inégalités[76] que révéler des informations sensibles permettant l'oppression de ces populations[30]. Outre les questionnements sur la vie privée ou la sécurité des données, se pose aussi la question de la légitimité d'une telle collecte. Les systèmes d'IA étant développés majoritairement par et pour l'occident, la collecte de données au Sud est parfois décriée comme une forme de néo-

[30] Voir les critiques de COMPAS : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

colonialisme[113; 2]. Ces collectes interrogent aussi la souveraineté digitale des pays concernés, et soulèvent donc de nouveaux défis sur le plan éthique, même dans le cadre d'une meilleure représentativité de ces pays dans les bases de données d'image.

2.5.1.3 Adaptation des méthodologies de la collecte pour le Sud

Des initiatives voient le jour pour mettre au point des outils permettant d'adapter les techniques existantes pour la collecte de données inclusives[50]. Nous avons dans cette thèse proposé notre propre approche pour la collecte massive de données d'une manière plus inclusive. Nous avons adapté la méthodologie de génération des bases de données comme MS COCO[106], utilisant des requêtes sur la base de données Flickr[31], pour inclure des données en provenance de toutes les zones géographiques du monde[18]. Nous avons publié une base de donnée extraite grâce à cette méthodologie, et annotée pour 400 données dans chacune des 23 zones géographiques utilisées lors de la collecte[16]. Si cette méthodologie a la convénience de pouvoir passer à l'échelle et d'être utilisée dans de nombreux contextes, n'étant pas rattachée à une base de donnée ou une architecture particulière, elle rencontre cependant des soucis de qualité de la donnée, la provenance des données étant encore plus remise en cause que dans le cas du tri de données. De nouvelles voies sont encore à explorer pour la génération de données inclusives, et la manière dont cette génération peut prendre en compte les individus concernés par ces collectes de données.

2.5.2 Le développement de modèles

L'étape du développement de modèles au Sud n'est pas exempt de nouveaux défis à surmonter. Outre les conséquences dans les modèles développés du manque de données locales, on retrouve de nombreux biais qui interviennent dans la conception des modèles, comme la langue anglaise généralement utilisée pour les annotations (intégrant ainsi les concepts anglo-saxons), le choix des catégories, des applications, et autres biais d'aggrégation. Les pratiques et normes dans le développement de modèles correspondent pour le moment à des contextes socio-culturels occidentaux, pénalisant les autres pratiquants. Cvitkovic (2019) en fait mention dans son appel, et Sambasivan et al. (2020) souligne comment des normes éthiques ont besoin d'être adaptées dans le contexte indien.

On retrouve également l'attachement dans le développement de modèles à un contexte technique qui n'est pas partagé mondialement. Les modèles développés sont ainsi de plus en plus gourmands en énergie et en capacité de calcul, et les capacités de calcul augmentent plus rapidement au Nord. Par exemple, il n'y a à l'heure actuelle qu'un calculateur de haut niveau en Afrique Subsaharienne, au Sénégal. Il est donc difficilement envisageable aujourd'hui de développer des modèles à l'état de l'art sans utiliser les ressources des pays du Nord, une situation qui avantage ces derniers. L'adoption de ces technologies se fait au dépend d'une autonomie locale pour les pays n'ayant pas encore les infrastructures permettant des développements en interne.

2.5.3 Déployer une IA au Sud

Les performances annoncées des modèles sur leurs jeux tests laissent percevoir des prouesses, pourtant leur déploiement sur le terrain peut parfois être compliqué, à cause de conditions qui diffèrent entre laboratoire et terrain. Cette situation, qui a initié les travaux sur les glissements de contexte, est à l'origine des benchmarks WILDS[92] qui mettent l'accent sur les différences entre jeux d'entrainement des modèles, jeux de tests et données considérées "terrain". Si ce phénomène est source de défis partout dans le monde, certains sont spécifiques au Sud.

2.5.3.1 Le glissement de contexte

Ce glissement survient quand les données utilisées pour entraîner et tester un modèle possèdent un contexte différent des données sur lesquelles le modèle est déployé pour une application quelconque.

[31] https://www.flickr.com/

Ce phénomène est en premier lieu observé de manière systématique par Torralba and Efros (2011), qui constatent qu'il est possible d'entraîner des modèles à retrouver la base de données dont sont issues des images, et qu'il existe donc un biais spécifique aux bases de données utilisées, issu du contexte de conception de ces bases de données. Rapporté au déploiement au Sud, ce glissement de contexte a pour conséquence une sur-évaluation des performances d'un modèle, qui pourra se tromper beaucoup plus fréquemment sur des contextes non rencontrés durant l'entraînement. Spécifiquement au Sud, ce phénomène est accentué par l'absence presque totale de données issues du Sud dans les bases de données d'image, le glissement de contexte étant alors poussé au paroxysme.

2.5.3.2 Les défis au Sud

Cvitkovic (2019) identifient en Afrique de l'Est une difficulté à intégrer les données terrains dans les modèles, trop brutes et nécessitant des pré-traitements coûteux. Ils identifient aussi que beaucoup de bases de données y sont stockées sous format SQL, et que la plupart des efforts pour déployer des modèles se retrouvent dans la traduction de ces bases de données vers le format plus classique vectoriel en AM. La faible disponibilité des ressources en terme de volume de données ou de capacité computationnelle est aussi abordée ; insistant sur l'importance des modèles pour les terminaux à faible dimension et consommation énergétique, et sur les systèmes permettant l'utilisation de faibles volumes de données. Ces travaux mettent non seulement en lumière des oppositions entres tendances actuelles en AM (augmentation des volumes de données et des capacités computationnelles) et réalité terrain en Afrique de l'Est, mais éclairent aussi sur les défis rencontrés dans une région particulière concernée par le déploiement de modèles dans des contextes Sud.

Les défis cités ici ne le sont pas de manière exhaustive, mais permettent de mieux cerner comment le déploiement de modèles est complexifié par des contextes socio-techniques spécifiques. Trivedi et al. (2019) soulignent encore cette complexité, montrant comment des données ouvertes en Inde peuvent mener à des conclusions erronées et dangereuses lors d'utilisation de modèles et métriques sans compréhension du contexte local. Gonzalez et al.; Okolo (2019; 2020) quant à eux, soulignent comment certaines applications se montrent surprenamment peu effectives dans le contexte des pays du Sud, soulignant la nécessité d'une réflexion particulière lors de l'utilisation d'apprentissage par transfert dans ces cas précis.

2.5.4 Le caractère occidental de l'IA

2.5.4.1 Définitions

Nous avons évoqué en section 1.2.1 que l'occident n'a pas une définition précise ou scientifique, et que nous faisions le choix dans cette thèse de considérer comme occidental tout ce qui est produit culturel de l'Amérique du Nord, de l'Europe, de l'Australie et de la Nouvelle Zélande. Les données ainsi que modèles et méthodologies issues de ces zones géographiques sont donc considérées occidentales, en opposition avec ce qui est issu du reste du monde.

2.5.4.2 Impact sur les modèles

L'AM se repose sur l'apprentissage de caractéristiques particulières dans les données pour résoudre les tâches qui lui sont fournies. Ce faisant, un modèle n'apprend pas seulement la tâche qui lui incombe, mais les biais spécifiques du contexte d'entraînement[157; 123]. Ce contexte d'entraînement regroupe tout le processus de développement du modèle, inclue toutes les personnes et ressources nécessaires au développement du modèle. Plus ces ressources sont ancrées dans un contexte occidental, plus le modèle sera ancré dans ce contexte occidental, et en aura donc les biais, tant que ceux-ci ne sont pas identifiés, cernés, et palliés. Hors occident, d'autres zones géographiques s'impliquent dans le développement des technologies de l'IA (en Asie de l'Est, en Russie, etc..), mais aucune de ces zones ne semble inclure le Sud dans ses préoccupations pour le développement de la technologie de manière uniforme. Les modèles développés restent non inclusifs, et excluent les pays défavorisés.

De plus, les efforts pour développer des modèles favorisant le développement des plus démunis se fait souvent avec une vision extérieur à leurs contextes, provoquant possiblement des effets non anticipés, souffrant d'un manque de validation des outils employés, d'algorithme biaisés et de manque de régulations. Blumenstock (2018) souligne ces possibles écueils et encourage à sortir d'une vision occidentale et d'augmenter les partenariats et initiatives internationales pour pallier ces problèmes.

2.5.4.3 Pallier la non inclusivité

La réponse classique, technique, au problème de la non inclusivité des modèles, est l'utilisation de stratégies de généralisations issues des domaines de l'adaptation ou de la généralisation de domaines. Par exemple, Kalluri et al.; Dubey et al. (2023; 2021) proposent des benchmarks en adaptation et généralisation de domaines pour améliorer les capacités des modèles dans les zones géographiques de ces benchmarks. L'intégration de ces nouvelles données plus inclusives dans les données d'entraînement des modèles fondateurs est une voie pour la démocratisation de modèles plus inclusifs. Mais comme évoqué plus tôt, les pays du Sud sont exclus du développement de ces modèles, et si les performances des modèles peuvent être améliorées sur des modèles généraux, ils ne pourront pas pour autant être considérés inclusifs sans intégrer des logiques et des considérations plus spécifiques aux contextes du Sud.

2.5.5 Les approches participatives

Évoquée en section 2.4.4.2, la recherche participative permet d'inclure les partenaires locaux dans les prises de décision dans un processus de recherche. L'intégration des problématiques et des considérations des acteurs sur le terrain permet une meilleure intégration de l'outil dans la culture liée à ce terrain. Les approches participatives sont ainsi un excellent moyen de pallier la non inclusivité et de mettre en avant les problématiques spécifiques au contexte des populations ciblées. Cette approche, si elle nécessite de ralentir le rythme de développement, intègre une dimension sociale qui assure la pertinence du modèle développé.

Durant cette thèse, nous avons proposé une approche participative pour le développement de modèles de détection de pirogues sur les ports de pêche artisanaux des côtes du Sénégal[17]. Pour ce faire, nous avons mis en place un groupe de travail constitué de représentants des pêcheurs, de scientifiques de l'océan et des pêcheries, et d'officiels gouvernementaux et militaires issues des instances de gestion des aires marines protégées. De nombreux aspects de la gestion des ressources halieutiques ont été discuté durant ces réunions de travail, définissant le cadre et l'intérêt d'un modèle capable de compter les pirogues d'un port de pêche pour les relevés de certains indices de suivi économique et halieutiques. Cette approche s'est reposée en plus sur une génération de données locales et un développement dynamique, prenant en compte les ressources limitées et les spécificités des terrains concernés. Cette initiative intègre les critiques ciblant le caractère non inclusif des modèles grands publics, et est un premier jalon vers une IA qui soit plus intégrée dans le paysage du Sud. Plus qu'une application majeure, c'est une preuve de concept, un exemple permettant de montrer qu'il est possible de pratiquer une IA autrement qu'en intégrant uniquement des concepts définis par le Nord.

Chapter 3

Contribution méthodologique

Chapter Summary

Résumé du chapitre

On a pu voir en section 2.2.3.2 que les systèmes d'IA étaient décriés pour leur manque de "justice algorithmique", et que cette justice algorithmique prenait diverses formes. L'une de ces formes est l'inclusivité des différentes cultures et des différentes zones géographiques. Atwood et al. (2020) sont les premiers à lier la notion d'inclusivité à celle d'origine culturelle ou géographique d'une donnée, en définissant des images inclusives comme des images issues d'un endroit ou d'une culture peu ou pas représentée dans les bases de données d'images génériques. L'inclusivité d'un système est donc étroitement liée au biais géographique auquel ce système est soumis.

Le biais géographique est alors avancé comme un obstacle à l'inclusivité des systèmes d'IA, altérant les performances des modèles dans les contextes des pays du Sud. Les stratégies de mitigation de biais sont multiples, et soulignent un point important : la nécessité de précisément identifier le biais et ce qu'on cherche à mitiger. Une méthodologie précise pour caractériser un biais est donc nécessaire, et plus particulièrement pour le biais géographique.

On propose dans cette section une telle méthodologie, permettant la caractérisation du biais géographique dans les systèmes de classification d'image, à travers un protocole nommé STI. Ce protocole est d'abord contextualisée en section 3.1, grâce à une étude des précédentes évaluations du biais géographique et une identification des manques dans ces études. La section 3.2 présente le protocole

STI, de manière globale en premier lieu, puis point par point, en explorant les trois caractéristiques d'un biais selon ce protocole : Source, Type et Impact. On y présente ensuite des considérations supplémentaires sur l'isolation d'un biais, la caractérisation d'un glissement ou encore la manière dont on procède au report de résultat. On discute enfin des choix réalisés dans le protocole dans la section 3.3, ainsi que des opportunités que ce dernier ouvre, avant de conclure en section 3.4.

3.1 Contexte et définitions

Les systèmes de vision par ordinateur sont très populaires, et utilisés dans de nombreuses solutions, destinés à être utilisés partout dans le monde. Ces systèmes ne sont pas exempts de biais, et notamment de biais géographique. La caractérisation du biais géographique dans ces systèmes permettra d'informer les usagers des performances du système dans les contextes de déploiement considérés, et possiblement d'encourager ou de décourager le déploiement de certains systèmes en fonction de leur inclusivité.

On souhaite donc dans cette partie étudier le biais géographique des systèmes de classification d'image, et être capable d'évaluer son impact sur les performances de ces systèmes.

3.1.1 Précédentes évaluations du biais géographique

Le biais géographique est introduit par Shankar et al. (2017), qui tracent la source des données de deux bases de données d'images publiques et populaires, ImageNet et Open Images. Ils font le constat que les données sont majoritairement issues d'Amérique du Nord et d'Europe. En comparaison, la Chine et l'Inde, qui sont les deux pays les plus peuplés au monde, sont représentés par au maximum 2.1% des données dans les deux bases de données. Les auteurs proposent en conséquence une analyse des performances d'un modèle réalisant de la classification d'images sur des images comprenant une plus grande diversité. Ils créent pour cela un jeu de données à l'aide de personnes issues de communautés non américaines ou européennes, auxquelles ils demandent de récupérer en ligne des images correspondant à des catégories spécifiques. Ils mixent ces données avec d'autres données récupérées par un processus de récupération automatique de données en ligne. Ils observent sur cette base de données diversifiée des erreurs de classification de la part de systèmes de classification usuels, et constatent que les modèles de classification semblent se concentrer sur le visage des personnes sur les images plutôt que les attributs concernés par la catégorie (comme les habits pour les catégories "groom" et "bridegroom"). Ils réalisent enfin une comparaison de performances de modèles entraînés sur Imagenet et Open Images entre les données issues des jeux d'évaluation d'Imagenet et Open Images, et celles issues de la base de données diversifiée. Pour faire cette comparaison, ils utilisent des graphes de log-likelihood pour trois classes sélectionnées pour la forte disparité de résultats obtenus entre données usuelles et diversifiées (voir la figure 19). Ils montrent que les performances des modèles sont moins bonnes sur les données diversifiées, et concluent sur la nécessité de créer des bases de données plus diversifiées et inclusives.

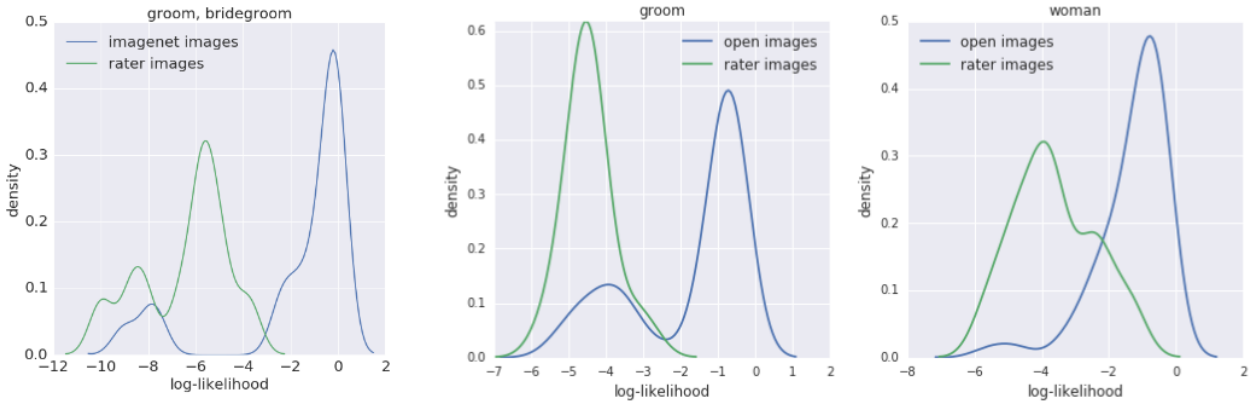
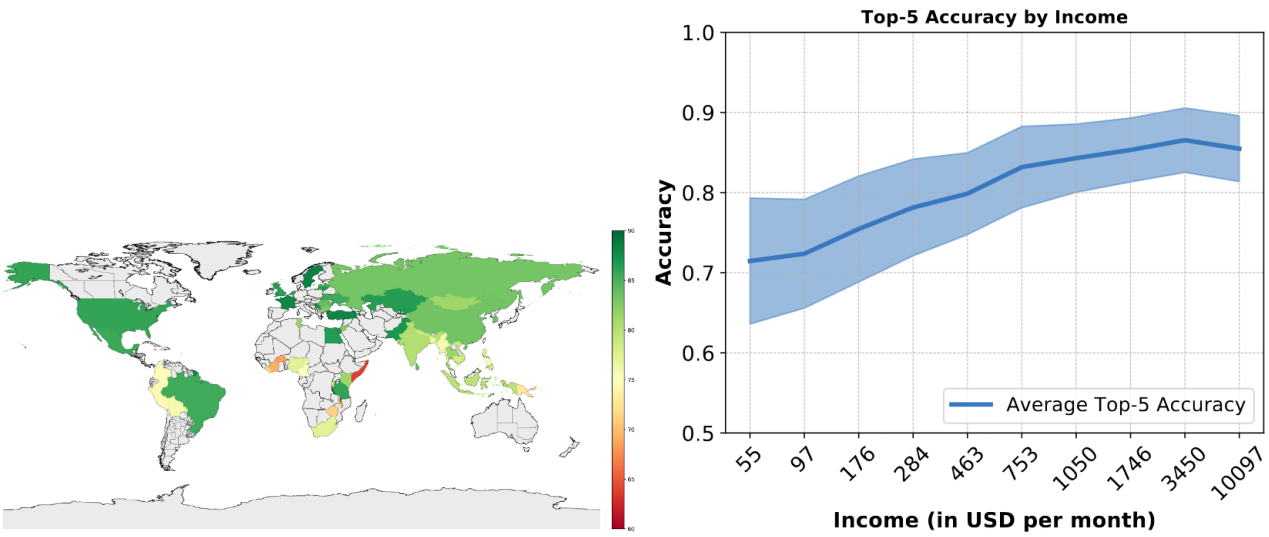


Figure 19: Graphes de densité de log-likelihood pour les catégories "groom, bridegroom", "groom" et "woman" comparant entre des bases de données génériques (Open images, Imagenet) et un jeu de donnée diversifié (rater). Crédit Shreya Shankar[147].

DeVries et al. (2019) ont poursuivi ces travaux en reprenant le même schéma opérationnel. Ils commencent également par tracer la source des bases de données populaire en vision par ordinateur, en rajoutant MS COCO dans la liste des bases de données manquant de diversité. Ils procèdent ensuite à une analyse des performances de plusieurs systèmes de classification d'images sur une base de données appelé Dollar Street Dataset. Cette base de donnée contient des images de la vie de tous les jours de cultures diverses, mettant en lumière la diversité des modes de vie à travers le monde, via l'initiative de Gapminder[1]. L' analyse de performance se base sur l'évaluation par des experts des prédictions des systèmes. Les auteurs de ces travaux proposent deux manières de présenter les performances des systèmes sélectionnés sur ces données. La première est une carte du monde présentant les performances par pays. La seconde est un graphe présentant les performances en fonction des revenus des foyers dont sont issues les données images. Ces deux représentations visibles en figure 20 permettent d'illustrer des phénomènes particuliers : ils illustrent une corrélation entre performance des systèmes et origine géographique d'une donnée en premier lieu, une corrélation entre performance des systèmes et revenus des foyers dont sont issues les données images en second lieu. Ils proposent deux justifications principales à ces corrélations : le manque de représentativité des zones géographiques hors-occident dans les bases de données d'images les plus populaires, et l'utilisation de la langue anglaise pour la collecte de données.

[1] https://www.gapminder.org/dollar-street



(a) **(b)**

Figure 20: Ces figures présentent les résultats de DeVries et al. (2019), qui ont testé plusieurs systèmes de reconnaissance d'objets dur les données du Dollar Street Dataset[65]. (a) Carte présentant la précision moyenne de six systèmes de reconnaissance d'objets par pays. (b) précision moyenne de six systèmes de reconnaissance d'objets en fonction du revenu associés aux données (en US$ par mois). Crédit Terrance Devries[49].

S'appuyant également sur les observations de Shankar et al. (2017), Atwood et al. (2020) organisent une compétition de classification sur des images inclusives, et définissent ces images comme "des images issues d'emplacements et de contextes culturels qui sont mal représentés ou absents des bases de données utilisées en entraînement". Cette compétition a pour but d'encourager au développement de modèles capables de mieux généraliser sur des données inclusives, et considère comme acquis le fait qu'il y ait un glissement géographique entre données d'entraînement et données de test. Dans la configuration proposée, c'est au modèle de mitiger les conséquences du biais géographique assumé dans la compétition, en obtenant les meilleures performances possible sur le jeu de données de test.

Le benchmark WILDS[92] étend ces efforts, en nettoyant et regroupant plusieurs bases de données contenant des glissements sous un unique benchmark pour évaluer la capacité des modèles à être déployés "sur le terrain". Parmi ces bases de données, plusieurs comportent des biais géographiques, et permettent d'illustrer les problématiques liées à l'entraînement de modèles sur des données présentant un manque de diversité. Le benchmark défini précisément les stratégies d'entraînement et de test pour chaque base de donnée, et propose notamment deux stratégies pour l'évaluation de l'impact d'un glissement : la différence de performance d'un unique modèle entre les deux sous-ensemble de données présentant le glissement, ou la différence de performance entre deux modèles, chacun entraîné sur un des sous-ensemble de données présentant le glissement. Ces deux stratégies sont dépendantes de facteurs extérieurs, comme la disponibilité de données, la capacité computationelle disponible. La deuxième approche permet notamment de prendre en compte la différence de difficulté entre les deux sous-ensembles de données présentant le glissement, effaçant de fait un biais inhérent à la première approche.

Dans leur outil REVISE, Wang et al. (2022) proposent des analyses des bases de données visuelles pour y traquer les biais et offrir une analyse la plus compréhensive possible de la source des biais dans les images et des stratégies possible pour la mitigation de ces derniers. Ils proposent notamment une analyse géographique des données, en étudiant leur distribution géographique, celle des catégories et objets dans les images, mais aussi la couleur de peau des personnes dans les images. Ils étudient aussi la langue associée aux données en assumant une source touristique si cette dernière n'était pas liée à la localisation de la donnée,la distribution des données en fonction d'un indice de niveau de vie, et la distribution géographique des conditions météo dans les images. Toutes ces analyses permettent de

mettre en évidence des biais présents dans les bases de données qu'ils utilisent, qui sont de nature et de source différentes, et qui chacun constituent une composante du biais géographique dans les bases de données visuelles.

Face au biais géographique, plusieurs approches sont donc proposées pour améliorer l'inclusivité des systèmes de vision par ordinateur, illustrées en figure 21. La première approche se concentre sur les données, et promeut une meilleure représentativité des différents contextes culturels et emplacements dans les bases de données. Une deuxième approche concerne les modèles, et promeut des modèles ayant des meilleures capacités d'adaptation et de généralisation, en supposant qu'il y aura toujours un biais dans les données. Une troisième approche promeut l'analyse de performance des modèles déployés aujourd'hui, pour encourager une meilleur prise en compte de la problématique de l'inclusivité dans les modèles. C'est cette troisième approche, jusqu'ici peu empruntée, que nous abordons avec ce protocole.

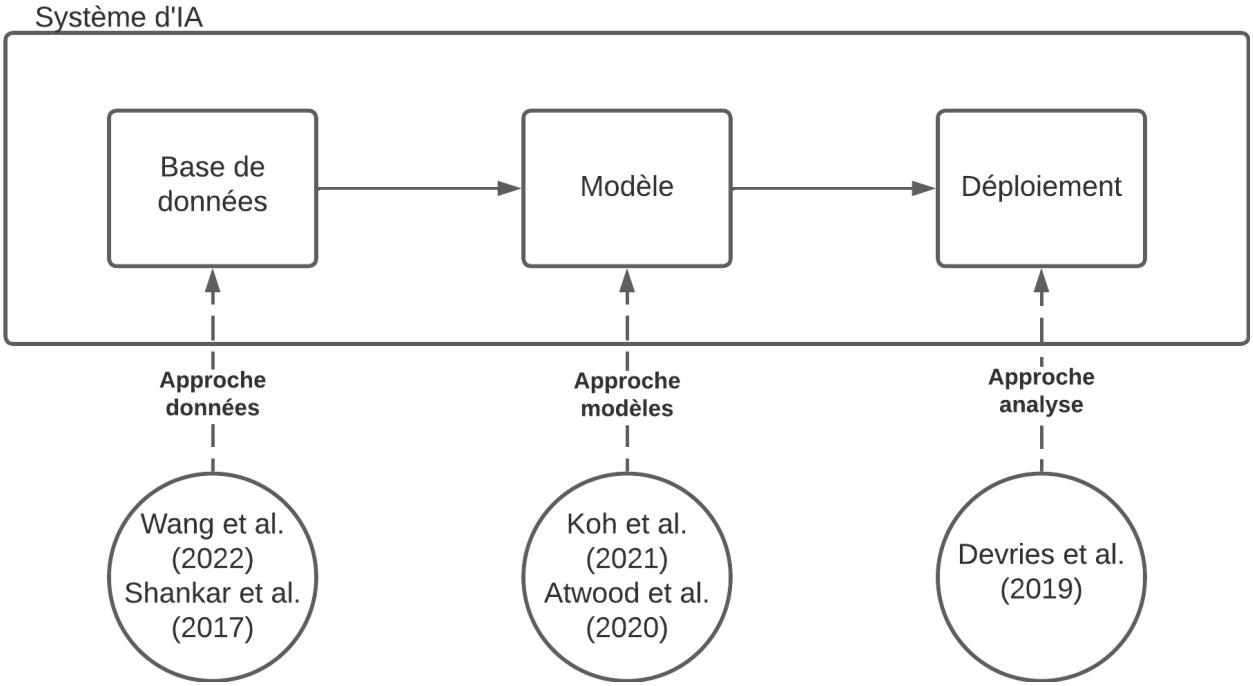


Figure 21: Illustration des différentes approches pour la mitigation du biais géographique. Les systèmes d'IA sont découpés en leur trois composantes principales : les bases de données, le modèle, et le déploiement.

3.1.2 Manques des travaux précédents

Les initiatives précédentes, si elles ouvrent la voie, manquent pourtant d'une certaine rigueur dans leur approche pour caractériser le biais géographique. Les analyses proposées par Shankar et al. (2017) et DeVries et al. (2019) ne se reposent pas sur des approches classiques de caractérisation des biais, et leurs outils et protocoles sont présentés sans comparaison à l'existant. De fait, ils ne font pas d'analyse critique de ces derniers, des pré-requis pour les employer, de leur capacité à être décliné dans d'autre contextes ou de leur passage à l'échelle. L'outil REVISE[165] quant à lui se limite aux bases de données visuelles, et n'explore donc pas les biais inhérents aux modèles ou au déploiement de ces derniers.

L'utilisation des graphes de densité de log-likelihood par pays pour des classes spécifiques et la représentation des valeurs de log-likelihood par pays pour ces classes dans les travaux de Shankar et al. (2017) (figure 19) ne permet pas de passer à l'échelle. Ces représentations nécessitent la sélection d'une ou de plusieurs classes en particulier, puis l'observation visuelle de ces résultats. Le fait de devoir sélectionner des classes à observer introduit un biais de sélection (on peut par exemple sélectionner uniquement des classes pour lesquelles la théorie est vérifiée ; alors que d'autres classes

pourraient aller à l'encontre de cette théorie), tandis qu'une vérification visuelle introduit des complications d'ordre temporel (dans le cas de bases de données avec de nombreuses classes) et d'objectivité (un chercheur ou expert peut être biaisé vers un résultat en particulier et voir dans les données une réalisation de ce biais sans pour autant que les données reflètent le résultat en question). De plus, ces représentations nécessitent un accès à la dernière couche du modèle, ce qui est une condition parfois difficile à atteindre lors d'utilisation de systèmes de vision.

L'utilisation par DeVries et al. (2019) d'une carte du monde présentant les performances des modèles par pays et d'un graphe présentant les performances en fonction des revenus associés aux données images présente aussi des points faibles. Ces deux représentations sont soumises à analyses visuelles, et donc à des biais d'objectivité lors de l'analyse. Ensuite, l'utilisation d'une carte importe un certain nombre de biais, comme le choix du centre de la carte (Europe, Asie, Amérique ?), le type de projection, la taille des pays (et donc l'impression d'une importance plus grande des pays les plus vastes). Pour autant, ces outils semblent plus adaptés à la caractérisation des biais géographiques que les précédents. L'outil REVISE[165] fait un pas vers la conception d'outil adaptés à la caractérisation du biais géographique, mais se limite aux biais dans les bases de données. En outre, les multiples approches présentées dans l'outil peuvent faire face aux même critiques que précédemment sur l'utilisation de cartes ou de graphes. On peut surtout pointer dans tous ces travaux comme faiblesse l'absence de formalisation de la définition du glissement de contexte géographique. Koh et al. (2021) font état de diverses configurations possibles pour les mesures d'écart de performance lorsqu'on témoigne d'un glissement, et formalisent ces différentes configurations. Dans leurs analyses, Shankar et al. (2017) et DeVries et al. (2019) mettent en avant les écarts de performance des modèles utilisés dans les différentes zones géographiques considérées, mais ne précisent pas comment ces mesures témoignent d'un biais géographique.

Enfin, les modèles utilisés dans les initiatives précédentes le sont sans justifier d'une politique particulière pour la sélection d'algorithmes, de modèles, d'hyperparamètres lors d'entraînement. Gulrajani and Lopez-Paz (2021) soulignent l'importance de décrire précisément les stratégies relatives à ces choix, pour éviter les irrégularités dans les performances et les résultats communiqués.

3.1.3 Caractérisation : Source, Type, Impact

La caractérisation d'un biais correspond à l'identification des caractéristiques d'un biais. Nous proposons dans cette thèse de définir trois caractéristiques pour un biais : la source du biais, la manière dont il est introduit dans le système, et l'impact qu'il a sur le système. La définition de la source du biais relève de l'observation du biais, et de l'interprétation qu'en font les auteurs de l'observation. La liste établie en partie 2.2.2 permet de faciliter cette identification, mais nous proposons ici des outils plus complets pour mieux comprendre comment la source d'un biais est ancrée dans un ensemble de considérations qui dépendent en partie de l'observateur. Les cadres introduits en partie 2.2.1.1 permettent d'identifier la manière dont le biais s'introduit dans le système, et nous proposons dans la suite notre propre cadre. La mesure de l'impact d'un biais est introduite en section 2.2.3 pour l'évaluation de la dernière caractéristique d'un biais. Le protocole de caractérisation de biais proposé dans cette thèse se concentre sur le biais géographique, dont la définition est apportée en section 2.2.6. Il est nommé après les trois caractéristiques sur lesquelles il se repose : **S**ource, **T**ype, **I**mpact.

L'identification de ces caractéristiques pour le biais géographique permettra de formaliser clairement le glissement à l'oeuvre dans un système et de proposer une classification claire de ce dernier ainsi qu'une méthode éprouvée pour témoigner de son impact. Une telle analyse rigoureuse ouvrira la voie à de futurs travaux sur ce biais en capitalisant sur les observations et implémentations réalisées durant cette thèse.

3.2 Le protocole STI

On présente dans cette section le protocole STI avancé pour la caractérisation de biais. On commence par présenter les trois étapes dans une vue globale en section 3.2.1 qui permettent de déterminer les

trois caractéristiques d'un biais : la source, le type, et l'impact. La section 3.2.2 présente les arbres d'identification de source qui devront être implémentés pour chaque biais. Dans la section 3.2.3, nous reprenons les cadres d'identification du type de biais des travaux de Suresh and Guttag (2021) et proposons une évolution mineure de leur cadre. En section 3.2.4, nous illustrons les différentes définitions de l'impact d'un biais, les multiples manières de mesurer cet impact et comment choisir l'un d'eux. Enfin, en section 3.2.5, nous revenons sur des considérations particulières, comme les notions d'isolement d'un biais, de glissement ou de report de résultats.

3.2.1 Vue globale du protocole STI

La figure 22 illustre le protocole STI, qui se découpe en trois étapes :

- L'identification de la source d'un biais correspond à la recherche par l'observateur des causes du biais. Cette identification peut être réalisée à différentes échelles d'abstraction de concepts, et n'est pas indépendante des choix réalisés par les auteurs de l'analyse. Nous proposons dans la suite pour cette identification l'utilisation d'un arbre qui illustre ces différentes échelles d'abstraction et la nature subjective du choix d'une source.
- L'identification de la manière dont le biais est introduit dans le système est illustré par les différent cadres présentés en section 2.2.1.1, et correspond au type du biais. De multiples approches sont possibles pour déterminer le type de biais, et nous proposons dans la suite notre propre cadre en figure 24 pour l'identification du type de biais.
- L'identification de l'impact du biais sur le système peut concerner les performances du modèle, des sorties injustes pour certaines catégories de population, ou encore des a-priori injustifiés pour des classes ou groupes particuliers dans les données. Il existe différentes manières de mesurer l'impact d'un biais, et nous présentons une approche compréhensive de cette mesure d'impact.

Ces étapes n'ont pas d'ordre particulier ; on peut réaliser la présence d'un biais par son impact lors du déploiement, ou au contraire intuiter la présence de celui-ci et suivre le protocole pour le déterminer à priori. Dans tous les cas, une compréhension des implications des sources jusqu'à l'impact est nécessaire pour une mitigation efficace du biais.

Pour mitiger un biais, on peut vouloir intervenir sur l'une de ces trois étapes. On peut intervenir à la source du biais en modifiant les données ou les sources, empêcher le biais d'être introduit dans le système en modifiant les étapes de conception de ce dernier, ou empêcher le biais d'avoir un impact en limitant le déploiement d'un système ou en modifiant les sorties de ce dernier.

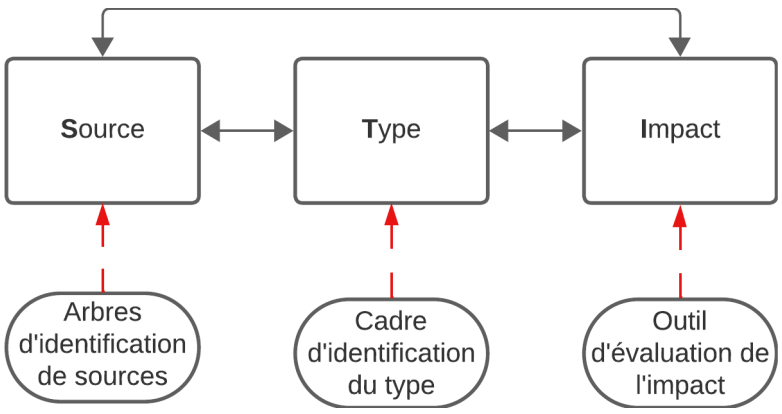


Figure 22: Protocole STI (Source, Type, Impact) pour la caractérisation d'un biais. Ces trois caractéristiques sont dépendantes les unes des autres, et peuvent être déterminées dans n'importe quel ordre. Nous proposons des outils pour identifier chacune des caractéristiques d'un biais : des arbres de recherche pour identifier la source, un cadre pour l'identification du type de biais, et des outils pour l'évaluation de l'impact d'un biais.

Nous décrivons dans la section suivante les trois étapes du protocole STI et illustrons ces dernières par des outils adaptés à la caractérisation d'un biais.

3.2.2 Première caractéristique : Source du biais

Tout biais prend sa source dans le contexte culturel et socio-technique mondial. Mais on peut déterminer plus finement la source d'un biais afin de comprendre ce qui l'a fait apparaître dans le système d'IA étudié en tentant de remonter les étapes entre l'observation du biais et la source de ce dernier. Lors de l'étude et la caractérisation d'un biais, identifier et fixer une source est un moyen de se concentrer sur un aspect particulier, tout en admettant de manière compréhensive les limites de l'approche. Il existe différents niveaux d'abstraction pour la source d'un biais, et bien qu'il soit impossible d'identifier précisément les liens de causalité entre un biais et ses sources du fait de la complexité du monde réel, on peut tenter de tracer un ensemble de sources possibles. Benjamin (2021) proposent ainsi une cartographie des relations sociotechniques en vision par ordinateur, utilisant cinq niveaux de structure sociale pour les différents acteurs et groupes représentés sur la cartographie. Si nous ne prétendons pas produire de cartographie aussi complexe, cette dernière est utile pour comprendre la complexité reliant un biais et la source de ce dernier. Nous illustrons ce fait avec la figure 23, en proposant un arbre d'identification des sources pour le phénomène de glissement de données. Cet arbre est construit en s'appuyant sur les travaux précédents[92; 145; 174; 56; 69; 114; 122], et se lit en interprétant une flèche d'un phénomène à un autre comme "à cause de" ou "prend sa source dans". Ainsi, différentes échelles techniques et sociales sont représentées dans cet arbre, reflétant les multiplicités des cas et des approches possibles pour prendre en compte ce biais.

La lecture d'un tel arbre ne doit pas permettre d'argumenter contre la présence d'un biais qui ne serait pas encore identifié ou documenté dans un système particulier ; il permet seulement d'obtenir une vue plus compréhensive du phénomène observé, en identifiant les multiples causes et niveaux où ce phénomène est déjà documenté. Aussi, l'absence ou la présence de lien entre deux artefacts dans un tel arbre, ou leur mise sur un même niveau, est la conséquence d'une lecture particulière des auteurs de l'arbre en question d'une revue de littérature sur un biais spécifique. Un protocole plus sérieux et défini, ou l'utilisation de revues automatiques par des algorithmes identifiant les concepts et artefacts liés à un phénomène de biais dans des revues, pourraient permettre une construction moins orientée de tels arbres de sources. De tels travaux constituent de futures investigations du protocole STI.

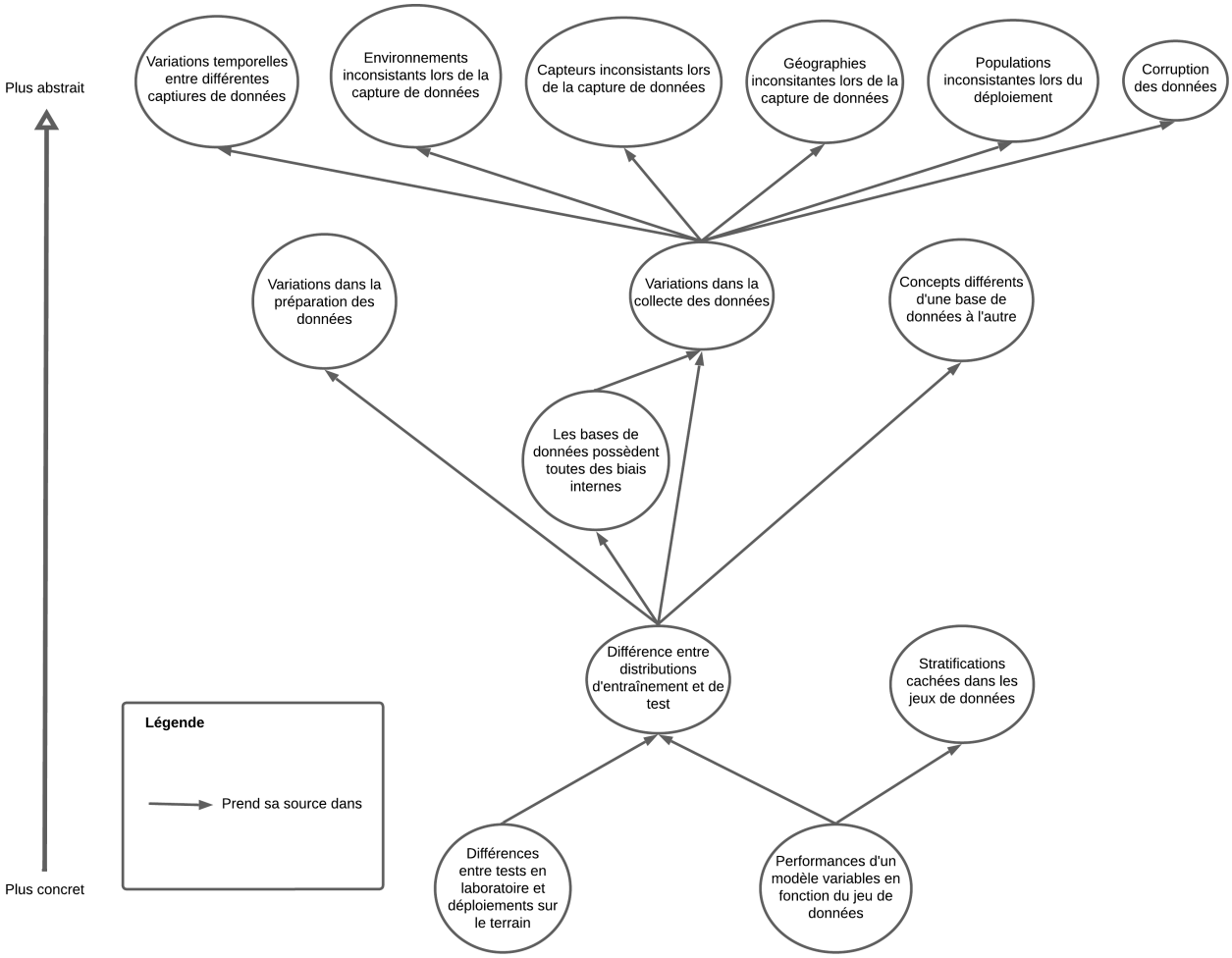


Figure 23: Illustration d'un arbre d'identification du biais géographique, utilisant un ensemble de travaux de recherche[92; 145; 174; 56; 69; 114; 122]. Cet arbre n'est pas exhaustif, mais illustre comment les sources s'imbriquent à différentes échelles et émanent toutes du contexte socio-culturel.

Ce procédé apporte plusieurs avantages :

- L'identification de différentes sources possibles permet une étude ultérieure en isolant une des sources par rapport aux autres identifiées et d'en déterminer l'importance dans le biais identifié.
- L'identification de ces sources peut aussi servir d'inspiration pour cibler les efforts à fournir pour pallier le biais identifié.
- L'identification de liens de causalité entre sources permet une compréhension des interactions entre les différents choix réalisés aux différentes étapes de développement du système d'IA où le biais est identifié.

Un tel arbre d'identification peut être utilisé de différente manière :

- De haut en bas : en partant des concepts les plus abstraits et en allant vers les concepts les plus concrets, on peut utiliser cet arbre pour chercher des biais géographiques dans un système d'IA. La lecture de cet arbre peut aider un acteur à comprendre quel biais peut intervenir dans la construction du système d'IA qu'il utilise, et où et comment ce biais prend source. Ainsi, tout acteur peut grâce à cet arbre identifier des tests à réaliser pour vérifier si certaines sources potentielles de biais impactent leur systèmes.
- De bas en haut : en partant des concepts les plus concrets et en allant vers les concepts les plus abstraits, on peut comprendre comment une situation rencontrée dans un système peut prendre sa source à différents niveaux du systèmes et à différentes échelles de compréhension et

d'action. La lecture de cet arbre peut aider un acteur à comprendre le comportement du système qu'il utilise, et identifier des points bloquants ou sensible sur lesquels concentrer l'attention.

- Horizontalement : en identifiant à quel niveau on est capable d'agir sur un système, cet arbre peut être utilisé pour identifier les sources potentielles de biais à une échelle identifiée et d'agir en conséquence. Ainsi, un acteur pourra utiliser cet arbre pour identifier les sources de biais auquel son système est soumis et utiliser des stratégies pour mitiger ces biais au niveau auquel il peut agir.

Cet arbre propose donc différents niveaux de lecture et autant de moyens d'être utilisés pour des acteurs ou des utilisateurs de systèmes d'IA. La construction d'un tel arbre requiert une revue de littérature autour du biais identifié, ainsi qu'un découpage en différentes échelles des sources de ces biais. Nous réalisons une telle implémentation pour le biais géographique en section 4.1.1

3.2.3 Deuxième caractéristique : Type du biais

Le cadre proposé pour l'identification du type de biais divise la chaîne de traitement de conception et de déploiement d'un système d'AM en trois sous-parties : la construction des bases de données, la construction des modèles et le déploiement du modèle. Cette division reflète une réalité dans l'utilisation des systèmes d'IA : les données, modèles, et les contextes de déploiement sont interchangeables facilement pour les différents acteurs du domaine. Ce cadre, présenté en figure 24 souligne donc cette division, et permet de mieux isoler les biais relatifs à chacun des éléments de cette division.

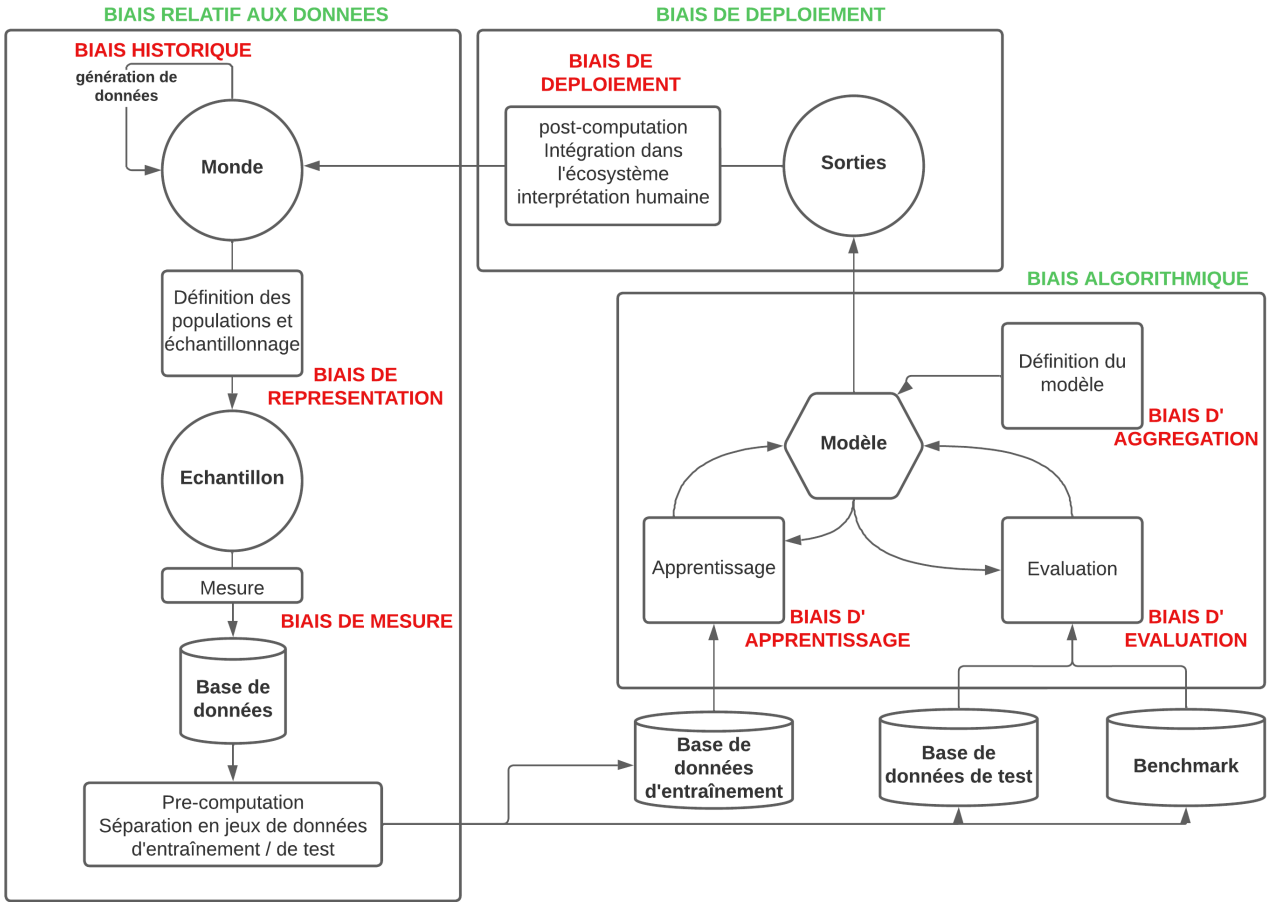


Figure 24: Le cadre proposé pour l'identification du type de biais, qui permet de comprendre la manière dont le biais survient dans un système d'IA. Ce cadre étend celui de Suresh and Guttag (2021), avec deux ajouts : une échelle de lecture supplémentaire illustrant les éléments constituant les systèmes d'IA et une boucle illustrant l'aspect cyclique des biais qui se renforcent au fur et à mesure que les systèmes sont déployés.

Ce cadre n'est pas spécifique au biais géographique, mais illustre les différentes manière dont ce dernier s'introduit dans un système. L'utilisation de différentes échelles permet de procéder à un granularisation du type de biais. L'aspect cyclique de ce cadre souligne le fait que les systèmes participent aujourd'hui à la construction du monde et à la reproduction des biais historiques, engendrant un effet de renforcement de certains biais.

Il peut être utilisé par différents acteurs (concepteurs, développeurs, utilisateurs) d'un système d'IA pour les aider à comprendre comment le biais s'introduit dans le système.

La lecture de ce cadre peut répondre à de nombreux besoins :

- Un acteur faisant face à un biais dans son système peut utiliser ce cadre pour tenter d'identifier comment le biais s'y est introduit, en interrogeant l'une des trois étapes de la conception du système en premier lieu, et les biais spécifiques à l'étape identifiée en second lieu.
- Un acteur désireux d'avoir le système le moins biaisé possible pourra utiliser ce cadre pour réaliser des tests aux différents niveaux de granularité des types de biais pour s'assurer des résultats de ce dernier.
- Un acteur agissant à une étape spécifique du développement du système pourra utiliser ce cadre pour identifier quelles sont les risques encourus par le système sur lequel il travaille et mettre en oeuvre des stratégies pour pallier ces risques.
- Un acteur souhaitant mitiger un biais pourra utiliser ce cadre pour identifier des stratégies potentielles pour la mitigation du biais identifié.

Ce cadre n'est pas exhaustif, puisqu'on pourrait descendre à des niveaux de granularité plus fin pour la description des type de biais et des étapes de conception et de développement des systèmes. Néanmoins, il permet une identification rapide et claire des types de biais et donc de mieux comprendre comment un biais peut survenir dans un système.

3.2.4 Troisième caractéristique : Impact du biais

On a vu en section 2.2.3 que les biais peuvent avoir des impacts directs sur la performance du modèle, ou des impacts indirects qui heurtent des groupes ou populations particulières sans que cela ne se reflète sur la métrique de performance ou sur les tests de performance réalisés pour le système. Nous utiliserons ici les termes impact direct et impact indirect pour décrire ces deux types d'impact. Nous abordons dans cette partie l'évaluation de ces deux types d'impacts pour un biais identifié, qui peut être réalisée de nombreuses manières différentes.

3.2.4.1 Dans le cas où l'impact du biais est direct.

La revue compréhensive des types d'impact en section 2.2.2 et de la manière d'évaluer l'impact d'un biais en section 2.2.3 permet de dresser un arbre de décision pour les configurations à choisir pour l'évaluation d'impact en fonction du type de biais. Cet arbre de décision est présenté en figure 25.

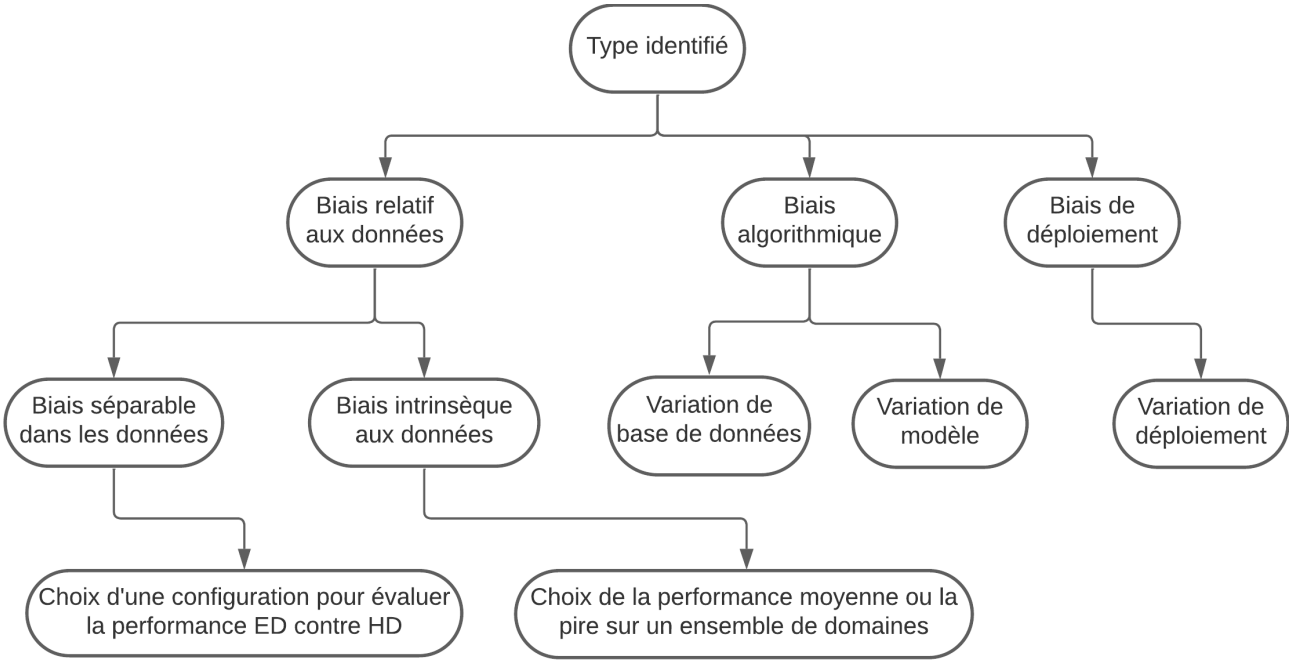


Figure 25: arbre illustrant les différentes configurations possibles pour l'évaluation de l'impact d'un biais.

On décrit ci-dessous les différentes choix et configurations représentées sur les feuilles de cet arbre :

- Dans le cas de biais relatifs aux données séparables entre deux jeux de données ED et HD, on pourra choisir une des configurations illustrées en figure 15. Ces deux configurations proposent une comparaison des performances du système sur les données ED et HD, mais différentes définitions existent pour ces performances. Le choix d'une des deux configurations dépend de facteurs tel que le volume de données disponibles et la capacité computationnelle disponible.
- Dans le cas de biais relatifs aux données et intrinsèques à celles-ci, on pourra choisir entre deux méthode d'évaluation de l'impact, la performance moyenne ou la pire performance sur un ensemble de jeux de données séparés par une variation du biais identifié.
- Dans le cas d'un biais algorithmique, on pourra évaluer l'impact du biais en comparant les performances de différents modèles en faisant varier parmi ces derniers l'apprentissage, l'évaluation ou la définition du modèle. L'utilisation de multiples bases de données permettra aussi de caractériser le biais par les variations de performances sur les différentes données qui lui seront présentées.
- Dans le cas d'un biais de déploiement, on pourra évaluer l'impact du biais en comparant les performances d'un modèle dans différents contextes de déploiement intégrant le biais identifié.

Si le type de biais n'est pas identifié, il n'est pas possible de déterminer à l'avance quel type de test peut permettre de mettre à jour un biais. Mais dans le cas où le biais est découvert à travers son impact, le type de test réalisé pour découvrir le biais permet de lier directement l'impact au type de biais.

3.2.4.2 Dans le cas où l'impact du biais est indirect.

Le biais se mesurera de manière générale en étendant le processus d'évaluation du système ou en procédant à une évaluation complémentaire, via une modification ou un ajout dans les données utilisées pour l'évaluation ou l'ajout de métriques de performances supplémentaires. Cette mesure est illustrée en section 2.2.3.2. Si le biais n'a pas d'impact sur la performance du système, il faut construire une métrique adaptée pour la mesure du biais. Cette métrique sera donc spécifique à chaque

biais identifié et l'analyse de la performance du système sur cette métrique permettra de témoigner de l'impact du biais.

Il existe déjà une littérature très étendue en justice algorithmique autour de métriques pour la mesure de biais et le déploiement de ces dernières. Nous nous contentons de souligner ici que le protocole STI s'applique tout autant dans ce cas que dans le cas où l'impact du biais est direct.

3.2.5 Considérations supplémentaires

Nous revenons dans cette partie sur certains détails du protocole dont l'exploration mérite un intérêt particulier.

3.2.5.1 Données et isolation du biais

Dans le cas de biais relatif aux données, la section 3.2.4 propose de séparer les données selon la présence ou non de biais (dans le cas séparable) ou de faire varier ce dernier dans un ensemble de jeux de données. Ces deux stratégies se basent sur la même hypothèse : qu'il n'y ait pas d'autres facteurs qui permettent d'expliquer les variations de performance que la différence provoquée par la variation du biais ou de sa présence. On considère alors que le biais est isolé. Cette hypothèse est dans les cas de données complexes difficile à vérifier. De nombreux facteurs extérieurs à l'expérience et inhérents aux protocoles en AM sont toujours présents, et peuvent influer sur les variations de performance d'une condition d'expérience à une autre. Dans le cas où des travaux se basent sur cette hypothèse et où cette hypothèse n'est pas vérifiée, il peut alors être compliqué d'identifier l'impact recherché ou de réussir à cerner la source ou le type de biais. Une attention particulière doit donc être donnée à cette hypothèse.

3.2.5.2 Caractérisation d'un glissement

Il est nécessaire de reprendre la définition de ce qu'est un glissement de domaine à la suite de la définition d'un biais à l'aide des trois caractéristiques que nous proposons. Un glissement de domaine survient quand les données d'entraînement et de test appartiennent à des domaines différents, la notion de domaine étant vague et se rattachant à la nature et le contexte socio-technique et culturel des données. A la lumière de notre définition du biais, un glissement de domaines se trouve donc être un biais de déploiement s'il est découvert lors du déploiement d'un système, ou un biais relatif aux données s'il est présent lors de phases d'entraînement ou de test d'un modèle. L'identification de la source d'un glissement pourra se faire grâce à la construction d'un arbre d'identifications de sources spécifique au glissement mentionné, et son impact sera mesuré grâce aux métriques identifiées par son type.

3.2.5.3 Report de résultats

Si la manière de reporter des résultats est souvent considérée comme neutre, il est nécessaire d'identifier que certains biais de déploiement sont dus à la manière dont les résultats sont apportés aux utilisateurs. Les métriques et classements fournis sont des abstractions de réalités plus complexes, et la construction de ces derniers comporte des choix qui peuvent être nocifs si non pris en compte, comme le contexte dans lequel ces métriques sont construites. Les cartes et graphes possèdent de même des biais intrinsèques à leur utilisation. Un système qui se dit inclusif doit veiller à ce que la manière de fournir les résultats aux utilisateurs soit adaptée à ces derniers, et ne reflète pas une construction issue du contexte socio-technique des développeurs de l'outil.

3.3 Discussion

Cette section offre une discussion autour du protocole STI et des complications qui sont soulevées par les définitions et outils développés.

3.3.1 Définitions et outils.

La décomposition du biais en trois caractéristiques proposée en section 3.2 n'a pas la prétention d'être exhaustive, et de futurs travaux pourraient proposer des décompositions plus poussées. Cette décomposition est cependant jugée utile pour mieux permettre à la communauté scientifique de cerner les tenants et aboutissants des travaux sur la mitigation des biais et la prise en compte des différents aspects et inclinaisons dans les données, les modèles et les déploiements. De même, les outils proposés pour l'identification de chacune des caractéristiques se reposent sur des analyses détaillées des précédents travaux, mais pourront sûrement être améliorés dans des itérations futures. Ces travaux ont pour but de proposer un protocole qui gagnera à être exploité et amélioré par la communauté scientifique.

3.3.2 Apprentissage profond et inclusivité.

L'apprentissage profond, par nature, est un apprentissage statistique, qui se repose sur l'identification de représentations communes dans des ensembles de données. Un système inclusif, quant à lui, doit pouvoir faire la différence entre différents contextes et produire des résultats prenant en compte ces différences. De fait, l'AP se repose sur l'apprentissage de généralités tandis que l'inclusivité défait ces généralités. Ainsi, on peut établir qu'il pourrait y avoir un paradoxe entre AP et inclusivité. Un système totalement inclusif serait un système qui sait répondre parfaitement à chaque situation et chaque contexte, et possède donc l'information qui lie chaque entrée possible à sa sortie possible. C'est un oracle qui, dans son contexte de déploiement, connaît déjà la sortie à attribuer à chaque entrée, et est donc assimilable à un arbre de décision fini, bien que pouvant atteindre de très grandes dimensions. L'utilisation des réseaux de neurones est donc en réalité contre-productive lorsque l'on tente de construire un système inclusif. Les systèmes à base de réseaux de neurones et se voulant inclusifs doivent donc faire un compromis entre capacité à extraire des généralités et capacité à extraire des spécificités ; entre performance et inclusivité. Ce compromis pourra être exploré plus en avant dans de futurs travaux.

Il est à noter que de nouveaux paradigmes voient constamment le jour en AM, et que le compromis à réaliser dans un système respectant un paradigme n'est pas le même que dans un autre système basé sur une autre paradigme. Les tendances actuelles en AM ouvrent la voie à des modèles capables d'une grande généralisation et les modèles fondateurs sont capables de réaliser de nouvelles tâches sans réentraînement. Ces évolutions sont prometteuses pour l'inclusivité des systèmes. On peut néanmoins toujours identifier des artefacts témoignant d'un manque d'inclusivité dans un système, du fait que l'inclusivité est un fait social et politique avant d'être un fait technique. La caractérisation de cette inclusivité permet alors de rendre compte de cet état de fait et de souligner les efforts à réaliser, sur les divers plans techniques, sociaux et politiques.

3.3.3 Perspectives.

Ce nouveau protocole est aussi une contribution au domaine de la justice algorithmique[163], qui se concentre sur l'identification et la mitigation de biais ayant des conséquences sociales importantes. L'utilisation du protocole pourrait permettre de mieux comprendre les biais, et donc apporter aussi au domaine de l'explication de modèles. L'implémentation de ce protocole pour un biais en particulier permet aussi aux travaux futurs de déployer plus facilement le protocole, en se reposant ou en construisant à partir des arbres d'identifications de sources, de types et d'impacts déjà implémentés. C'est donc un protocole qui se veut évolutif, permettant la collaboration et simplifiant la réutilisation et l'apport de nouvelles contributions.

3.4 Conclusion

Le protocole STI proposée dans ces travaux répond à un duo de besoins : la décomposition d'un biais en ces caractéristiques pour mieux le comprendre, et la nécessité d'avoir des outils pour déterminer

ces caractéristiques. Pour être opérationnel, le protocole doit pouvoir être implémenté à l'aide d'une étude rigoureuse du biais, et pourra s'inspirer de précédentes implémentation pour des biais différents. Ce protocole permet de mettre en évidence certaines des lacunes des précédentes approches proposant la mitigation de l'impact des biais en AP, et propose un nouveau cadre permettant la construction de savoirs itératifs et évolutifs. Il est donc particulièrement adaptée à l'AP, qui a pour caractéristique d'évoluer rapidement et de se transformer radicalement au rythme des innovations algorithmiques et technologiques.

Chapter 4

Implémentation et validation sur données synthétiques

Chapter Summary

Résumé du chapitre

Dans ce chapitre, nous présentons l'implémentation du protocole STI, et la validation de la méthodologie associée. La section 4.1 présente l'implémentation du protocole pour le biais géographique et détaille les trois étapes de ce protocole. Un arbre d'identification est construit à partir d'une revue des productions académiques concernant le biais géographique. Cet arbre n'a pas pour prétention d'être exhaustif, mais de présenter une vue d'ensemble des connaissances relatives aux sources du biais géographique. On présente ensuite comment le biais géographique peut être de différents types, et comment l'impact du biais géographique peut intervenir sur les performances d'un modèle. La section 4.2 précise les bonnes pratiques pour la sélection de modèles ou l'entraînement de ces derniers. Ces bonnes pratiques sont cruciales pour la réplicabilité des résultats et le bon déroulement des expérimentations. En section 4.3, nous faisons le constat d'une absence de bases de données synthétiques à biais

géographique, et entreprenons de construire de telles bases de données en nous inspirant de précédentes variations de la base de données MNIST. Nous proposons ainsi GeoMNIST, une méthodologie qui permet l'introduction de biais géographique dans MNIST sous la forme de transformation liées à une donnée géographique pour chacune des données de la base. Nous proposons ainsi trois implémentations différentes de GeoMNIST, respectant des répartitions géographiques différentes : une répartition similaire à la répartition mondiale de population, une répartition uniforme, et une répartition similaire à celle des bases de données d'images publiques. Dans la section 4.4, nous expérimentons l'implémentation du protocole STI pour les bases de données GeoMNIST générées, en procédant à la caractérisation du biais géographique sans puis avec le protocole STI. Les résultats de ces expérimentations sont discutés en section 4.5, et nous apportons une conclusion sur les implémentations réalisées dans la section 4.6.

4.1 Implémentation du protocole STI

Le protocole STI se décline en trois étapes : identification des sources, identification du type, évaluation de l'impact. Cette section reprend ce découpage et présente comment les outils pour chaque étape sont implémentés et utilisés.

4.1.1 Arbre d'identification pour la source du biais géographique

Le biais géographique prend sa source dans le contexte culturel et socio-technique mondial ; l'accès au numérique, le développement des infrastructures, la présence de terminaux capables de générer des données, sont des facteurs qui jouent sur la disponibilité des données dans les différents pays. Pour déterminer les multiples sources du biais géographiques dans les systèmes d'IA, nous nous basons sur une revue de la littérature[49; 147; 92; 11; 174; 2; 51; 107; 145; 65; 157; 152; 35; 40; 88; 134; 164; 45; 66; 171]. L'arbre illustrant les différentes sources du biais géographique à partir de ces travaux est présenté en figure 26. Ce dernier n'est pas exhaustif, mais retrace l'ensemble des liens trouvés entre les artefacts à la source du biais géographique identifiés dans la littérature. L'absence de lien entre deux artefacts n'est donc pas une absence de causalité entre l'un et l'autre, mais l'absence de la documentation d'une telle causalité. L'absence d'une source particulière n'est également pas preuve de son absence d'impact sur le biais géographique, mais d'un manque de documentation de cette dernière.

La construction d'un tel arbre nécessite aussi de faire des choix dans le regroupement de plusieurs artefacts, puisque les contextes et normes d'identification de ces derniers diffèrent dans la littérature. Ces choix peuvent être discutés et contestés, et d'autres implémentations de cet arbre à partir d'une même littérature pourrait mener à un arbre sensiblement différent de celui proposé ici. Néanmoins, l'utilité de cet outil réside dans les multiples points de vue qu'il présente, et la possibilité d'adopter ces derniers pour la caractérisation d'un biais géographique :

- Les bases de données d'images étant majoritairement construites à l'aide d'images récupérées sur le web, les pays les plus représentés sur la toile ont une représentativité supérieure aux autres. La source des biais géographique se trouve donc de ce point de vue dans le manque d'inclusivité du web.

- Les chercheurs et acteurs dans le domaine de l'AM sont dans une grande majorité issus de pays du Nord ou de l'Asie de l'Est. Ces bases de données d'images sont donc conceptualisées et construites dans un contexte socio-technique particulier, excluant les pays du Sud. La source des biais géographiques se trouve donc de ce point de vue dans le manque d'inclusivité des acteurs en AM.

- Les outils et capteurs utilisés pour générer des données peuvent présenter des caractéristiques différentes selon les localités dans lesquelles ces outils sont utilisés. Ces particularités peuvent

produire des artefacts dans les données générées, liant localité et artefacts techniques. La source des biais géographiques se trouve donc de ce point du vue dans le contexte socio-technique.

Ces différents points de vue et d'autres sont représentés sur l'arbre en figure 26.

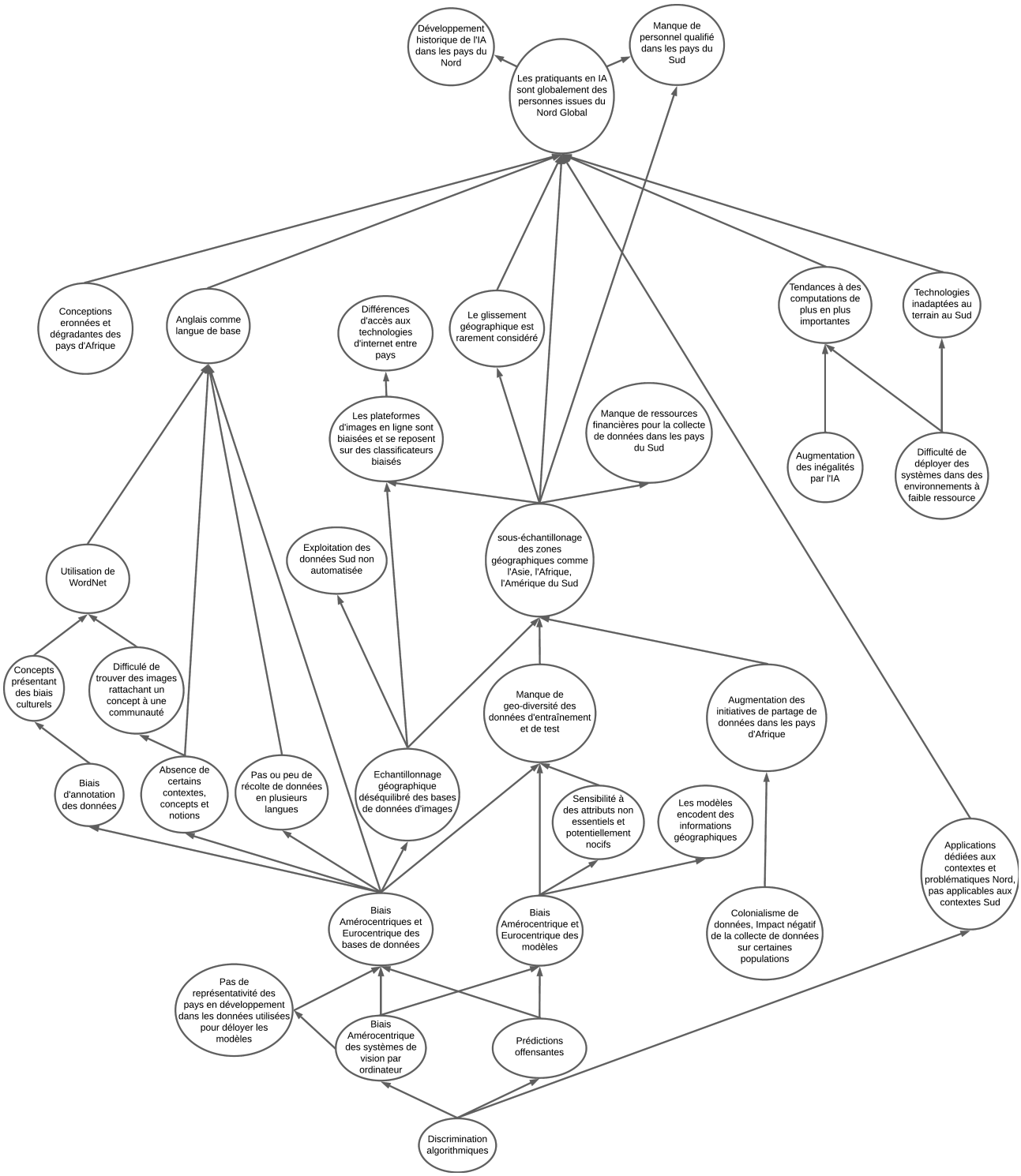


Figure 26: Arbre illustrant les différentes sources du biais géographique, et les relations entre ces dernières. La construction de cet arbre se repose sur la littérature liée au biais géographique dans les systèmes de vision par ordinateur[49; 147; 92; 11; 174; 2; 51; 107; 145; 65; 157; 152; 35; 40; 88; 134; 164; 45; 66; 171].

Sur cet arbre, on peut comprendre que le biais géographique prend sa source dans une imbrication complexe de raisons sociales, techniques, culturelles, économiques. Les différentes échelles illustrent les différentes conceptions que l'on peut se faire des causes d'un problème, et illustrent les multiples approches possibles pour la mitigation du biais géographique. Il est important de comprendre que ce

dernier n'est pas juste un artefact technique, mais qu'il est ancré dans nos sociétés indépendamment des données qu'on en extrait ou de l'aspect numérique de ces dernières.

4.1.2 Le type de biais géographique

Le cadre utilisé pour l'identification du type de biais géographique est le cadre général présenté en figure 24. Il a été implémenté grâce à l'outil lucidchart[1] et est largement inspiré des travaux de Suresh and Guttag (2021). Ce cadre présente sept types de biais différents, regroupés dans trois catégories de biais, représentant les trois étapes majeures de la conception et du développement d'un système d'IA : la construction de la base de donnée, la conception du modèle, et le déploiement de celui-ci. Le caractère cyclique de ce cadre souligne la manière dont les trois étapes sont aujourd'hui couplées et comment des biais peuvent se renforcer à cause de systèmes déjà biaisés abreuvant le monde de nouvelles données.

Pour comprendre comment le biais géographique est introduit dans les systèmes de vision, reprenons ce cadre et spécifions comment les différents types de biais identifiés peuvent correspondre à un biais géographique.

- **Le biais historique** : un biais géographique peut être de type historique quand il reflète une situation déséquilibrée dans le monde réel. L'accès à un haut débit est touché par ce biais, facilité dans les pays du Nord, permettant un plus grand partage de données que dans les autres zones géographiques. Les volumes de données générés et qui seront ensuite collectés pour construire des bases de données sont donc déséquilibrés, et le déséquilibre créé dans les bases de données pourra avoir un impact sur les performances du modèle dans les pays du Sud.

- **Le biais de représentation** : un biais géographique peut être de représentation quand les données sont échantillonnées de manière déséquilibrées dans les populations concernées. Les bases de données construites via la collecte d'information par capteurs numériques exclura de fait toutes les personnes n'ayant pas accès au numérique, implémentant un tel biais. Ces personnes sont géographiquement plus présentes dans les pays les moins développés. Les modèles construits sur ces bases de données comporteront de fait un biais de représentation.

- **Le biais de mesure** : un biais géographique peut être de mesurement lorsque les caractéristiques utilisées pour évaluer une métrique sont insuffisantes dans un contexte particulier. Les modèles prédisant le remboursement des prêts par des individus se basant sur des dossiers bancaires pourront présenter ce biais s'ils sont déployés dans des communautés non bancarisées, où la participation à des cercles de prêt et à d'autres systèmes moins centralisés permet de mieux prédire le remboursement des prêts.

- **Le biais d'aggrégation** : un biais géographique peut être d'aggrégation quand les concepts utilisés dans la préparation d'une base de données regroupent des populations ou des sous-concepts dont les différences sont effacées par la construction de la base de données. Par exemple, les personnes atteintes de diabète possèdent des morbidités différentes selon leurs ethnies et genres. La mesure des niveaux d'HbA1c, largement utilisée pour suivre le diabète, dépends de manière complexe des caractéristiques ethniques et de genre des individus. Un modèle faisant des prédictions sur l'état de santé d'un patient qui ne prendrait pas en compte ces variations mais utiliserait uniquement le niveau d'HbA1c aurait des prédictions erronées du à un biais d'aggrégation.

- **Le biais d'apprentissage** : un biais géographique peut être d'apprentissage lorsque le processus d'apprentissage induit des conséquences sur les performances d'un modèle sur des populations particulières. Certaines méthodes utilisées pour produire des modèles compacts peuvent amplifier les disparités de performances sur des données avec les attributs les moins représentés[80], témoignant alors d'un biais d'apprentissage.

[1] https://www.lucidchart.com/

- **Le biais d'évaluation** : un biais géographique peut être d'évaluation lorsque les benchmarks utilisées pour évaluer des modèles ne reflètent pas les populations d'usage des modèles. Par exemple, les algorithmes de reconnaissance faciale commerciaux n'étaient testés que sur des bases de données comprenant peu de femmes à peau foncée, ce qui n'a pas permis de détecter et pénaliser les algorithmes qui fonctionnaient moins bien sur ces populations.

- **Le biais de déploiement** : un biais géographique peut être de déploiement lorsque le système est utilisé d'une manière détournée par rapport à son but initial. Ainsi, un système d'évaluation de biodiversité à partir d'images satellites (utilisant par exemple des proxy comme le couvert forestier) entraîné sur des données européennes ne pourra pas être utilisé en Afrique de l'Ouest sans présenter un sérieux biais de déploiement, les environnements présentant des différences profondes.

Le biais géographique est un biais complexe, qui peut être introduit par toutes les manières différentes dans un système d'IA. Il est donc primordial quand on se confronte à un biais géographique, d'identifier à quel type de biais on a affaire, car la simple notion de biais géographique n'est en rien explicite sur la manière dont ce dernier est introduit dans le système.

4.1.3 Configurations pour l' évaluation de l'impact du biais géographique

Afin d'évaluer l'impact du biais géographique, on peut se servir de l'outil développé en section 3.2.4. Cet outil est constitué d'un arbre de décision pour guider un utilisateur sur le choix à réaliser en fonction du type de biais et du caractère direct ou indirect de l'impact de ce biais.

On a pourtant vu dans la sous-section précédente que le biais géographique pouvait être de tout type. A moins d'opérer sur un type particulier de biais géographique, il n'y a donc pas de réduction possible de l'ensemble des tests à un test spécifique pour évaluer l'impact du biais géographique. Néanmoins, dans cette partie, nous décrirons comment les tests proposés dans la section 3.2.4 peuvent être implémentés dans le cadre du biais géographique.

Dans le cas d'un biais à impact direct dont le type est identifié, la figure 25 présente cinq types de configurations possible pour l'évaluation de l'impact du biais. Dans le cas d'un biais géographique, ils peuvent être implémentés des manières suivantes :

- Dans le cas d'un biais relatif aux données et isolable, on pourra séparer la base de données en deux jeux, l'un considéré ED, l'autre HD. On compare alors les performances de modèles sur les deux bases de données. Une telle situation peut être due à une base de données comprenant deux localisations particulières (comme dans GeoNet[88], qui contient des données des USA et d'Asie) ou à une séparation plus conceptuelle, comme une séparation Nord/Sud entre des données.

- Dans le cas d'un biais relatif aux données et intrinsèque à celles-ci, on peut généralement séparer la base de données en k sous-ensembles de données. On évalue alors la performance moyenne ou la pire performance du système sur ces k ensembles. C'est ce que fait le benchmark FMoW-wilds[92], qui propose une configuration pour tester la capacité de modèles à opérer sous glissement de domaine pour la classification de bâtiments sur des images satellites. Ces domaines contiennent des images de différentes locations, réparties dans les jeux d'entrainement, de validation et de test. Une des métriques suivies pour la performance des modèles sur ce benchmark est la pire performance à travers les différentes régions des données test, pour souligner l'importance de produire des modèles capables d'opérer sous différentes géographies.

- Dans le cas d'un biais algorithmique, on peut évaluer l'impact du biais en comparant les performances de modèles de références avec les modèles proposés, introduisant une modification algorithmique permettant d'observer des variations de performances. Les différences de performance des modèles témoignent alors des variations du biais algorithmique.

- On peut aussi dans le cas d'un biais algorithmique faire varier les données utilisées pour entraîner ou évaluer un modèle, jouant sur les variations de bases de données pour souligner l'impact du biais.
- Dans le cas d'un biais de déploiement, on peut évaluer l'impact du biais en comparant les performances des modèles dans des environnements contrôlés ou ne présentant pas le biais identifié avec les performances des mêmes modèles dans des environnements ouverts ou avec présence du biais identifié.

Nous ne traitons pas ici le cas du biais à impact indirect, qui est déjà largement étudié dans la littérature autour de la justice algorithmique. Nous soulignons néanmoins que ces approches peuvent bénéficier de l'approche STI, en intégrant les outils de caractérisation du biais dans leurs démarches.

4.2 Stratégies pour les choix algorithmiques

Les choix algorithmiques dans l'implémentation de systèmes sont nombreux et ne peuvent être tous explicités ou justifiés. Néanmoins, certaines bonnes pratiques permettent aux implémentations d'être plus facilement compréhensibles et reproductibles. Nous précisons ici deux de ces bonnes pratiques, autour de la sélection de modèles et des stratégies d'entraînement.

4.2.1 La sélection de modèles

Il est commun pour les benchmarks de spécifier des stratégies particulières pour la sélection d'algorithme ou d'hyper-paramètres. Gulrajani and Lopez-Paz (2021) soulignent l'importance attachée à une description précise de ces choix et stratégies. D'autres benchmarks imposent des architectures et sélections d'hyper-paramètres, afin d'éviter les irrégularités.

On propose donc dans le cadre du protocole STI de suivre ces recommandations. Dans le cas où un benchmark ou une application n'a pas de stratégies imposées, tout résultat doit être accompagné d'un détail des choix réalisés lors de la sélection ainsi que d'une justification de ces choix.

4.2.2 Stratégies d' entraînement

De même que pour la sélection de modèles, la multiplicité des stratégies d'entraînement mérite de préciser celle sélectionnée lorsque le système nécessite l'entraînement d'un modèle.

Dans le cadre du protocole STI, on propose de situer ces stratégies d'entraînement dans celles qui sont décrites par Koh et al. (2021) dans leur benchmark WILDS. Ils dérivent cinq stratégie particulières pour l'entraînement d'un modèle, qui reflètent des situations particulières en regard de la disponibilité des données et des considérations faites pour la définition des données ED. Préciser quelle stratégie est utilisée pour un système permettra de cerner les choix réalisés lors de l'entraînement de son modèle et augmentera encore la réplicabilité des travaux entrepris.

4.3 Datasets pour la validation

Dans l'optique de valider l'approche STI, nous avons besoin de bases de données dont nous contrôlons les aspects biaisés pour être capable de comparer les résultats obtenus par l'approche testée à la réalité. La nature complexe des images réelles nous force à utiliser des données synthétiques pour la validation souhaitée. Nous nous concentrerons donc dans cette section sur les datasets synthétiques, c'est à dire composés de données synthétiques ou de biais contrôlés de manière algorithmique.

4.3.1 Construction de biais géographiques synthétiques

Il n'existe à priori pas dans la littérature de bases de données à biais synthétiques présentant un biais géographique. Ceci est du à la nature du biais géographique, qui prend sa source dans un contexte ancré dans le réel. Pour autant, on peut identifier des biais géographiques dans la création des bases de données même synthétiques : MNIST, par exemple, qui a été beaucoup repris pour la création de biais synthétique (RMNIST, Colored-MNIST, ...), est basé sur l'alphabet latin, employé en occident. D'autres alphabets (cyrillique, arabe, grec, kanji et autres) n'ont pas eu la même attention de la part de la communauté scientifique. Néanmoins, ce biais n'est pas variant dans les données de MNIST, et ne peux donc être mesuré à l'aide des données de MNIST uniquement. Nous commençons donc ici par créer des bases de données à biais géographiques liés aux données qui les composent.

4.3.1.1 Bases de données à biais géographique synthétique.

Nous nous inspirons pour la création de telles bases de données des variations de la base de données MNIST et combinons différentes variations pour la création d'une base de données biaisée à multiples entrées. Les variations de couleur du fond, des chiffres, ou la rotation centrale de ces derniers offre des occasions de faire intervenir des biais de manière guidée. Pour faire intervenir le biais géographique de manière synthétique, nous allons corréler des transformations des données de MNIST à des disparités géographiques. Pour ce faire, nous associerons à chaque donnée de MNIST un pays, et associerons des transformations possibles en fonction du pays associé à la donnée. La liste des pays qui seront liés aux données de MNIST est issue des données de la division statistique des nations unies, dans leur méthodologie M49[2]. Nous utiliserons tous les pays de la colonne "Country or Aera" de leur base de donnée.

Nous allons ainsi faire intervenir trois facteurs géographiques : la zone géographique (coordonnées en lagitude/longitude), le Produit Intérieur Brut par habitant à parité du pouvoir d'achat (PIB par habitant PPP), et la température moyenne dans les différents pays. Ces trois facteurs seront liés au pays associé à chaque donnée de MNIST de la manière suivante :

- **la zone géographique** sera liée à la coloration du fond de l'image de MNIST. La coloration du fond de l'image dépendra de la latitude et longitude de la capitale du pays qui lui est associé selon la formule suivante :

$$R = (\frac{lon+180}{360}) * 255$$
$$G = 127$$
$$B = (\frac{lat+90}{180}) * 255$$

  avec lat, lon les latitude et longitude de la capitale du pays associé à la donnée, et R,G et B les trois composantes rouges, vertes et bleues des pixels de couleur. Les formules et valeurs utilisées ici sont arbitraires et ont été sélectionnées pour offrir une variabilité continue sur le globe. La carte présente en figure 27 illustre ainsi la couleur associée à chaque pays du monde.

- **le PIB par habitant PPP** sera lié à la rotation centrale de l'image de MNIST. Plus précisément, l'axe maximal de rotation possible pour une donnée sera inversement liée au PIB par habitant PPP, par la formule suivante :

$$\Theta = (1 - (\frac{data_{PIB} - Min_{PIB}}{Max_{PIB} - Min_{PIB}})) * 90$$

  avec Θ l'angle maximal de rotation de la donnée en degrés, $data_{PIB}$ le PIB PPP du pays associé à la donnée, et Min_{PIB} et Max_{PIB} les minimum et maximum respectifs des PIB PPP des pays du monde. Ce lien reflète la réalité qu'une donnée issue de milieux pauvres a moins de chance d'être commune qu'une donnée issue de milieux plus aisés. La carte du PIB PPP par pays est

[2]UNSD Methodology M49 : https://unstats.un.org/unsd/methodology/m49/

présentée en figure 28. Les données utilisées seront celles de la banque mondiale, de l'année la plus récente pour chaque pays[3].

- **la température moyenne** sera liée à la coloration du digit de l'image de MNIST, par association entre température moyenne et couleur au moyen de la formule suivante :

$$R = (\frac{data_T - Min_T}{Max_T - MinT}) * k$$
$$G = 0$$
$$B = (1 - \frac{data_T - Min_T}{Max_T - MinT}) * k$$

avec $data_T$ la température moyenne du pays associé à la donnée, k la valeur initiale du pixel du digit, et R, G et B les trois composantes rouges, vertes et bleues des pixels de couleur. La carte présente en figure 29 représente la coloration associée à chaque pays du monde en fonction de sa température moyenne. Les données sur la température moyenne de chaque pays sont issues du Climate Research Institute[4], et récupérées via wikipedia[5].

Nous utilisons ces trois facteurs pour leur capacité à refléter des aspects différents du biais géographique : l'utilisation des latitudes et longitudes assure un biais géographique du à la localisation des pays. Les températures moyennes reflètent une différence géographique naturelle et indépendante des actions humaines. L'utilisation du PIB PPP reflète la disparité économique globale.

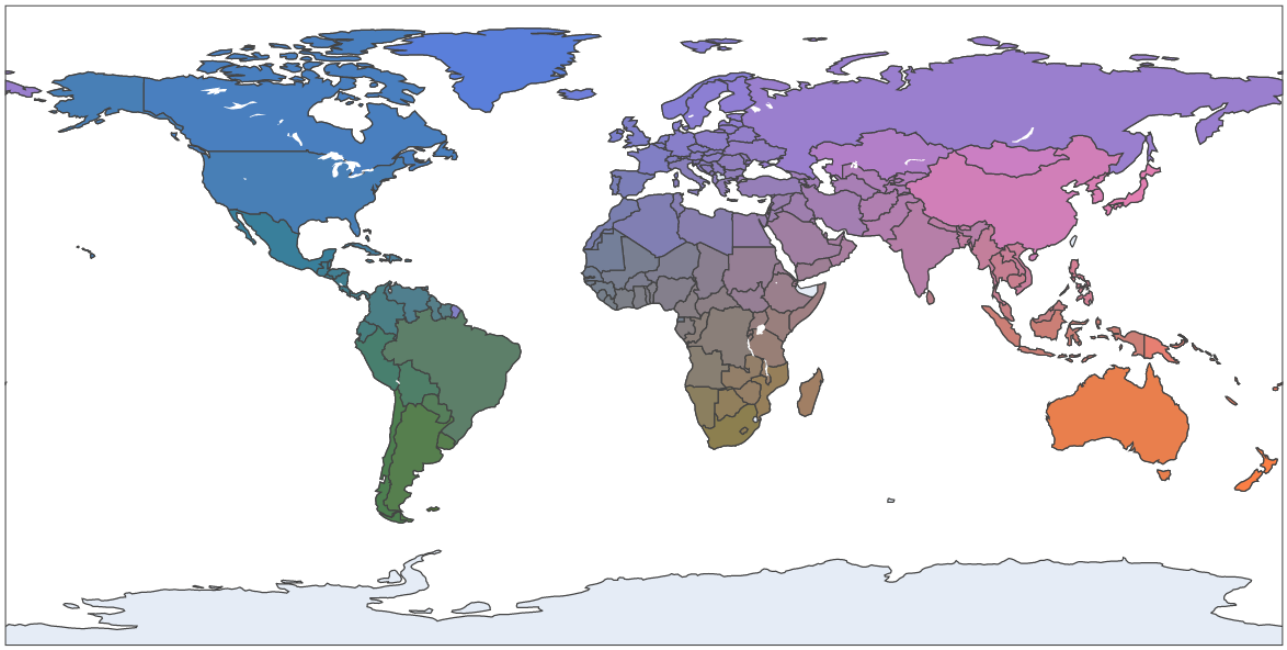


Figure 27: Carte montrant les couleurs attribuées à chaque pays en fonction de sa localisation.

[3]GDP per capita, PPP, current international $, World Bank : https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?most_recen
year_high_desc=true

[4]Climate Research Institute : https://www.uea.ac.uk/groups-and-centres/climatic-research-unit

[5]Wikipedia : "List of countries by average yearly temperature" : https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly

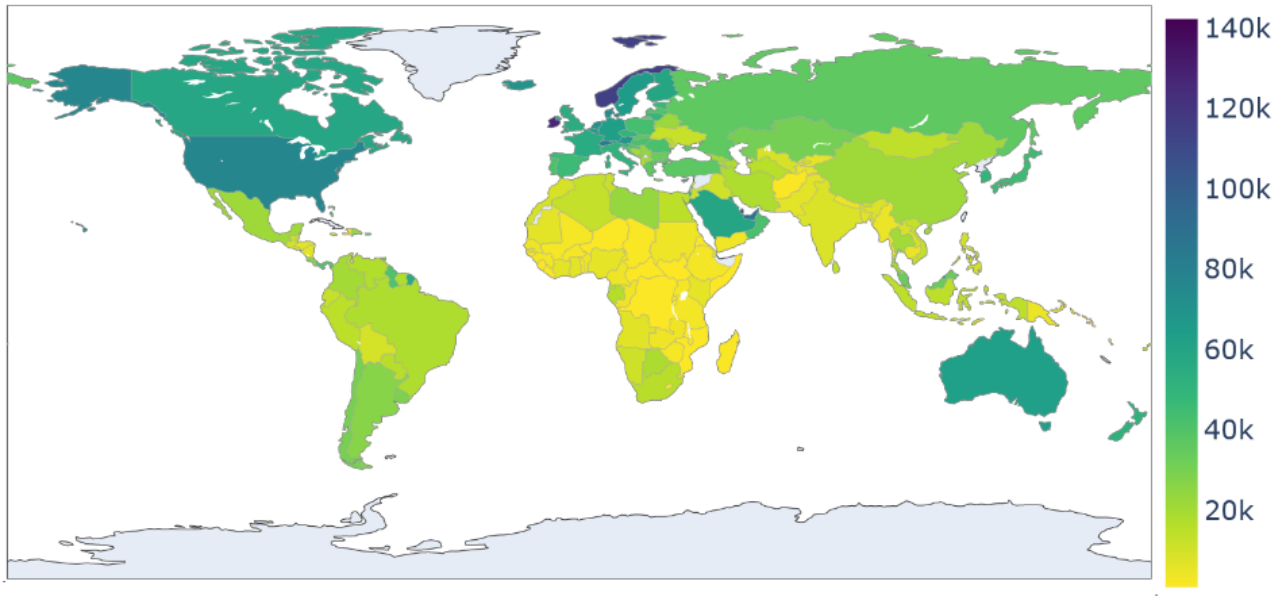


Figure 28: Carte du PIB par pays selon les données de la banque mondiale.

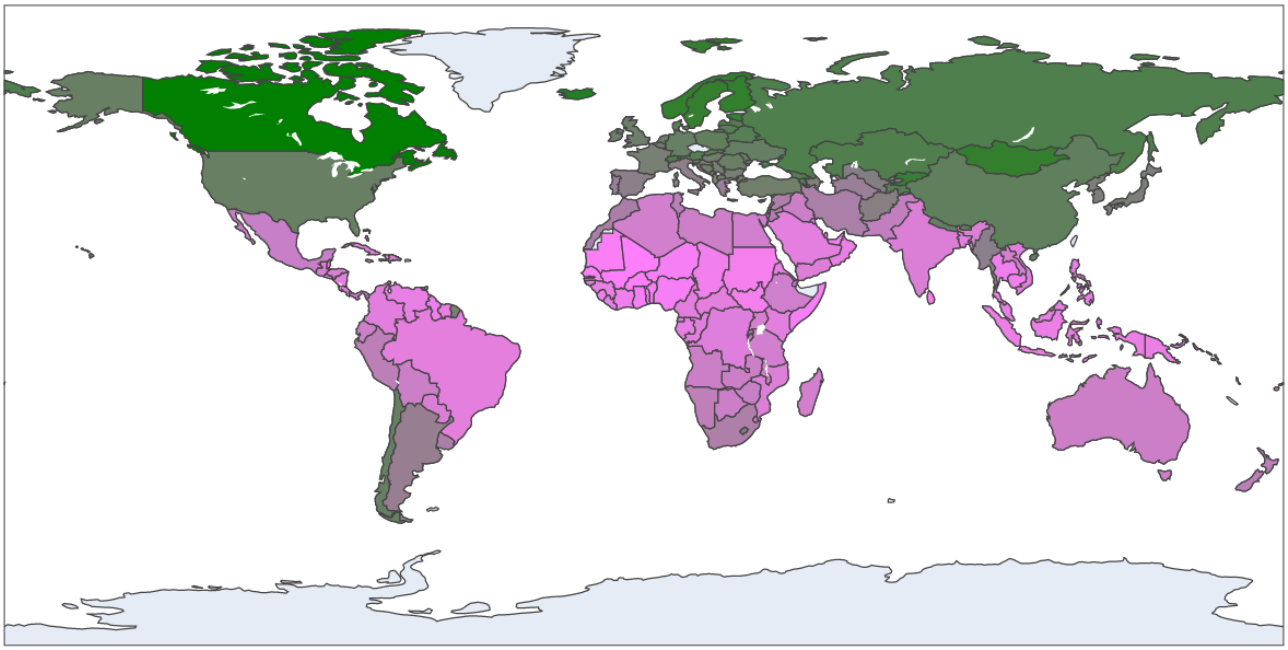


Figure 29: Carte montrant les couleurs attribuées à chaque pays en fonction de sa température moyenne.

Générer de telles modifications de la base de données MNIST nous permet d'introduire des biais géographiques synthétiques. Le choix dans la répartition des pays associés aux données MNIST permet de créer ainsi différentes bases de données avec différentes répartitions de ces biais géographiques. Pour illustrer différentes considérations dans la création de bases de données, nous proposons ainsi trois répartitions différentes : une répartition égalitaire (même nombre de données pour chacun des pays), une répartition conforme à la répartition de la population mondiale (nombre de données associées à un pays corrélé à sa population totale), et enfin une répartition conforme à la répartition des données dans les bases de données d'images génériques (nombre de données associées à un pays corrélé à sa représentativité dans les bases de données comme Imagenet ou OpenImage selon Shankar et al. (2017)). Nous appellerons ces trois bases de données respectivement GeoMNIST-A, GeoMNIST-B et GeoMNIST-C.

4.3.1.2 GeoMNIST-A.

Cette base de données sera composée d'une répartition égalitaire des données de MNIST dans les différents pays. Pour ce faire, nous répartissons chacune des 10 classes dans les jeux d'entraînement, de validation et de test de manière équitable dans les 249 pays. Cette répartition fait écho aux appels pour des bases de données uniformes épurées de tous biais de répartition.

4.3.1.3 GeoMNIST-B.

Cette base de données sera composée d'une répartition des données de MNIST dans les différents 249 pays dans une logique qui suivra la répartition de la population mondiale dans ces 249 pays. Les données pour la population de chaque pays sont issues elles aussi de la division statistique des Nations Unies[6]. Cette répartition fait écho aux appels pour des bases de données plus représentatives du monde réel. La figure 30 illustre la répartition des données de GeoMNIST-B selon le pays associé aux données. Cette répartition est partagée entre les données d'entrainement et les données de test.

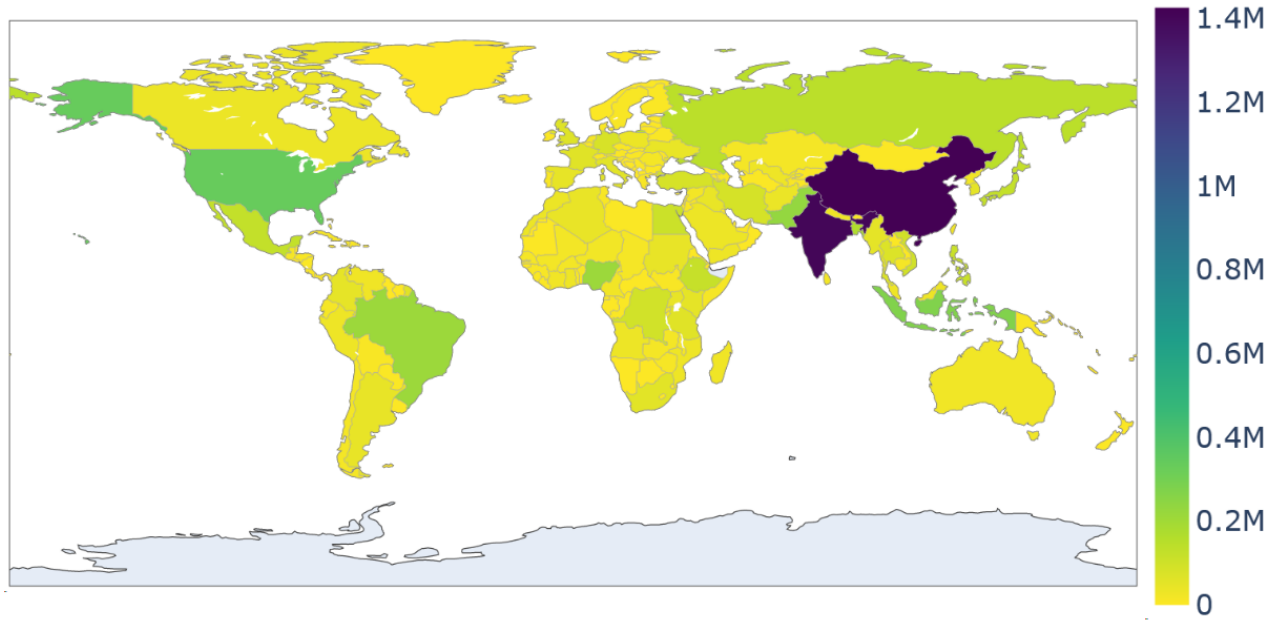


Figure 30: Carte de la répartition mondiale de la population, qui sera utilisée pour générer les données de GeoMNIST-B.

4.3.1.4 GeoMNIST-C.

Cette base de données sera composée d'une répartition des données de MNIST dans les différents 249 pays qui reproduira la répartition des données d'Imagenet dans les pays du monde selon les travaux de Shankar et al. (2017). Cette répartition fait écho aux bases de données existantes et illustre le biais occidental dans la génération de bases de données d'images par collection sur internet. La figure 31 illustre la répartition des données de GeoMNIST-C selon le pays associé aux données. Cette répartition est partagée entre les données d'entrainement et les données de test, et est construite à la main, à l'aide des données des travaux de Shankar et al. (2017).

[6]UNSD UN DATA : http://data.un.org/

Figure 31: Carte de la répartition des données dans les bases de données génériques, qui sera utilisée pour générer les données de GeoMNIST-C.

4.3.1.5 Génération et réplicabilité

Ces bases de données peuvent être générées à chaque utilisation en suivant l'algorithme de génération des données, en fonction des différentes répartition des pays et des transformations associées, mais cela induit une variation dans les bases de données d'une utilisation à une autre. C'est une propriété qui peut être intéressante pour la généralisation et la cross-validation par exemple, mais qui peut nuire à la réplicabilité des résultats des expérimentations. Dans nos implémentations, nous utiliserons donc pour faire office de base de données réplicable des fichiers csv qui attribueront à chaque donnée de MNIST un pays respectif suivant la répartition de la base de donnée choisie, ainsi que les transformations associées à cette donnée. Ainsi, ces bases de données pourront être réutilisées telles quelles pour répliquer les résultats.

4.3.2 Analyse préliminaires du biais géographique synthétique

Nous proposons plusieurs analyses préliminaires pour étudier l'intérêt de nos bases de données synthétiques. Nous commençons par une observation des données générées par les transformations proposées, puis procédons à des analyses par entraînement et tests de modèles sur ces données.

En premier lieu, nous proposons une analyse des performances de quatre modèles entraînés sur chacune des trois bases de données ainsi que la base de données MNIST sans transformation pour comparaison, et testés sur ces quatre bases de données à chaque fois. L' analyse des phénomènes de variation de performance en fonction de la base de données d'entraînement et de test apportera une preuve que ces bases de données construites avec les mêmes règles implémentent bien un biais géographique puisque la seule variation entre ces dernières est la répartition des données dans les pays (l'ajout d'un modèle ne comportant aucune transformation agit en tant que témoin pour comparaison avec nos données plus complexes).

En second lieu, nous proposons une analyse de performance du modèle entraîné sur GeoMNIST-C sur une base de données dont la provenance sera simulée en Afrique de l'Ouest, pour illustrer le cas d'un déploiement de modèle occidental dans un contexte hors-occident.

Pour ces analyses, nous utiliserons des algorithmes et hyper-paramètres standards dans la communauté scientifique. L'utilisation d'un algorithme standard permet d'un côté d'augmenter la réplicabilité des résultats, et de l'autre de renforcer l'influence des résultats en assurant que ces derniers ne sont pas dus à un modèle particulier. Pour chacune de ces analyses, nous utiliserons donc comme algorithme un Resnet-101 couplé à un bloc de classification. Ce bloc de classification sera composé

de deux couches, la première de taille 2048 pour correspondre à la sortie de l'architecture Resnet, et la seconde de taille 512 avec une sortie de taille 10 . Dans les analyses suivantes, l'appellation "le modèle" fera référence à cette architecture. C'est un algorithme standard qui permet un entraînement rapide tout en possédant une architecture assez complexe pour la classification d'images. Cet algorithme est illustré dans le schéma 32. Les données d'entraînement et de test étant définies par les bases de données MNIST, nous utiliserons la répartition par défaut sans plus de modifications. Les modèles sont entraînés pendant 200 epochs avec un batch de taille 1024 sur une carte graphique NVIDIA quadro P4000, et seule la meilleure itération sur le jeu de validation est conservée après ces 200 epochs.

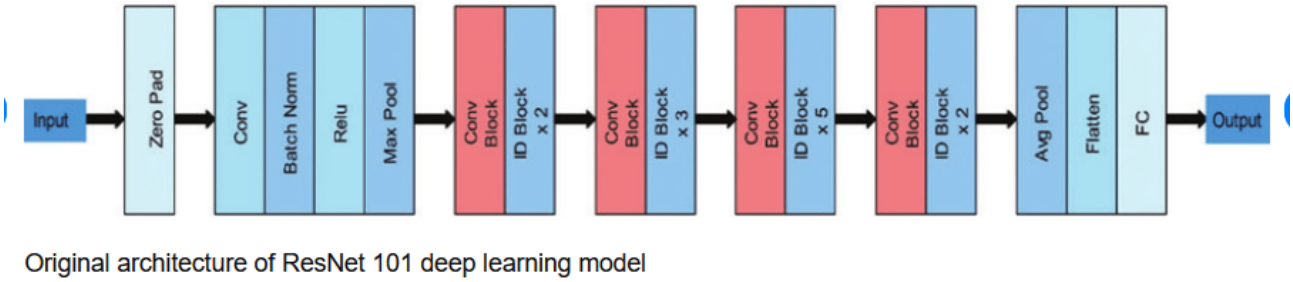


Figure 32: Présentation de l'architecture Resnet-101 issue de He et al. (2016).

4.3.2.1 Observation et complexité des données

On observe sur la figure 33 que dans certains cas, les rotations sont importantes et peuvent presque entièrement renverser la donnée de 90°, comme c'est le cas pour la Zimbabwe ou la Colombie. Dans d'autres, les couleurs du fond de l'image et du chiffre affiché se confondent presque, rendant difficile la distinction à l'oeil humain, par exemple pour les données associées à la Grèce. Ces données sont donc de complexité différente, la provenance géographique influant sur la difficulté de la donnée. Notons néanmoins qu'une variation de couleur entre fond de l'image et chiffre difficilement discernable pour l'oeil humain ne l'est pas forcément pour la machine, et que la notion de complexité d'une donnée peut varier de la perception humaine à la perception machine.

Il est aussi à noter que cette complexité n'est pas issue de la répartition de la base de données A, B ou C, mais bien des transformations liées aux données elle-mêmes, et donc uniquement des transformations apportées aux données en fonction du pays qui y sont associés. La complexité même de la donnée n'a donc rien à voir avec la répartition, mais la répartition peut ajouter à la complexité des données une complexité due à des pays moins représentés que d'autre, et donc favoriser certains types de données.

Figure 33: Quelques données issues de GeoMNIST-C, avec les pays attachés aux données. Les transformations des données issues de MNIST dépendent uniquement des pays rattachés aux données.

modèle	GeoMNIST-A	GeoMNIST-B	GeoMNIST-C	MNIST
A	**0.4114**	**0.4167**	**0.3514**	0.1157
B	0.1654	0.2411	0.1389	0.0958
C	0.0979	0.1071	0.2355	0.101
N	0.0997	0.0992	0.102	**0.6919**

Table 4: Tableau de report des performances des modèles entraînés sur les bases de données GeoMNIST et testées sur ces dernières. Le modèle A est entraîné sur GeoMNIST-A, et une logique similaire (i.e B pour GeoMNIST-B et C pour GeoMNIST-C) est appliquée pour les modèles B et C. Le modèle N est entraîné sur MNIST.

4.3.2.2 Première analyse : comparaison croisée des performances sur les bases de données synthétiques.

Pour cette analyse et les suivantes, nous nous référerons au modèle entraîné sur GeoMNIST-A par "le modèle A" et suivrons la même logique pour les appellations "modèle B" et "modèle C". Nous procédons pour cette analyse à l'entraînement de trois modèles sur chacun des trois jeux d'entraînement des bases de données, ainsi qu'un quatrième modèle sur le jeu de données MNIST sans modification autre qu'une projection sur 3 channels de la valeur de niveau de gris. Ce modèle est appelé "modèle N" pour "normal" dans la suite. Nous comparons ensuite les performances de ces quatre modèles sur chacun des trois jeux tests des bases de données et sur MNIST sans transformation. Les résultats sont reportés dans le tableau 4. En moyenne sur les quatre bases de données, le modèle A obtient une précision de 0.3238, le modèle B une précision de 0.1603, le modèle C une précision de 0.1354 et le modèle N une précision de 0.2482.

On observe que bien que les trois bases de données soient construites avec les mêmes règles, une variation de la répartition des données entraîne une variation de performance du modèle, toute chose égale par ailleurs. Le modèle A, construit sur une base de donnée équilibrée, est celui qui présente les meilleurs performances sur les trois bases de données. Les modèles B et C quant à eux semblent peiner à obtenir des performances intéressantes, toutes bases de données confondues, bien qu'ils performent un peu mieux sur leurs distributions d'entraînement respectifs. Si on regarde les matrices de confusion en figure 34, on observe que cela est du au fait que le modèle B semble fournir en output principalement la classe 2 tandis que le modèle C semble fournir en output principalement les classes 0, 3, 8 et 9. Pourtant, les classes sont équitablement réparties dans les différents jeux d'entraînement, de validation et de test. Ces trois modèles n'arrivent pas à obtenir de prédictions satisfaisantes sur la base de données MNIST, et à l'inverse, le modèle N n'arrive pas à obtenir de prédiction satisfaisantes sur les données GeoMNIST. Une version zoomée des prédictions du modèle A sur GeoMNIST-A est proposée en figure 35.

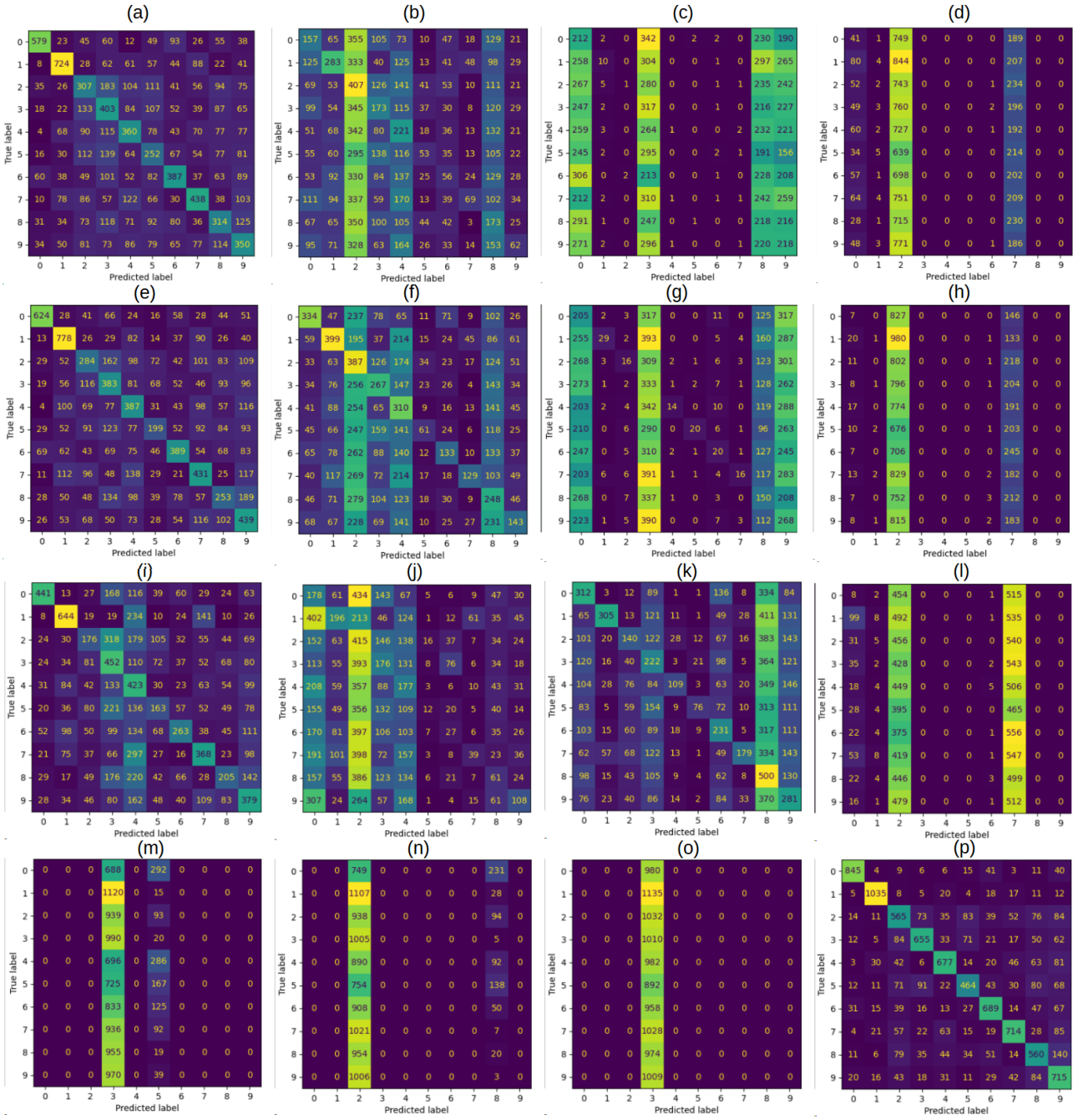


Figure 34: Matrices de confusion des classes pour les différents modèles et bases de données. (a) Modèle A sur GeoMNIST-A (b) Modèle B sur GeoMNIST-A (c) Modèle C sur GeoMNIST-A (d) Modèle N sur GeoMNIST-A (e) Modèle A sur GeoMNIST-B (f) Modèle B sur GeoMNIST-B (g) Modèle C sur GeoMNIST-B (h) Modèle N sur GeoMNIST-B (i) Modèle A sur GeoMNIST-C (j) Modèle B sur GeoMNIST-C (k) Modèle C sur GeoMNIST-C (l) Modèle N sur GeoMNIST-C (m) Modèle A sur GeoMNIST-N (n) Modèle B sur GeoMNIST-N (o) Modèle C sur GeoMNIST-N (p) Modèle N sur GeoMNIST-N.

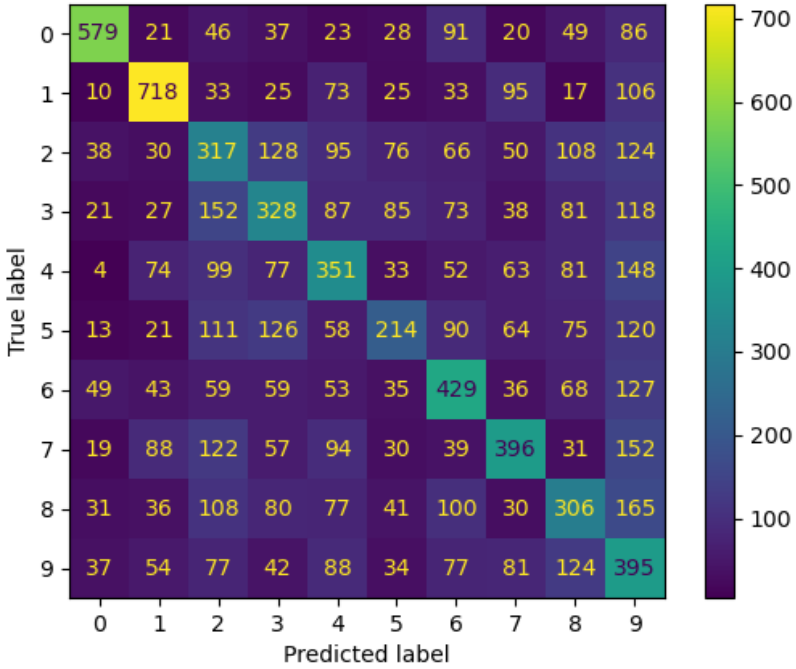


Figure 35: Matrices de confusion des classes pour le modèle A sur GeoMNIST-A.

Cela illustre comment les biais de répartition impactent les performances des modèles en vision par ordinateur. Un modèle issu d'un jeu de donnée dont la répartition géographique n'est pas optimale sera moins performant qu'un modèle entraîné sur une base de donnée à répartition géographique optimale, toute chose étant égale par ailleurs. Ce résultat peut être étonnant dans le sens où l'on s'attend à ce qu'un modèle entraîné sur une distribution de données et testé sur cette dernière soit plus performant qu'un modèle entraîné sur une autre distribution de données, pourtant en tout cas le modèle A est meilleur que les modèles B et C. Ces résultats illustrent l'importance de bases de données inclusifs et représentatifs dans la recherche de performances globales.

On observe aussi que les performances des modèles dépendent des bases de données utilisées. Les modèles A et B obtiennent leur meilleures performances sur GeoMNIST-B, tandis que le modèle C obtient ses meilleures performances sur GeoMNIST-C et le modèle N sur MNIST. On s'attend à ce qu'un modèle soit performant sur le jeu test correspondant à la distribution sur laquelle il a été entraîné, et on peut donc être surpris de voir que le modèle A obtient des meilleures performances sur GeoMNIST-B que sur GeoMNIST-A. Ce résultat s'explique par une variation dans la difficulté des données, les données à forte transformation (inclusives) étant plus représentées dans GeoMNIST-A que dans GeoMNIST-B. Ces résultats illustrent comment une variation dans les métadonnées géographiques associées aux données permet de générer des bases de données à difficulté variable.

Ces premiers résultats soulignent bien l'intérêt de tels bases de données à biais géographique synthétique : les variations de répartitions géographiques des données et les transformations associées aux pays du globe font des différentes versions de GeoMNIST des reflets utiles de situations réelles plus complexes.

4.3.2.3 Seconde analyse : illustration du déploiement d'un modèle hors-occident.

Dans cette seconde analyse, nous proposons de comparer les performances du modèles C sur son jeu test à ses performances sur des données dont la provenance est simulée en Afrique de l'Ouest. Nous appellerons ce jeu de données test "GeoMNIST-AO". Les résultats sont reportés dans le tableau 5.

modèle	GeoMNIST-C	GeoMNIST-AO
C	0.2355	0.0947

Table 5: Performance du modèle C sur les bases de données GeoMNIST-C et GeoMNIST-AO. GeoMNIST-AO simule des données GeoMNIST issues d'Afrique de l'Ouest.

On observe que les performances du modèle C ne se transposent pas sur GeoMNIST-AO. Ce phénomène, conséquence d'un glissement de répartition dans les données, souligne l'importance du contexte lié à la donnée. La seule variation entre GeoMNIST-C et GeoMNIST-AO étant une variation dans la répartition géographique des données, on a bien ici une diminution de performance liée à un biais géographique dans un jeu de données synthétique. Puisque l'on est capable de témoigner de l'impact du biais du modèle sur une zone géographique particulière, alors on doit être capable de caractériser ce biais sur l'ensemble du globe, c'est à dire de décrire l'impact du biais sur toutes les zones géographiques. C'est ce que nous proposons dans les prochaines expérimentations.

4.4 Expérimentations

Après ces analyses préliminaires, nous cherchons ici à démontrer l'intérêt du protocole STI sur les données synthétiques générées. Pour ce faire, nous allons chercher à caractériser le biais géographique du modèle C par deux moyens : sans le protocole STI, puis avec le protocole STI, et comparer les résultats obtenus. Le choix du modèle C est un choix réfléchi, censé refléter l'utilisation de modèles entraînés sur les grandes bases de données d'images comme Imagenet ou Open Images qui manquent de représentativité géographique.

4.4.1 Caractérisation sans le protocole STI

Dans cette première expérience, nous reportons les performances du modèle C sur les différentes zones géographiques représentées dans la base de données GeoMNIST-A. Nous proposons ainsi une caractérisation du biais en reportant les variations géographiques de performances d'un tel modèle. Les résultats sont présentés dans la carte 36. La performance moyenne du modèle C sur GeoMNIST A de 0.0979 est déjà reportée dans le tableau 4 et on peut déjà s'attendre à une variation de performance en fonction des pays puisque la performance du modèle C sur son propre jeu de données est de 0.2355.

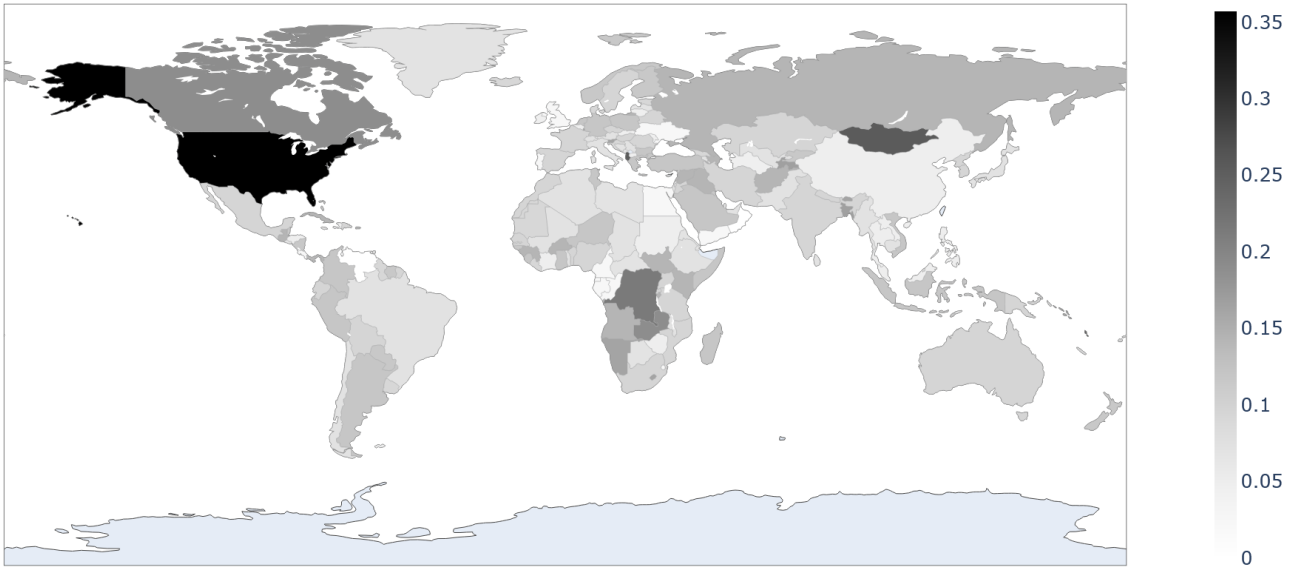


Figure 36: Carte de performance du modèle C sur GeoMNIST-A.

Cette carte de performance par pays est utilisée par DeVries et al. (2019) pour reporter une baisse de performance des API de grand groupes du secteur de l'IA dans les zones géographiques les moins

représentées dans les bases de données d'images. Ici, la carte de performance par pays nous permet de comprendre comment le modèle est influencé par deux facteurs : la difficulté des données et la variation de distribution entre données d'entraînement et données de test. Le modèle C est entraîné majoritairement avec des données dont la provenance est simulée occidentale, comportant donc des transformations plus légères que les données de GeoMNIST-A issues d'autres zones géographiques. On observe sur cette carte que le modèle C est particulièrement performant en Amérique du Nord avec une performance moyenne à 0.3571 aux État-Unis et 0.1905 au Canada. Il obtient aussi de bonnes performances dans d'autres pays comme en Mongolie ou en Albanie où le modèle atteint 0.2558 et en République Démocratique du Congo où il atteint 0.2142. Dans le reste du monde, les performances sont variables mais généralement basses, d'où une moyenne mondiale à 0.0979. Nous analysons plus en avant ces résultats en partie 4.5.

4.4.2 Caractérisation avec le protocole STI

Dans cette seconde expérience, nous implémentons le protocole STI pour identifier un biais spécifique dans le modèle C et déterminer la stratégie pour témoigner de l'impact de ce dernier. Commençons par identifier le biais en question, c'est à dire déterminer sa source et son type, avant de déterminer comment mesurer son impact.

La source du biais géographique que nous étudierons est le manque de représentation de données non occidentales dans la base de données GeoMNIST-C. Ce biais peut être de multiples types : un biais historique dans les données dû aux choix d'échantillonnage réalisés lors de la conception de la base de données, ou encore un biais de représentation dans la définition des populations et l'échantillonnage. Nous considérerons ici un autre type de biais, sans pour autant en changer la source, pour illustrer la particularité du protocole STI.

Le type du biais géographique sur lequel nous nous concentrons ici est le biais d'évaluation du modèle. Les autres types pourraient aussi être abordés, mais dans un soucis de synthèse et dans l'esprit du protocole STI, nous identifions un type de biais en particulier sur lequel travailler. Ce biais d'évaluation intervient au moment d'évaluer les performances du modèle, et donc de l'inférence du modèle entraîné sur le jeu de test. Le biais se situe alors soit dans le processus d'évaluation (tâche évaluée, métrique d'évaluation), soit dans le jeu de données de test (biais historiques, de représentation, de mesurement dans la construction de la base de données test). Nous travaillerons sur la caractérisation du biais d'évaluation en explorant deux choix de l'évaluation, la métrique d'évaluation du modèle et le jeu de données de test.

L'impact du biais géographique devra donc permettre de témoigner du biais d'évaluation issu du manque de représentativité des données. Suivant l'arbre 25, cet impact pourra se mesurer en procédant à des variations de modèle ou de base de données. Nous procéderons ici à une variation dans les processus d'évaluation du modèle via des modifications de métriques et du jeu de donnée de test. Nous pourrons comparer les performances issues des différentes variations pour chiffrer le déplacement et établir une mesure de la distance entre évaluation initiale et variation de l'évaluation. Cette mesure sera établie comme témoin et impact du biais géographique d'évaluation issu du manque de représentativité des données.

Implémentation et description : nous identifions le glissement géographique par les variations de performance du modèle C sur GeoMNIST-A comparé à ses performances sur GeoMNIST-C. Cette différence est de 0.1376, mais cette seule valeur ne permet pas de caractériser le glissement géographique. Pour caractériser ce biais, nous reprenons la carte du modèle en fonction des pays, mais prenons comme valeur pour chaque pays la différence entre les performances du modèle C et les performances du modèle A. Cette comparaison pays par pays prend en compte les variations de difficulté entre les données de GeoMNIST-A et GeoMNIST-C, et isole donc le biais géographique. Les valeurs seront majoritairement négatives, puisque le modèle A est plus performant que le modèle C ; mais cela permettre d'identifier plus facilement les biais du modèle C en exposant les pays pour lesquels il est plus performant que le modèle A. La carte est représentée en figure 37.

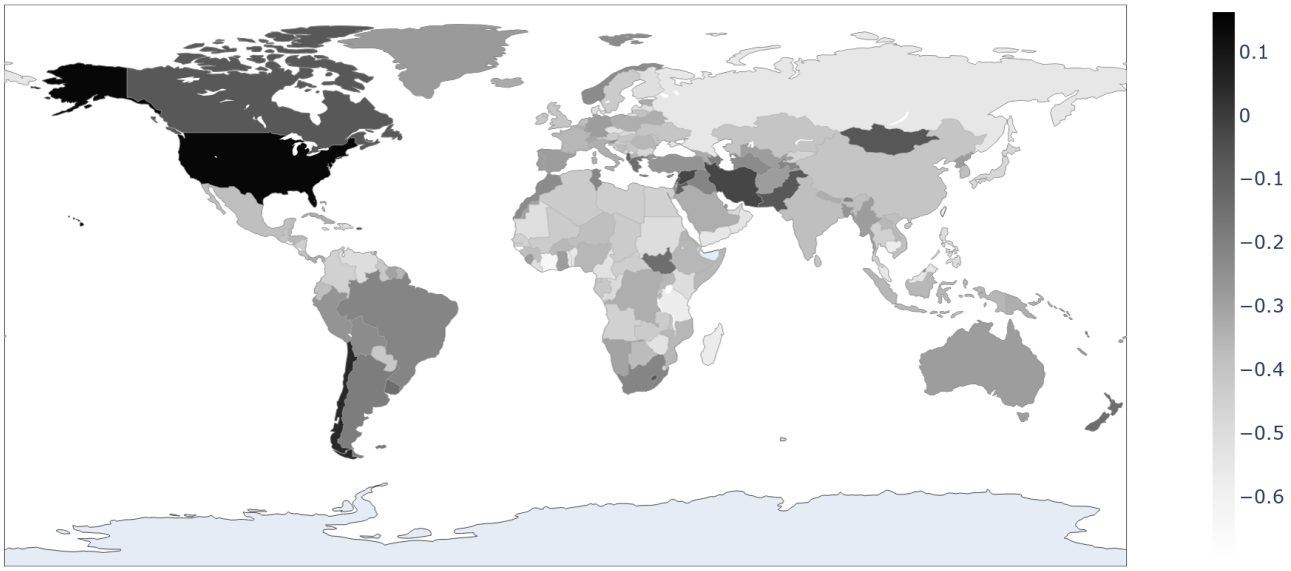


Figure 37: Carte montrant la différence de performance par pays entre les modèles C et A sur GeoMNIST-A. Cette visualisation permet de s'affranchir de la difficulté des données et d'isoler le biais géographique associé au modèle C.

On observe que certains pays ont des valeurs positives (États-Unis, Chili, ainsi que certaines îles comme les Samoa, Tonga, Îles Caimans, Niué, les Îles Cook) tandis que la plupart des pays ont une valeur négative. Cela reflète le fait que le modèle A est plus performant que le modèle C au niveau global sur l'ensemble des données mondiales. Les performances du modèle C sur les État-Unis et certaines îles reflètent le biais géographique intégré dans le modèle, qui est entraîné majoritairement sur des données issues des États-Unis. Les îles pour lesquelles le modèle C est meilleur que le modèle A sont groupées à l'est de l'Australie, et donc géographiquement proche d'un des pays les plus représentés dans GeoMNIST-C. Ces résultats sont donc cohérents avec l'observation attendue : le modèle C possède un biais géographique, et il est aligné avec la répartition des données dans GeoMNIST-C.

L'intérêt de cette approche est de pouvoir comprendre où le modèle est performant, où il ne l'est pas, et là où les efforts sont à fournir pour améliorer le modèle. La comparaison avec le modèle A permet de souligner les biais de la distribution apprise par le modèle C. Nous avons ici deux modifications à l'évaluation du modèle C : une modification de la base de donnée test utilisée (de GeoMNIST-C à GeoMNIST-A) et une modification de la métrique utilisée (des performances moyennes globales à l'utilisation d'une carte de performance). Pour autant, l'utilisation d'une carte comporte des désavantages : subjectivité des observations, nécessité d'une interprétation humaine[19]. On peut avoir un intérêt particulier à utiliser plutôt des métriques numériques, pour pouvoir plus facilement réaliser des comparaisons et mettre en place des benchmark.

D'autres métriques comme la pire performance par pays, utilisée par Koh et al. (2021) par exemple dans certains benchmarks de WILDS, permettent de porter une attention particulière aux classes défavorisées par les modèles, puisque ce sont ces dernières qui définissent la performance évaluée. Le tableau 6 contient les performances des différents modèles entraînés sur les différentes bases de données utilisées en utilisant pire performance par pays comme métrique de performance des modèles.

modèle	GeoMNIST-A	GeoMNIST-B	GeoMNIST-C	MNIST
A	0,0238	0	0	0,0233
B	0	0	0	0
C	0	0	0	0
N	0	0	0	0,4524

Table 6: Tableau de report des performances des modèles entraînés sur les bases de données GeoMNIST et testées sur ces dernières. Le modèle A est entraîné sur GeoMNIST-A, et une logique similaire est appliquée pour les modèles B et C. Le modèle N est entraîné sur MNIST. La métrique utilisée ici est la pire performance par pays.

Ce tableau illustre le fait que la pire performance n'est ici pas gage d'une bonne métrique pour suivre le glissement géographique. Le nombre important de pays ainsi que les problèmes de manque de représentation de certains pays dans GeoMNIST-B et GeoMNIST-C lisse les pire performances vers 0 dans de nombreux cas. Il est donc nécessaire de créer d'autres métriques numériques que la pire performance par pays.

On peut s'inspirer des statistiques et des utilisations des quartiles pour exprimer des mesures d'intérêts. On peut ainsi reporter les performances des modèles sur différents sous-ensembles de pays par exemple, en les séparant par zone géographique ; ou reporter les performances aux différents quartiles pour donner une idée de la manière dont le modèle performe globalement. On présente dans le tableau 7 l'implémentation d'une telle métrique, en reportant pour chaque combinaison de modèle et base de donnée la performance moyenne, la meilleure performance, la moyenne des 10% des pays les plus performants, la moyenne des 10% des pays les moins performants, et la pire performance.

Modèle	Dataset	Moyenne	Best	Top 10%	Worst 10%	Worst
A	GeoMNIST-A	0,4118	0,8140	0,6821	0,0748	0,0238
A	GeoMNIST-B	0,3977	1	1	0	0
A	GeoMNIST-C	0,4175	1	.880	0	0
A	GeoMNIST-N	0,1158	0,2381	0,2146	0,0349	0,0233
B	GeoMNIST-A	0,1657	0,5476	0,4096	0,0279	0
B	GeoMNIST-B	0,1807	1,0	1,0	0	0
B	GeoMNIST-C	0,1848	0,8	0,68	0	0
B	GeoMNIST-N	0,0958	0,2093	0,1868	0,0187	0
C	GeoMNIST-A	0,0979	0,3571	0,2371	0,0163	0
C	GeoMNIST-B	0,0754	1	0,5569	0	0
C	GeoMNIST-C	0,0885	0,6	0,4207	0	0
C	GeoMNIST-N	0,101	0,2326	1915	0,021	0
N	GeoMNIST-A	0,0997	0,2326	0,2005	0,0212	0
N	GeoMNIST-B	0,0995	1	0,7429	0	0
N	GeoMNIST-C	0,1117	0,4	0,4	0	0
N	GeoMNIST-N	0,6921	0,9048	0,8669	0,5139	0,4524

Table 7: Tableau de report des performances par pays des modèles entraînés sur les bases de données GeoMNIST et testées sur ces dernières. Différentes métriques sont utilisées ici pour tracer les performances des modèles sur les différents GeoMNIST.

On observe que les valeurs des performances communes avec le tableau 4 divergent - c'est dû à la méthode de calcul différente, qui utilise les scores par pays au lieu de se reporter au score par donnée. Dans cette configuration, tous les pays ont la même importance, alors qu'autrement, les pays les plus représentés ont plus d'importance que les autres dans le calcul de la métrique.

A partir de toutes ces données, nous pouvons proposer plusieurs moyens de témoigner de l'impact du biais d'évaluation géographique du modèle C:

- La carte 37, qui nécessite une interprétation humaine, mais offre une bonne granularité pour comprendre les variations de performance par pays.
- La différence de performance entre le modèle A et le modèle C sur la base de donnée GeoMNIST-A (performance moyenne par exemple, ou une autre des mesures de performance utilisée dans le tableau 7). Cette métrique permet de s'affranchir de facteurs tels que la difficulté différente des données en fonction des pays et de complètement isoler le biais géographique.
- La différence de performance entre le modèle C en moyenne sur les données du jeu test de GeoMNIST-C et le même modèle en moyenne par pays sur les mêmes données. Cette métrique permet de s'affranchir de la nécessité d'avoir d'autres données et d'autres modèles que le modèle et les données initiales, et est donc minimaliste, tout en soulignant les écarts entre performance moyenne et performance moyenne par pays.

4.5 Discussions et analyses

Dans les expérimentations sans puis utilisant le protocole STI précédentes, nous caractérisons le glissement géographique du à la répartition des pays dans GeoMNIST-C. Puisque c'est à travers les modèles que nous observons le biais géographique, nous confondrons le fait de caractériser un biais et celui de caractériser le biais appris par le modèle dans la suite. Dans la première expérimentation, nous avons employé les moyens classiques, et repris la méthodologie de DeVries et al. (2019) : l'utilisation d'une carte de performance par pays des modèles. Dans la seconde, nous rajoutons l'analyse STI pour caractériser le biais et identifions particulièrement ce qui est fixe et ce qui est variable dans l'analyse.

4.5.1 Caractérisation du glissement géographique

Nous tentons dans la première expérience une caractérisation en utilisant un outil utilisé par DeVries et al. (2019), une carte de performance par pays (figure 36). Cette carte, si elle possède certains désavantages, permet de saisir globalement les tendances qui existent dans le modèle utilisé. Mais comment expliquer les performances du modèle C en Mongolie, en Albanie ou en République Démocratique du Congo ? Cette carte ne permet pas d'apporter de réponse, simplement de poser l'hypothèse que c'est du à une différence de difficulté entre les données de ces pays et celles du reste du monde hors pays représentés dans GeoMNIST-C.

La seconde expérimentation propose un découpage du glissement géographique selon le protocole STI proposée, et commence par sélectionner une source et un type de biais pour ensuite choisir un impact approprié. Les choix réalisés reflètent au mieux les cas d'utilisations de modèles de classifications dans le monde, en ne remettant pas en question les données d'entraînement ou les principes d'entraînement du modèle (qui sont généralement fixés et inaccessibles dans le cas d'API), mais en proposant d'autres moyens d'évaluer le modèle sélectionné (et donc en remettant en cause le processus d'évaluation du modèle, en identifiant et caractérisant ses biais). C'est pourquoi l'enjeu du protocole STI est alors d'identifier des métriques d'impact capable d'isoler et de témoigner du biais géographique pour pouvoir le caractériser de manière précise.

La carte proposée en figure 37 est inspirée d'une définition donnée par Koh et al. (2021) d'un glissement entre domaines, qui doit être la différence entre les performances atteignables sur le domaines cible et les performances atteintes par le modèle sur ce domaine cible. De fait, les performances atteignables sont interprétées par les performances du modèle A, entraîné sur GeoMNIST-A puisque c'est le jeu de test de cette base de données qui est utilisé en tant que domaine cible. Cela permet de prendre en compte la difficulté des données, et d'isoler le biais géographique du modèle C comme seul facteur des différences de performance entre les modèles, puisque ces données seront

également difficiles pour les deux modèles, et la différence de performance ne dépendra alors pas de cette caractéristique. On retrouve des résultats cohérents, comme le fait que les performances du modèle C sont meilleures sur les États-Unis que le modèle A, témoignant d'une spécialisation du modèle sur des données sur-représentées dans GeoMNIST-C. D'autres pays (îles du Pacifique notamment) témoignent du même phénomène, et ne sont pas le fruit d'une spécialisation sur ces pays précis mais de la proximité des données avec d'autres données fortement représentées (notamment issues des États-Unis ou de l'Australie).

Pour autant, cette carte ne permet pas de construire des benchmarks ou d'automatiser des comparaisons de modèles, puisqu'elle est soumise à une interprétation visuelle et donc subjective et humaine. Nous proposons dans le tableau 6 une autre métrique, qui est la pire performance par pays, utilisée dans d'autres benchmarks comme certains de WILDS[92]. Cette métrique se montre inefficiente dans le cas présent, et nous supposons que c'est dû à un grand nombre de pays et à une faible représentation de certains pays dans les données de GeoMNIST-B et GeoMNIST-C. Cette faible représentation entraîne un faible nombre de données sur lesquelles il est facile de n'avoir aucune prédiction valide, entraînant un résultat commun nul sur ces bases de données.

Nous proposons en réponse un report plus complet de performance, en reportant les meilleures performances, les pires performances, ainsi qu'une moyenne des 10% meilleures et 10% pires performances du modèles, visible dans le tableau 7. Ce tableau permet de mieux cerner le comportement des modèles dans les différentes bases de données de GeoMNIST sans pour autant surcharger le lecteur d'informations. Il permet aussi d'observer à quel point les modèles sont consistants - en comparant par exemple les meilleures et pires performances, ou les 10% meilleures et 10% pires performances. Un tel report en utilisant les performances par pays modifie aussi les performances moyennes précédemment reportées, et diminue les performances du modèle C sur son propre jeu de données. Ainsi, en passant d'une performance moyenne sur l'ensemble des données à une performance moyenne par pays, on observe déjà une chute de performance qui témoigne de l'impact du glissement géographique, puisque donnant autant d'importance à chacun des pays au lieu de donner plus d'importance aux pays les plus représentés.

Si le tableau détaillé des performances par pays ne permet pas de suivre les performances de manière géographique, il permet d'observer et d'établir des métriques pertinentes pour évaluer le biais géographique, et de faire des choix réfléchis à partir de ces métriques. Les métriques pour témoigner du biais proposées reflètent des situations et cas d'utilisations différents. La carte suppose une intervention humaine et subjective, et donc élimine la possibilité de l'automation de la comparaison à d'autres modèles. Néanmoins, elle permet une vue d'ensemble et une compréhension plus complète du phénomène de glissement géographique que ses contreparties. La différence de performance entre le modèle A et le modèle C sur les données test de GeoMNIST-A permet d'isoler le biais géographique du modèle C pays par pays, ce qui en fait une métrique précise pour témoigner du biais géographique ; néanmoins, elle nécessite un nouveau jeu de données (GeoMNIST-A) et un modèle supplémentaire (le modèle A) pour établir la métrique désirée. La troisième métrique est plus minimaliste, nécessitant uniquement les données du jeu test GeoMNIST-C, et comparant les performances par donnée et les performances par pays. Cette métrique a pour autant un désavantage, celui de risquer d'être biaisé par une trop faible représentativité dans certains pays, et donc de donner énormément d'importance à quelques données en particulier ; il serait facile alors d'améliorer les performances d'un modèle utilisant cette métrique en se spécialisant sur quelques données en particulier plutôt que d'améliorer la capacité générale du modèle.

4.5.2 Pertinence du protocole de validation

Comment évaluer alors le protocole STI dans sa capacité à caractériser le biais géographique sur cet exemple ? Reprenons ce que nous apportent les deux expérimentations. La première expérimentation, grâce à la carte en figure 36, informe que le modèle C possède un biais occidental fort, avantageant en particulier les données Nord-américaines. Nous concluons de cette expérience que le modèle C aura des performances moindres dans les autres zones géographiques, et qu'il faudra probablement

opérer des transformations et améliorations pour le déployer hors de son contexte d'entraînement. La seconde expérimentation prend du recul sur le glissement cible en le découpant en source, type et impact. On situe la source dans la répartition des pays dans GeoMNIST-C, et on s'intéresse uniquement au biais d'évaluation, c'est à dire au biais dû à l'évaluation du modèle C sur un jeu test possédant la même distribution que son jeu d'entraînement. L'exploration des métriques d'impact possible offre de nombreuses informations supplémentaires. La carte en figure 37 nous informe que le modèle C est spécialisé sur les données des États-Unis, et sous-performe en moyenne dans les autres pays du monde, ce qui nous confirme que le modèle a été influencé par la répartition de sa base de de données, mais aussi qu'il peut être amélioré en utilisant des données mieux réparties. Cette carte nous informe donc d'une piste possible pour l'amélioration du modèle. Le tableau 7 est porteur non seulement d'information sur la stabilité des modèles à travers les différents jeux de tests de GeoMNIST, mais aussi des tendances de ces modèles sur ces jeux tests. La seconde métrique proposée nous donne une mesure numérique de la capacité d'amélioration du modèle C sur la base de donnée GeoMNIST-A, et témoigne donc de l'impact de la répartition de GeoMNIST-C sur les capacités de généralisation d'un modèle entraîné sur cette dernière. Cette mesure souligne donc la manière dont l'évaluation du modèle C ne permet pas d'évaluer la capacité de généralisation du modèle. La troisième métrique proposée a l'avantage de ne reposer que sur les données de GeoMNIST-C et du modèle C. En utilisant la différence entre performance moyenne et performance moyenne par pays, on écrase l'influence des pays les plus représentés et on augmente fortement l'influence des pays les moins représentés. Bien que plus facilement "hackable" en spécialisant le modèle sur les quelques données des pays les moins représentés, cette métrique a l'avantage de remettre en question la prédominance des localisations occidentales dans l'évaluation d'un modèle.

Il n'y a pas de choix final de métrique, car ce n'est pas l'objet de cette expérience ; cela serait le cas dans le cas de l'établissement d'un benchmark ou d'une étude en situation réelle. Nous conservons les trois métriques dans un soucis d'illustration de la manière de réaliser des choix avec le protocole STI, et de montrer qu'elle permet de trouver des métriques d'impact sous différentes contraintes. Toutes les métriques proposées se reposent néanmoins sur l'utilisation de la métadonnée géographique adjointe à une donnée. Cette métadonnée est nécessaire à l'évaluation du biais géographique, ce qui souligne la difficulté de caractérisation de ce biais. Au final, l'utilisation du protocole STI apporte un lot d'informations et de précisions supplémentaires, et permet particulièrement de situer les efforts réalisés dans des cadres bien spécifiques. En précisant la source et le type de biais, on fixe le cadre de l'étude et on limite les variables possibles. On peut alors se concentrer sur l'évaluation, évaluant comment cette évaluation est influencée par le glissement identifié et en proposant, après un processus de recherche et d'exploration, des métriques d'impact appropriées aux source et type sélectionnés. L'utilisation du protocole STI est plus coûteuse temporellement qu'une caractérisation plus simple, mais elle apporte des précisions qui peuvent se relever très importantes pour la compréhension des phénomènes de glissement et la lutte contre les conséquences de ces derniers.

4.6 Conclusion

Nous implémentons dans cette section le protocole STI pour le glissement géographique, et appliquons cette implémentation sur des bases de données à biais géographique synthétique que nous avons développé en nous inspirant des transformations de MNIST. Cette implémentation en premier lieu permet d'avoir une vue d'ensemble du biais géographique, en explorant à travers les différents travaux de recherche les sources et types de biais géographiques, avec comme produit final la figure 23 qui permet l'identification d'une source pour un biais géographique. En second lieu, on y explore comment le biais géographique peut embrasser les différents types identifiés dans le protocole STI, et donc comment ce dernier peut se glisser dans un système d'IA ; informant sur les possibles incursions du biais aux différents niveaux de développement du système. On y étudie ensuite comment témoigner de l'impact du glissement géographique en fonction du type de biais sélectionné, soulignant l'importance du choix justifié d'une métrique d'impact.

Pour valider cette implémentation du protocole STI, nous développons une nouvelle base de données, GeoMNIST, qui est la première base de données à biais géographique à caractère synthétique. Cette base de donnée est déclinable à l'infini en choisissant différents paramètres de transformations et de répartition de données, et nous construisons trois répartitions particulières : une répartition uniforme (GeoMNIST-A), une répartition suivant la répartition mondiale de la population humaine (GeoMNIST-B) et une répartition suivant celle des données dans les bases de données d'images (GeoMNIST-C). Ces trois bases de données sont ensuite analysées pour s'assurer de leur intérêt dans notre processus de validation. Deux expérimentations sont ensuite réalisées à partir de ces bases de données et de modèles entraînés sur ces bases de données : une caractérisation avec et une caractérisation sans le protocole STI. La validation du protocole STI est apportée par la différence apportée par cette méthodologie dans les caractérisations effectuées dans ces deux expérimentations.

Nous implémentons puis testons donc l'implémentation du protocole STI sur des données synthétiques, et observons un intérêt à utiliser cette implémentation. Bien que nécessitant un investissement plus important dans la démarche de l'évaluation du glissement concerné, ce protocole permet une meilleure compréhension du glissement en précisant ses caractéristiques, apportant ainsi des clés pour lutter contre ses conséquences.

Chapter 5

Expérimentation sur données réelles

Chapter Summary

Résumé du chapitre

Dans ce chapitre, nous déployons le protocole STI pour la caractérisation du biais occidental dans les modèles de vision sur des données réelles. Nous commençons par présenter le protocole expérimental et les hypothèses sous-jacentes en section 5.1. La méthodologie utilisée, de la sélection des bases des données et modèles à l'implémentation du protocole STI, est décrite en section 5.2. L'expérimentation est ensuite conduite en section 5.3, conduisant à la présentation de divers résultats allant à l'encontre des hypothèses formulées et de l'état de l'art. Les discussions de la section 5.4 analysent ces résultats et font le constat de la nécessité d'analyser plus en détail les prédictions obtenues. Nous concluons en section 5.5 en constatant les faiblesses du protocole STI, qui s'est montré insuffisant pour la tâche identifiée dans l'expérimentation.

5.1 Contexte

Bien que le protocole STI fonctionne sur les tests précédents, son intérêt ne peut être démontré que sur des cas d'utilisation pratique, sur des données réelles. Nous abordons maintenant la problématique de la caractérisation du biais géographique dans des données images réelles. La complexité d'une telle caractérisation étant illustrée dans la section précédente, nous restreindrons ici la caractérisation du

biais géographique à une opposition entre données occidentales et données non occidentales, telles que définies dans le premier chapitre. Pour ce faire, nous commençons par reprendre l'état de l'art en caractérisation de biais géographique, avant d'analyser sur quelles hypothèses se reposent les précédentes expérimentations et comment ces dernières se traduisent dans le protocole STI. Enfin, nous précisons et formalisons les cadres de l'expérimentation de cette section.

5.1.1 Etat de l'art

De nombreuses publications ont démontré comment le transfert en condition opérationnelle de systèmes d'IA pouvait être problématique si certains biais n'étaient pas pris en compte[35]. DeVries et al. (2019) et Shankar et al. (2017) expliquent que ce phénomène émerge lors d'un glissement géographique, en montrant que les performances de modèles de classification sont moindres pour des données situées dans des pays non représentés dans les bases de données d'images classiques. Ils utilisent pour faire ce constat des représentations graphiques ; une carte de performance par pays et un graphe de performance par revenu pour DeVries et al. (2019), des courbes de densité de log-vraisemblance pour des classes choisies par Shankar et al. (2017). Ces travaux exploratoires, bien que permettant d'illustrer les conséquences du biais géographique dans les systèmes de vision ne permettent pas de caractériser précisément ce biais.

Le biais géographique est abordé dans de multiples travaux et bases de données[134; 88; 11; 65; 51]. Il s'agit de définir des stratégies et des architectures capables de réduire l'écart de performance entre les différents domaines considérés, et non pas de caractériser le biais géographique en lui-même. Cette caractérisation est nécessaire pour éviter l'écueil des déploiements non fructueux dus à une mauvaise prise en compte des spécificités de l'application cible et du contexte de son déploiement[67]. Il est aujourd'hui majeur de considérer cette caractérisation, et de combiner les différentes approches entreprises pour avancer une meilleure prise en compte du biais géographique dans les systèmes de vision.

5.1.2 Hypothèses

L'application du protocole STI nécessite l'identification d'un biais et la mise à disposition d'un moyen pour mettre ce biais en évidence. Loin d'être des faits, dans le cas du glissement géographique, nous identifions quelles sont les hypothèses concernant les données et les modèles qui sont réalisées a priori dans les travaux à l'état de l'art :

- **Hypothèse 1** : il existe des bases de données d'images possédant des informations sur leurs origines géographique, avec une diversité plus grande que les bases de données d'images génériques. Sans métadonnée géographique, il est compliqué de réussir à estimer le biais géographique d'un modèle. Sans des bases de données à forte diversité géographique, ce biais géographique peut être observé (comme on l'a vu dans les expériences de la section précédente), mais à l'aide de métriques particulières.

- **Hypothèse 2** : Les systèmes de classification d'image génériques sont soumis à un biais géographique lorsqu'ils sont exposés à des bases de données inclusives[1]. Nous faisons ici l'hypothèse que les systèmes de classification sont entraînés sur des données non inclusives, et que les processus d'entraînement ne prennent pas en compte ce biais géographique pour le pallier. C'est une hypothèse s'appuyant sur les travaux de DeVries et al. (2019) et Shankar et al. (2017), et qui suppose que les modèles utilisés ne sont pas issus de méthodologies utilisant des techniques de DA ou DG.

- **Hypothèse 3** : Les métriques de performance usuelles d'un système de classification peuvent permettre de rendre compte du biais géographique. On doit donc pouvoir observer le biais géographique sans avoir à concevoir de métrique spécifique, ce qui simplifie le processus de

[1] Revoir la définition d'une donnée inclusive en chapitre 1

caractérisation du glissement géographique. Cette hypothèse est aussi appuyée par les travaux de DeVries et al. (2019).

- **Hypothèse 4** : Le biais géographique intervient dans les systèmes de classification d'image génériques utilisés sur des bases de données inclusive a un impact sur les métriques de performance utilisées. C'est l'hypothèse la plus forte, qui nécessite que les trois hypothèses précédentes soient vérifiées.

Ces quatre hypothèses appuyées par les travaux précédents peuvent être traduites dans le cadre du protocole STI, en utilisant les différents cadre d'identification de types de biais :

- La première hypothèse suggère qu'il existe des bases de données dont on a en partie mitigé les caractères historique, de représentation et de mesurement des biais géographiques. Cette hypothèse assure donc en parallèle la présence de biais géographique relatif aux données dans les bases de données d'images.
- La deuxième hypothèse soutient que le biais géographique est présent dans les systèmes classiques de classification d'image, au moins sous la forme de biais algorithmique.
- La troisième hypothèse affirme que le biais géographique qui survient a un impact directement sur la métrique de performance du modèle, lors de l'évaluation ou du déploiement de ce dernier. Elle nous guide donc vers un choix particulier dans la détermination de l'impact du biais dans le protocole STI.

Cette traduction éclaire les choix qui seront réalisés pour l'identification d'une source, d'un type et de l'impact dans l'implémentation du protocole STI.

5.1.3 Tâche et notations

La caractérisation du biais géographique est une tâche complexe, comme nous l'avons illustré dans le chapitre 4. Nous faisons le choix de réduire cette complexité en réduisant le champ du biais considéré dans cette expérimentation. Nous procédons ainsi à une mise en situation du protocole STI pour l'évaluation du caractère occidental de systèmes de vision par ordinateur. Nous identifions ainsi deux domaines socio-géographique d'intérêt : le domaine occidental et le domaine non occidental. Cette réduction à deux domaines d'intérêt permet de réduire la complexité du témoignage et de simplifier la tâche à réaliser. Plus particulièrement, nous portons notre attention sur la tâche de la reconnaissance d'objets communs dans des images. Ce choix se justifie par l'apport théorique plus fort du choix d'objets communs par rapport à une application spécifique, et de la limitation possible due au contexte particulier de l'application[158]. Nous situons l'intérêt de nos expérimentations dans les pays du Sud, et ne considérerons donc que des algorithmes et modèles nécessitant un minimum de capacité de calcul ou d'investissement financiers, en regard des conditions de déploiement des modèles sur le terrain[27; 45]. Les choix de la méthodologie refléteront ces considérations particulières.

Nous formalisons la classification multi-classe avec comme espace d'entrée $X \subset \mathbb{R}^d$ et comme espace de sortie $Y \subset \{1, \cdots, k\}$. A chaque donnée (x, y) définie sur $\mathcal{X} \times \mathcal{Y}$, on peut associer un contexte $c \in \mathcal{C}$ qui inclue tous les aspects sociaux, culturels et techniques associés à la donnée. Une base de données $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ est échantillonnée à partir d'une distribution spécifique P_S définie sur $\mathcal{X} \times \mathcal{Y}$ et est sujette au contexte C_S, qui peut lui-même être catégorisé par des sous-contextes $C_S = \{C_{S_1}, \cdots, C_{S_n}\}$. Un glissement de donnée survient quand un modèle M_S entraîné sur la base de données S est utilisé sur le dataset T sujette au contexte C_T avec $C_S \neq C_T$. Les performances d'un tel modèle sur son jeu de donnée d'entraînement $R_S(M_S) = \mathbb{E}_{z \sim \mathbb{P}_S} l(z, M_S)$ sont meilleures en général que ses performances sur le jeu de donnée cible $R_T(M_S) = \mathbb{E}_{z \sim \mathbb{P}_T} l(z, M_S)$ avec $z = (x, y)$, et l une fonction de perte. Des cas particuliers peuvent subvenir, comme des jeux de données cibles plus "faciles" que les données d'entraînement du modèle, ou encore une répartition particulière des classes ou des domaines qui avantagent les performances du modèle sur les données cibles.

Pour souligner le caractère occidental d'un modèle, on peut assigner à une base de données son contexte C_S qui peut lui-même être catégorisé par des sous-contextes spécifique à la caractérisation du caractère occidental $C_S = C_{OCC} \cup C_{\overline{OCC}}$. Ces contextes témoignent des caractères occidentaux et non occidentaux d'une base de données. La performance d'un modèle M sur la base de données S est alors $R_S(M) = \mathbb{E}_{z\sim\mathbb{P}_S} l(z, M) = \mathbb{E}_{z\sim\mathbb{P}_{S_{OCC}}} l(z, M) + \mathbb{E}_{z\sim\mathbb{P}_{S_{\overline{OCC}}}} l(z, M)$ avec $z = (x, y)$, S_{OCC} l'ensemble des données de S associé au contexte C_{OCC} (et réciproquement pour $S_{\overline{OCC}}$) et l la fonction de perte associée à la métrique de performance choisie.

5.2 Méthodologie

Nous commençons ici par déterminer quelles bases de données et modèles peuvent être considérés comme candidats avant d'implémenter le protocole STI et de réaliser la sélection finale parmi les candidats. La mise en place de cette méthodologie doit respecter les hypothèses précédemment posées et proposer une expérience permettant de caractériser l'occidentalité des systèmes de vision qui seront sélectionnés.

5.2.1 Bases de données

Nous considérons comme base de données candidate les bases de données d'images contenant des métadonnées géographiques ou culturelles, annotées pour la tâche de classification multiclasse. Des candidates potentielles peuvent être trouvées dans différents benchmarks, comme WILDS[92], avec iWildCam2020-wilds, Camelyon17-wilds et PovertyMap-wilds, qui possèdent des images de nombreux pays et localités. DomainBed[69] propose également un benchmark pour le glissement de distribution, utilisant plusieurs bases de données dont TerraIncognita, qui comprend des métadonnées géographiques. Ces bases de données reflètent des applications de vision par ordinateur sur le terrain, et sont donc spécifiques à ces applications, qui concernent l'analyse géographique de la pauvreté, l'analyse de cellules, ou l'identification d'animaux sauvages. Pour conserver une vue d'ensemble sur la vision par ordinateur, nous préférons considérer d'autres bases de données constituées d'images plus communes, comme Dollar Street Dataset[65] (DSD), COCO World URLs[18] (COCOWU), Geo-YFCC[51] (GYFCC), GeoImNet (GIN) issu du benchmark GeoNet[88] ou encore MaRVL[107]. Ces bases de données d'images contiennent des informations sur des concepts communs provenant de différentes localités, et peuvent donc correspondre à nos besoins. Elles sont listées dans le tableau 8.

Table 8: Comparaison des bases de données candidates pour l'expérimentation.

BDD	# données	# domaines	Métadonnée géographique	# classes	# domaines occidentaux	# données occidentales
DSD	38 479	63	photos des maisonnées	289	15	7 594
COCOWU	9 200	23*	recherche par mots-clé	80	6*	2400
GYFCC	1 147 059	62	localisation par geo-tag	1 261	28	529 690
GIN	250 050	2**	localisation par geo-tag	600	1	171 692
MaRVL	4914	5***	recherche par mot clé en différentes langues	429	0	0

*Les 249 pays dont sont issues les données de COCOWU sont répartis dans 23 zones géographiques, avec 6 zones considérées occidentales.
**Les données de GIN sont issues de deux zones géographiques, les USA et l'Asie. Le nombre de pays dans chacune de ces zones n'est pas précisé.
*** Les données de MaRVL proviennent de recherches sur tous les médias confondus en 5 langues différentes.

Dollar Street Dataset (DSD) est une base de données publiée par MLCommons, bien qu'issue de l'initiative de la communauté Gapminder[2]. Cette base de données contient des photographies prises avec l'intention de montrer la vie et les objets communs de familles issues du monde entier. Ces données proviennent ainsi de 63 pays, et comprennent des photos d'objets du quotidien décidés par les équipes de GapMinder. Cette base de données totalise 38 479 images issues de 404 foyers et annotées avec 289 classes. En plus du pays d'origine, les métadonnées contiennent le revenu mensuel de consommation de la famille photographiée en dollars ajustés à la parité de pouvoir d'achat (PPP).

[2]https://www.gapminder.org/

Après la première étape de nettoyage de la base de données par MLCommons, DeVries et al. (2019) décrit une seconde opération de nettoyage pour retirer les classes abstraites, comme "most loved item", jugées inadéquates pour la tâche à réaliser. Nous procédons de même et obtenons une base de données finale composée de 27 418 images issues de 23 pays et annotées avec 115 différentes classes. Cette base de données d'images n'a pas de partition officielle en jeux d'entrainement, de validation ou de test. Des exemples de données de DSD sont fournies en figure 38.

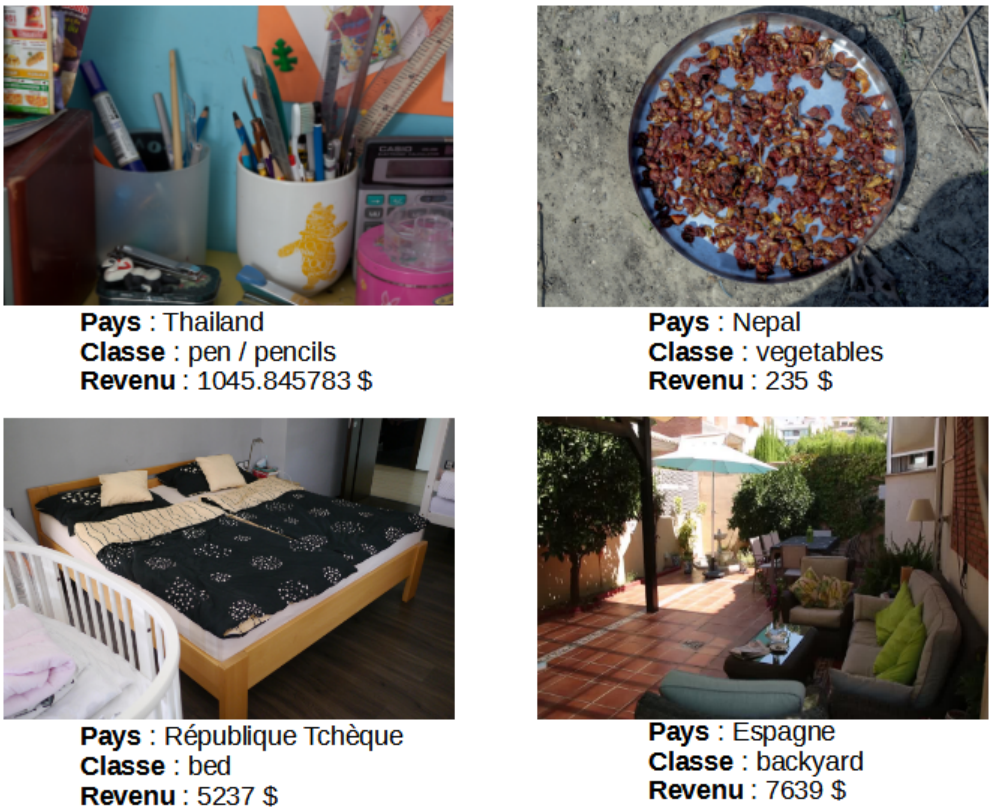


Figure 38: Illustration des données contenues dans Dollar Street Dataset. Les métadonnés associées à une image comprennent les classes associées, les revenus du foyer et le pays dont l'image est issue.

Cette base de données possède des métadonnées de très grande précision, puisque les données ont été prises directement auprès de foyers particuliers dans des zones spécifiques, et que ces informations sont conservées avec les données. Elle possède un volume moyen de données, et de nombreuses classes. Le côté abstrait de certaines des classes et la possibilité d'un manque d'annotation de certaines données sont des possibles écueils à l'utilisation de cette base de donnée.

COCO World URLs (COCOWU) est une base de données publiée par nos soins durant cette thèse. Nous avons répliqué la méthode de collecte de données de MS COCO[106] et l'avons modifié afin d'inclure des images inclusives pour chacune des 23 zones géographiques identifiées par la norme M49 des Nations Unies. Ces données sont issues de la base d'images Flickr, et 400 d'entre elles sont annotées avec les 80 classes de MS COCO pour chacune des zones géographiques. Des exemples de ces données sont fournies en figure 39.

Zone Geographique :
Northern Europe
Classe : bottle, cup, fork, knife, spoon, bowl, plate

Zone Geographique :
Micronesia
Classe : person, hat, shoe

Zone Geographique :
Eastern Africa
Classe : person, backpack, hat, shoe, bottle, chair

Zone Geographique :
Central America
Classe : person, shoe, desk, chair, door,

Figure 39: Illustration des données contenues dans COCO World URLs. Les métadonnées associées à une image comprennent la zone géographique associée à la donnée et les classes apparaissant à l'image.

Les métadonnées géographiques étant induites par les recherches sur la plateforme Flickr, elles sont considérées de faible qualité, car rien ne prouve qu'une donnée récupérée avec un mot clé est bien lié à ce mot clé. Cette base de données d'images n'a pas de partition officielle en jeux d'entrainement, de validation ou de test. L'intérêt de cette base de données est sa compatibilité avec les modèles entraînés sur la base de données MS COCO et sa grande représentativité géographique puisqu'elle couvre l'ensemble du globe.

Geo-YFCC (GYFCC) est un sous produit de la très large base de données YFCC100M[155], comprenant plus de 100 millions d'images. Les auteurs de GYFCC ont extrait les images de YFCC100M possédant un géo-tag, qui est une position (latitude, longitude) associée à la donnée. Il associent à chaque donnée le pays correspondant au géo-tag, et sélectionnent ensuite les pays possédant plus de 10 000 données et restreignent chaque pays à un maximum de 20 000 images. Cette base de données est au final composée de 1 147 059 images issues de 62 pays, annotées avec 1 261 classes issues des 4 000 catégories d'Imagenet-5K qui ne sont pas présentes dans ILSRVC12. Ils partitionnent ensuite ces données en 45 pays pour l'entraînement, 7 pays pour la validation et 15 pays pour le jeu test, en retenant pour chaque pays 3000 données pour la création d'un jeu de test par pays. La figure 40 illustre les données composant GYFCC.

Figure 40: Illustration des données contenues dans GYFCC. (figure extraite de Dubey et al. (2021)).

La précision des métadonnées obtenues par geo-tag est forte, même si ces données peuvent encore être biaisées par une contextualisation particulière (comme par exemple des photos de touristes plutôt que de personnes locales).Cette base de données est opportune pour la caractérisation du biais géographique en raison de la qualité des métadonnées, et de son volume important de données.

GeoImNet (GIN) est issue du benchmark GeoNet proposé par Kalluri et al. (2023) pour l'adaptation géographique, composée d'images en provenance des USA et d'Asie. Toutes les images de cette base sont issues de la base de données WebVision[99], et donc de la plateforme Flickr, à l'aide de requêtes construites à partir des classes d'Imagenet-5K. Les coordonnées GPS sont extraites à partir des Flickr-id associées aux données. Une phase supplémentaire de sélection de classe permet d'extraire 600 classes, avant de séparer la base de donnée en jeu d'entraînement (85% des données) et jeu test (15% des données). GIN est composé de 171 692 images issues des USA et de 78 358 images issues d'Asie. Ces images sont illustrées en figure 41.

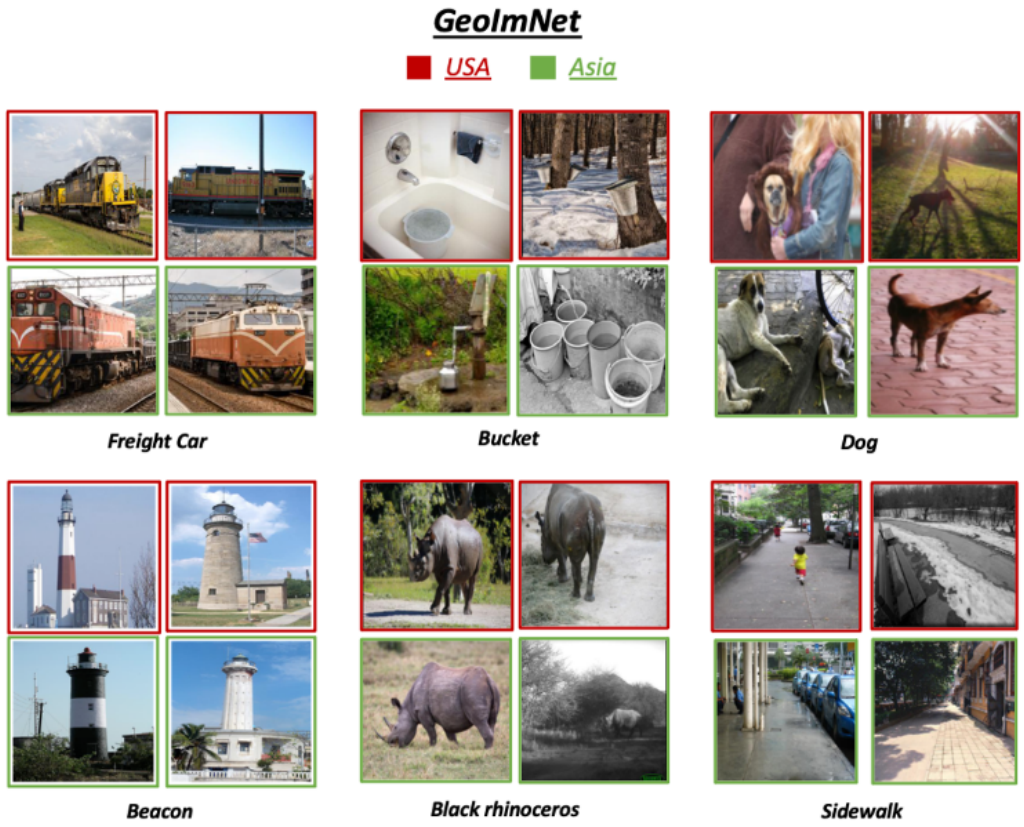


Figure 41: Illustration des données contenues dans GIN. Image issue de la publication originale de la base de données[88].

L'avantage de cette base de données est de proposer un important volume de données avec une bonne qualité de métadonnée, puisque les géo-tag sont considérés comme précis. Les désavantages

de cette base de données sont l'utilisation de Flickr et de requête par mot-clé, comportant son lot de biais dans la collection des images, ainsi que le nombre limité de domaines.

MaRVL est une base de données mêlant images et textes, associant à un couple d'images une description textuelle de ces images. Pensée pour représenter des cultures non représentées dans les précédentes bases d'images, elle est construire à partir de 5 langues : l'indonésien, le mandarin chinois, le swahili, le tamil et le turque. Toutes les étapes de la conception de la base de données (sélection des concepts, sélection d'images candidates, nettoyage des images, annotation) sont repensés pour être désignés de manière inclusive. Il en résulte une base de données composée de 4914 images, annotées avec des textes présentant ces dernières dans 5 langues différentes, centrés sur 429 concepts. Une illustration de ces données est présentée en figure 42.

(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுப்பட்டிருப்பதை காணமுடிகிறது. ("In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming."). Label: TRUE.

(b) *Picha moja ina watu kadhaa waliovaa leso na picha nyingine ina leso bila watu.* ("One picture contains several people wearing handkerchiefs and another picture has a handkerchief without people."). Label: FALSE.

Figure 42: Illustration des données contenues dans MaRVL. (a) Images issues de la culture Tamil, liées au concept "Jallikattu", partie d'une festivité indienne. (b) Images issues de la culture Swahili, liées au contexte "leso", ou mouchoir en français. Illustration reprise de la publication originale de la base de données[107].

Cette base de données contient des métadonnées d'une très grande qualité, puisque directement annotées par des personnes natives des cultures considérées. Néanmoins, le faible volume de données et l'annotation par description textuelle plutôt que par classes sont des faiblesses en comparaison aux autres bases de données proposées. L'absence de données occidentale, par choix de design, est aussi problématique pour l'estimation du caractère occidental d'un modèle.

5.2.2 Les systèmes de vision

Les systèmes de vision prennent de multiples formes et obéissent à diverses contraintes. Ces contraintes sont généralement d'ordre financière, computationnelle, et énergétique. Du côté financier, le coût d'utilisation des API ou l'investissement dans du matériel coûteux peuvent être des freins important à l'utilisation ou au déploiement de certaines solutions. Sur le plan computationnel, la nécessité

d'entraîner ou de peaufiner les modèles peut être un mur infranchissable dans des situations où les ressources calculatoires sont maigres. Enfin, l'accessibilité à des sources d'énergies fiables et abondantes peut être une condition difficile à atteindre dans certaines zones du monde. Ces contraintes se combinent généralement dans les contextes du Sud, limitant les possibilités d'utilisation de bon nombre de solutions en vision. Par exemple, l'emploi de modèles fondateurs est encore complexe ; de par leur nature extrêmement grande, il est nécessaire de posséder d'importantes ressources calculatoires et énergétiques pour les faire tourner en local. A l'opposé, l'utilisation de ces dernières par API requiert des ressources financières parfois importantes. La nécessité de disposer en plus de personnel capable d'adapter ces modèles fondateurs à des applications spécifiques rend ces dernier difficile d'utilisation dans de nombreux cas d'application au Sud.

Pour l'expérience de la caractérisation de l'occidentalité des modèles, nous désignons en premier lieu comme modèles candidats les modèles validant des critères de sobriété en terme d'investissement financiers, computationnels et énergétiques, pour mieux s'inscrire dans les problématiques du Sud. Nous ne considérerons donc pas les modèles fondateurs et autres systèmes de vision-langages qui se développent beaucoup ces derniers temps. Ces derniers nécessitent des ressources importantes et une expertise particulière pour pouvoir être utilisés, des pré-requis qui se montrent difficile à concevoir dans un laboratoire de recherche au Sud. De même, les API des grand groupes comme Amazon, Google, IBM, Microsoft ou Clarifai utilisés dans les travaux de DeVries et al. (2019) ne seront pas considérés comme candidats ici, à cause de leur coût financier important. Nous sélectionnerons un algorithme à l'état de l'art en classification multi-classe, CSRA[177]. Cet algorithme a l'avantage d'intégrer une architecture de transformer, d'être à l'état de l'art en terme de performance pour la tâche de classification multi-classe, et de disposer en ligne des poids correspondant à l'entraînement d'un modèle à partir de cet algorithme sur la base de données MS COCO, réduisant la charge computationnelle nécessaire pour l'emploi de ce dernier. Des expérimentations futures pourront élargir le champ de sélection de modèles en se basant sur un autre ensemble de contraintes.

5.2.3 Implémentation du protocole STI

La tâche à réaliser dans cette expérimentation est la caractérisation de l'occidentalité du modèle sélectionné. Le protocole STI suppose, avant toute chose, l'identification des trois éléments nécessaires à cerner le biais en question : la source, le type, et l'impact du biais. Cette section implémente la sélection de ces trois éléments.

5.2.3.1 Identifications de la source et du type de biais

Nous avons vu au chapitre 4 que la source du biais géographique est multiple, et dépend moins du contexte que des choix et considérations des expérimentations réalisées. Dans notre cas, nous utilisons un modèle entraîné sur la base de données MS COCO. Les données d'entraînement du modèle possèdent une distribution géographique non uniforme et certaines zones géographiques sont moins représentées que d'autres. C'est la source sélectionnée dans le cadre de l'implémentation du protocole STI : la distribution géographique inégalitaire de MS COCO, ayant servi de base de données d'entraînement pour le modèle utilisé dans l'expérimentation.

Du point de vue du modèle, cela induit un biais d'apprentissage et un biais d'évaluation puisque ce dernier est entraîné et évalué sur des distributions biaisées. Dans le cadre de la caractérisation de l'occidentalité du modèle CSRA, nous caractériserons comme dans les expériences précédentes le biais d'évaluation. Le second critère du protocole STI est donc défini, et il ne reste que la métrique d'évaluation à sélectionner. La source de ce biais étant la distribution géographique non uniforme des données d'entraînement, on s'attend à ce que les performances de CSRA soient atténuées sur des bases de données inclusives. C'est en tout cas une tendance démontrée par les travaux précédents en classification mono-classe grâce à une évaluation manuelle.

5.2.3.2 Évaluation de l'impact du biais

Nous cherchons à déterminer l'impact du caractère occidental du modèle CSRA entraîné sur MS COCO dans le cadre d'un biais d'évaluation issu d'une distribution non uniforme des données d'entraînement du modèle. Les données des bases de données sélectionnées pour l'expérimentation seront réparties en deux domaines, en fonction de leur provenance occidental ou hors occident. Une métrique du caractère occidental du modèle sera donc la différence de performance entre données à provenance occidentale et données provenant hors de l'occident. Une telle métrique est pratique, car simple, numérique, facilement calculable, interprétable. Le dernier critère à sélectionner pour l'implémentation du protocole STI est donc identifié, et après avoir déterminé comment générer cette métrique, nous pourrons passer à la mise en place de l'expérimentation.

Pour obtenir cette métrique, on peut se baser sur les métriques automatiques utilisées en classification multi-classe, comme la précision top-1 ou la précision top-5. DeVries et al. (2019) utilisent pour leur part un outil d'évaluation manuelle, qui permet l'évaluation de systèmes de vision ne partageant pas forcément les mêmes ensembles de classe. Nous proposerons des analyses en même temps automatiques et manuelles, la comparaison des deux étant intéressante pour la compréhension de certains biais présents dans les processus des deux types d'évaluation. Nous proposerons comme métrique utilisée la performance moyenne par classe, permettant d'éviter les écueils des métriques de précision top-1 ou top-5.

5.2.4 Outil d'évaluation manuel des prédictions

Inspiré de l'outil développé par DeVries et al. (2019) pour l'évaluation manuelle des prédictions d'un modèle, un outil simple est conçu pour l'évaluation de prédictions sur des données images, en faisant défiler des données annotées de leurs annotations et prédictions. Cet outil parcours une base de données à partir d'une classe sélectionnée, permettant à l'observateur d'interpréter la relation entre les prédictions d'un modèle et les données liées à la classe considérée.

Chaque donnée image est visualisée dans son entièreté, avec en ajout (en haut à gauche de l'image, dans l'ordre) la source géographique de l'image, la classe sélectionnée, l'annotation liée à l'image, et les prédictions réalisées par le modèle sélectionné. Si la classe sélectionnée est présente en annotation, celle-ci est écrite en verte ; en rouge autrement (et réciproquement pour les prédictions). L'interface est illustrée en figure 43.

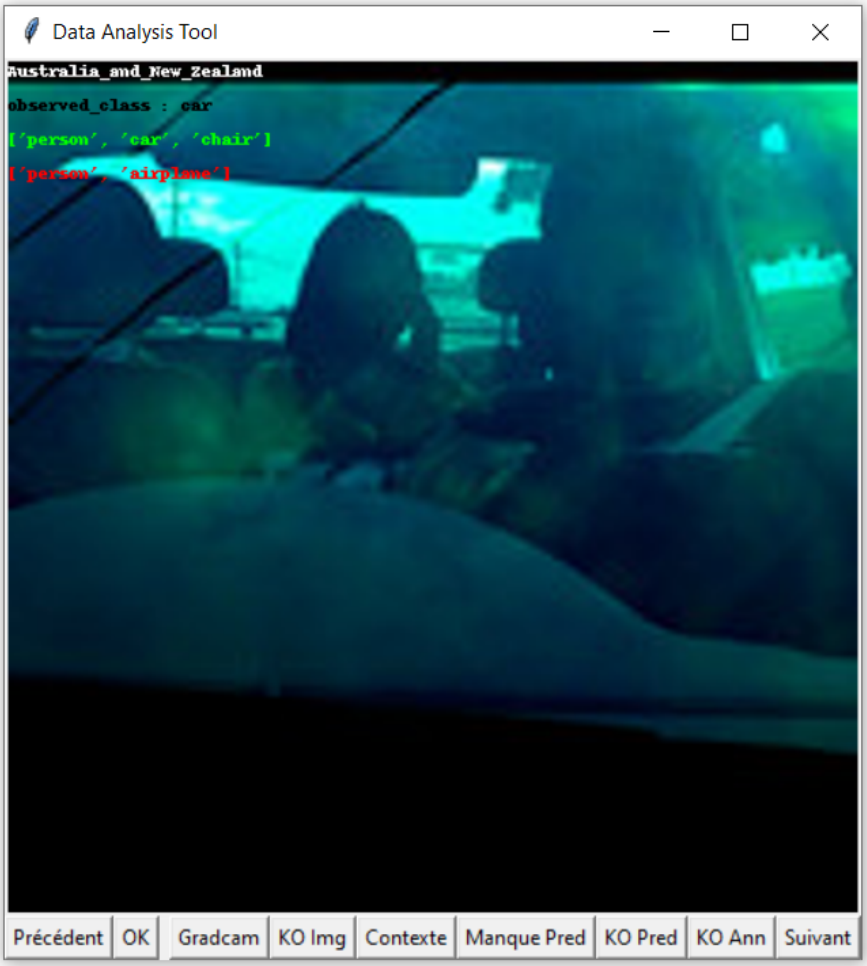


Figure 43: Interface de l'outil d'évaluation présentée à l'opérateur et développé par mes soins. De multiple boutons permettent de naviguer et d'annoter les données visualisées ainsi que les annotations et prédictions pour chacune de ces données.

Les boutons en dessous de l'image permettent de naviguer entre les données et de lier des observations à ces dernières. Les boutons "Précédent" et "Suivant" permettent simplement de parcourir les images. Les autres boutons servent à l'évaluation des données, de la manière suivante :

- si la prédiction du modèle est bonne (la classe est prédite et apparaît à l'image), l'opérateur presse le bouton "OK".
- si la prédiction du modèle est mauvaise (la classe est prédite par le modèle mais n'apparaît pas à l'image), l'opérateur presse le bouton "KO Pred".
- si le modèle n'a pas prédit la classe alors qu'elle est à l'image, l'opérateur presse le bouton "Manque Pred".
- si l'image ne contient pas la classe sélectionnée mais est annotée comme telle, l'opérateur presse le bouton "KO Ann".
- si l'image ne devrait pas être dans la base de données (donnée offensante, image ouvertement non réelle, ...), l'opérateur presse le bouton "KO Img".
- si une variation de concept entre annotation et prédiction est détectée (par exemple, une image d'une maquette de train ou d'un jouet, qui déclenche la prédiction d'un train à l'image mais n'est pas annotée comme telle), l'opérateur presse le bouton "Contexte".

- dans le cas où l'opérateur cherche à mieux comprendre la raison derrière une prédiction réalisée par le modèle, il peut presser le bouton "Gradcam". Cela ajoutera à la donnée réalisée un filtre utilisant la technique SmoothGradcam[124] pour comprendre quelles zones de l'image sont utilisées pour la prédiction de la classe sélectionnée par le modèle. SmoothGradcam nécessite l'accès aux couches internes du modèle, et l'utilisation de cette fonctionnalité n'est donc disponible que dans le cas où les couches du modèle sélectionné sont accessibles.

Cette interface permet ainsi à l'opérateur d'évaluer les prédictions, annotations, et données, et de produire des analyses statistiques par classe des prédictions pour les bases de données utilisées. Les évaluations réalisées par l'opérateur sont sauvegardées dans des tableurs csv, et différentes métriques de performances telles que la précision, le rappel, ou encore le F1-score sont calculées à partir de ces évaluations manuelles. Le parcours et l'évaluation de données avec cet outil permet donc de générer des statistiques d'évaluation manuelle de performance des modèles, ainsi que des statistiques sur les observations réalisées durant le parcours des données, comme le taux de mauvaises données ou le taux de variations de concept entre annotation et prédictions. L'outil est codé en python, à l'aide de la bibliothèque PIL[3]. Il prend en entrée un tableur csv contenant, pour chaque image, une ID, un chemin vers l'image, les annotations et les prédictions associées à l'image, et sa provenance géographique. Il produit en sortie des tableurs excel contenant un résumé des observations manuelles et des statistiques computées à partir de ces observations. Le code est ouvert et accessible via github[4].

5.3 Évaluation des performances de CSRA

On réalise des évaluations manuelle et automatique des prédictions du modèle CSRA pour caractériser l'occidentalité de ce dernier. Nous commençons par préciser les détails de l'expérience avant de présenter les résultats obtenus. Une discussion permet ensuite d'analyser et de proposer des explications à ces résultats.

5.3.1 Expérience

Pour caractériser l'occidentalité du modèle CSRA, il est nécessaire de se reposer sur une des bases de données candidates présentée plus haut. Parmi les cinq bases de données candidates, nous choisissons d'utiliser COCOWU pour les raisons suivantes :

- l'annotation manuelle est un processus chronophage. Les bases de données à large volume sont donc proscrites à cause du temps nécessaire à l'évaluation des performances du modèle.

- l'utilisation d'un ensemble de poids correspondant à l'entraînement de CSRA sur MS COCO comme paramétrisation du modèle permet de partager le même ensemble de classe entre CSRA et COCOWU. La sélection d'une autre base de données introduit un glissement d'ensemble de classes, et donc d'autres biais et une complexité supplémentaire dans le processus d'évaluation.

L'ensemble de la base de données COCOWU sera analysée, classe par classe, et les prédictions du modèle CSRA seront comparées aux annotations réalisées sur COCOWU. Les performances seront reportées sur les deux différents domaines identifiés, les données considérées occidentales et les données considérées hors occident. Un tableau présentant ces deux résultats et la différence de performance du modèle entre les deux domaines permettra de caractériser l'occidentalité du modèle. Ce tableau permettra aussi de comparer les performances obtenues entre évaluation manuelle et évaluation automatique.

Après cette évaluation générale, nous nous concentrerons sur certaines classes sélectionnées pour leur intérêt particulier. Ainsi, les classes associées au transport dans COCOWU ("car", "motorcycle", "airplane", "bus", "train", "truck", "boat") seront ensuite analysées plus en profondeur, puisqu'elles

[3] https://he-arc.github.io/livre-python/pillow/index.html
[4] https://github.com/tbayetird/WCBENCH

représentent une bonne partie de la base de données et combinent des caractéristiques intéressantes qui seront explicitées durant les expérimentations.

Afin de pouvoir proposer un parallèle avec une autre base de données et de proposer des résultats indépendants de biais présents dans la base de données COCOWU, les expérimentations seront prolongées sur une autre base de donnée, DSD. Ce choix repose sur un rapprochement avec les travaux de DeVries et al. (2019) et sur un partage de certaines classes associées au transport. Ainsi, les classes associées au transport dans DSD ("bicycle", "car", "motorcycle") seront aussi analysées pour les prédictions du modèle CSRA. Ces observations permettent de comparer les résultats obtenus sur la base de données COCOWU et de pérenniser les observations réalisées sur le modèle.

5.3.2 Résultats

Les résultats des évaluations générales et approfondies sur certaines classes des prédictions du modèle CSRA sur les bases de données COCOWU et DSD sont présentés dans cette section. Ces derniers sont rapidement commentés ici et plus amplement discutés dans la section suivante.

5.3.2.1 Prédictions de CSRA sur COCOWU

Le tableau 9 présente une comparaison des évaluations manuelles et automatiques des prédictions de CSRA sur la base de données COCOWU.

Table 9: Comparaison des évaluations manuelles et automatiques des prédictions du modèle CSRA sur la base de données COCOWU pour la métrique de précision moyenne par classe. Les données de COCOWU sont séparées entre les données occidentales (ED) et les données non occidentales (HD).

mode	ED	HD	diff
Automatic	0.3416	0.3643	-0.0227
Manual	0.5526	0.5860	-0.354

Les évaluations manuelles et automatiques ont des différences notables en terme de performance, et ce phénomène est en partie du à la prise en compte des mauvaises annotations dans les évaluations manuelles. Ces erreurs d'annotations sont particulièrement présentent dans la classe "motorcycle", avec 58 itérations d'une bonne prédiction du modèle alors que l'annotation manque sur la donnée, étant donc comptée comme faux positif dans le cas de l'annotation automatique. La différence de performance évaluée manuellement entre données ED et HD n'est pas négligeable comme dans le cas de l'évaluation automatique, soulignant une meilleure performance du modèle sur les données non occidentales.

Les évaluations automatiques et manuelles proposent des résultats allant à contre courant de l'état de l'art : elles attestent d'une meilleure performance du modèle CSRA sur les données non occidentales de COCOWU comparée à la performance du même modèle sur les données occidentales. Ce résultat est surprenant, et nous pousse à réaliser des analyses plus poussées, afin d'être en capacité d'expliquer pourquoi un modèle peut paraître plus performant sur des données considérées hors distribution. Nous analysons donc les performances classe par classe pour un ensemble de classe qui est très présent dans COCOWU : les classes de transport.

5.3.2.2 Prédictions de CSRA sur les classes de transport de COCOWU

Une analyse plus poussée, classe par classe, permettra d'évaluer si ces tendances se retrouvent uniformément dans chacune des classes ou si elles sont portées par quelques classes spécifiques. Nous examinons les classes liées au transport ("car", "motorcycle", "airplane", "bus", "train", "truck", "boat") dans le tableau 10. Ces classes sont sélectionnées pour leur grand nombre d'itérations et leur popularité à travers les différentes bases de données. Ce tableau présente des résultats hétérogènes. Le premier résultat frappant est celui de la classe "motorcycle", qui obtient une précision nulle dans le

cas de l'évaluation automatique, alors qu'une évaluation manuelle donne 67% de précision pour cette classe. La différence est due à une conjonction surprenante entre une mauvaise annotation (58 données comportant la classe n'étant pas annotées comme telles) et des prédictions manquantes sur toutes les données annotées. D'autres résultats sont étonnants, comme les différences de performances de la classe "bus" (témoignant d'une bien meilleure performance sur des données non occidentales) et de la classe "train" (témoignant d'une bien meilleure performance sur des données occidentales). Ces classes seront visuellement explorées dans le chapitre suivant dans le but de proposer des explications à ces résultats.

Table 10: Précision moyenne par classe évaluée manuellement et automatiquement pour les classes de transport de la base de données COCOWU. N est le nombre de données par classe. Les résultats en gras correspondent aux différences les plus importantes positivement et négativement entre performances ED et HD.

classe	N	précision (manuelle)	précision (automatique)	précision ED (automatique)	précision HD (automatique)	diff (automatique)
bicycle	126	0.8043	0.5873	0.6667	0.5591	0.1076
car	563	0.7044	0.6075	0.6765	0.5682	0.1083
motorcycle	88	0.6747	0	0	0	0
airplane	107	0.7851	0.5794	0.5882	0.5753	0.0129
bus	68	0.5933	0.5294	0.2381	0.6596	**-0.4215**
train	148	0.5853	0.4257	0.5862	0.3222	**0.2640**
truck	254	0.2551	0.1693	0.1591	0.1747	-0.0156
boat	357	0.6300	0.4706	0.4348	0.4830	-0.0482

Ces résultats plus poussés illustrent un phénomène étonnant : il y a une forte hétérogénéité dans les résultats en fonction de la classe considérée. Certaines classes présentent de fortes disparités de performances entre occident et non occident, tandis que d'autres présentent des différences négligeables. Dans les disparités importantes, certaines sont positives (à l'avantage des données occidentales) et certaines négatives (à l'avantage des données non occidentales). Nous discutons plus amplement ces résultats dans la section suivante.

Ces résultats surprenants sont-ils le fait de la base de données COCOWU, à des biais particuliers dans la conception de cette base de donnée, ou bien sont-ils le fait des mécanismes internes du modèle utilisé ? Pour explorer cette question, nous élargissons l'expérimentation à la base de données DSD.

5.3.2.3 Prédictions de CSRA sur DSD

L'utilisation de l'outil d'évaluation manuel sur la base de données DSD fournit une première observation marquante : le nombre de données non annotées est étonnamment élevé. Le tableau 11 illustre ce phénomène pour les classes de transport, révélant que jusqu'à 55% des annotations sont manquantes pour la classe "car". Une évaluation automatique des performances d'un modèle ne détecte pas ce phénomène et juge alors la prédiction du modèle mauvaise pour ces données. Ce phénomène est du à la construction de la base de données DSD, initialement mono-classe, avec une collection de donnée orientée vers des objet du quotidien au centre de l'image par GapMinder. Ainsi, d'autres classes peuvent apparaître dans les images sans pour autant être annotées comme telle. Le passage en multi-classe et l'opération de nettoyage de la base de données réalisé par ML Commons ne pallie pas ce problème. Ce phénomène a un impact fort sur l'évaluation automatique de la performance d'un modèle, et peut porter atteinte à la capacité de la méthodologie STI d'évaluer le caractère occidental d'un modèle.

Table 11: Ratio d'annotations manquantes (classe prédite et présentes dans l'image mais absente dans les annotations) pour le modèle CSRA pour les classes de transport de la base de données DSD.

classe	taux d'annotations manquantes
bicycle	0.33
car	0.55
motorcycle	0.37

Nous procédons tout de même à une évaluation manuelle et automatique des classes de transport de la base de données DSD (classes "bicycle", "car", "motorcycle"). En comparant ces résultats avec ceux obtenus sur la base de données COCOWU, nous pourrons analyser l'impact de la base de données sur le témoignage du caractère occidental du modèle.

5.3.2.4 Prédictions de CSRA sur les classes de transport de DSD

Le tableau 12 présente une analyse plus poussée des classes de transport pour la base de données DSD. Ce tableau permet de comparer les performances moyennes manuelles et automatiques pour chacune de ces classes, et les performances occidentales et non occidentales par classe de l'évaluation automatique. Ce tableau souligne une nouvelle fois une grande différence de performance entre évaluation manuelle et automatique, justifiée par des annotations de basse qualité. Les résultats sont pour autant toujours aussi peu consistants dans les tendances entre performance sur données occidentales et non occidentales.

Table 12: Précision moyenne par classe évaluée manuellement et automatiquement pour les classes de transport de la base de données DSD. N est le nombre de données par classe.

classe	N	précision (manuelle)	précision (automatique)	précision ED (automatique)	précision HD (automatique)	diff (automatique)
bicycle	243	0.8723	0.5391	0.6862	0.5000	0.1862
car	285	0.9170	0.3193	0.3627	0.2951	0.0676
motorcycle	183	0.6710	0.4262	0.3846	0.4294	-0.0448

Nous observons dans les résultats de ces expérimentations des tendances hétérogènes, entre appui à l'hypothèse d'un caractère occidental du modèle sélectionné pour certaines classes, et refus de cette hypothèse pour d'autres classes. La section suivante offre une analyse de ces résultats et ouvre la discussion sur de possibles explication des phénomènes observés.

5.4 Discussions

La précision manuelle par classe est, pour le modèle et les bases de données explorées, toujours meilleure que la précision automatique. Les différences entre ces différentes métriques souligne l'importance du nettoyage de données dans les bases de données et d'approches impliquant des opérateurs humains. Ceci plaide en faveur de l'utilisation de bases de données construites avec plus de soin, mais de telles bases de données comprenant des biais géographiques sont aujourd'hui difficiles à trouver.

Pour les deux bases de données et les deux évaluations manuelles et automatiques, certaines classes performent mieux sur des données occidentales tandis que d'autres performent mieux sur des données non occidentales. Ces résultats vont à l'encontre des hypothèses énoncées en début de section qui supposent une meilleure performance sur des données occidentales, et qui sont appuyées par l'état de l'art en caractérisation du biais géographique. Ces résultats sont donc particulièrement

surprenants, et il est nécessaire d'expliquer ces derniers pour comprendre comment les hypothèses sont remises en question.

Les biais et problèmes identifiés dans les évaluations ne suffisent pas à expliquer l'absence de tendance claire sur les performances des données occidentales et non occidentales. Les processus de collecte de données pour COCOWU et DSD étant très différents (récolte d'images via requêtes en ligne sur la plateforme flipper versus photos réalisées par des professionnels ou volontaires dans des maisonnées à travers le monde), ces résultats plaident vers une explication autre que la présence de biais occidentaux dans ce processus. L'inconstance des résultats en fonction du caractère occidental ou non occidental des données pousse à chercher une autre explication pour la performance du modèle. Une première thèse que l'on pourrait avancer pour expliquer ce phénomène est une facilité relative des données en fonction des classes, bien que rien ne puisse justifier de cette facilité relative. Une interprétation visuelle des données et prédictions est nécessaire pour comprendre d'où proviennent les résultats obtenus.

La caractérisation de l'occidentalité du modèle CSRA semble compromise. Les résultats obtenus tendent à exprimer une absence d'un tel caractère occidental, contrairement aux hypothèses et résultats académiques. Pour mieux comprendre les mécanismes à l'oeuvre à l'origine des phénomènes observés, nous procédons dans le chapitre suivant à des interprétations visuelles des données et des prédictions de CSRA sur ces données.

5.5 Conclusion

Le déploiement de la méthodologie STI sur des données réelles, initialement pensé pour caractériser l'occidentalité des systèmes de vision, produit des résultats contradictoires avec la théorie et l'état de l'art. Il est important d'identifier ce qui produit ces résultats, et comment notre méthodologie est différente des précédentes pour pouvoir fournir une explication à ces résultats surprenants. Nous identifions dans la méthodologie certains aspects qui diffèrent de l'état de l'art, notamment la tâche (classification multi-classe) et la métrique de performance. L'impact de ces écarts à l'état de l'art est non mesuré, et des observations supplémentaires sont nécessaires pour estimer leur rôle dans les résultats observés.

Ces expérimentations mettent en lumière une faiblesse du protocole STI : le fait que ce protocole repose sur des hypothèses qui doivent être vérifiées pour être employée, le fait qu'il peut mener à des conclusions surprenantes lorsque utilisé avec des paramètres non maîtrisés. Sans état de l'art, cette expérimentation mènerait à penser que les systèmes testés généralisent particulièrement bien sur les données des bases de données testées. Ces expérimentations mettent néanmoins aussi en avant les forces du protocole STI : une attention particulière aux comparaisons entre différents paramètres, à la compréhension du phénomène observé, et l'identification d'hypothèses qui peuvent être validées, ou dans notre cas, invalidées.

Il est nécessaire d'explorer plus en détails les prédictions des modèles testés pour mieux comprendre les résultats obtenus lors de nos expérimentations. Pour ce faire, nous utilisons notre outil d'évaluation pour visualiser les données, les annotations et les prédictions obtenus par différents modèles, afin de produire des observations visuelles de ces prédictions, et proposer des interprétations de ces dernières. Cette analyse visuelle est proposée dans le chapitre suivant.

Chapter 6

Exploration visuelle et interprétation de prédictions

Chapter Summary

Résumé du chapitre

Dans ce chapitre, nous explorons visuellement les données de l'expérimentation précédente pour tenter d'expliquer les résultats obtenus. La section 6.1 présente la méthodologie utilisée pour ces explorations visuelles, qui sont réalisées sur des classes particulières, afin de limiter les contraintes temporelles et humaines. Ces observations discutées en section 6.2 mènent à la découverte de biais concomitants, influençant les prédictions du modèle sur des axes différents de la séparation entre occident et non occident dans les données. Fort de ce constat, nous étendons nos expérimentations à de nouveaux modèles et bases de données en section 6.3, et montrons que des résultats similaires sont obtenus, soutenant l'hypothèse que les biais concomitants sont présent de manière générale dans les systèmes de vision, et que la caractérisation du biais géographique est une tâche plus ardue qu'on ne l'imagine. En descendant en section 6.4 à une granularité plus fine au niveau des domaines, on obtient toujours les mêmes résultats, montrant que ce phénomène n'étant pas réduit au caractère

occidental des modèles mais étendu au biais géographique. Nous concluons sur l'importance des biais concomitants et leur influence sur les prédictions des modèles dans nos expérimentations.

6.1 Méthodologie et observation des données

Les statistiques et comparaisons générées dans le chapitre précédent ne suffisent pas à expliquer pourquoi les prédictions des modèles sont bonnes ou mauvaises, ou encore quels contextes associés à une donnée impactent la prédiction des modèles. Ces statistiques sont dans cette section combinées à des observations visuelles réalisées par un opérateur humain pour produire des observations et interprétations visuelles des phénomènes à l'oeuvre dans nos expérimentations.

6.1.1 Idée et principes

L'évaluation manuelle des prédictions des modèles a permis de souligner dans le chapitre précédent des écarts entre métriques manuelles et automatiques, dues à des erreurs durant l'annotation des données utilisées. Les résultats présentés dans le chapitre précédent vont pour autant à l'encontre d'hypothèses et de résultats académiques à la base des travaux sur le glissement géographique. De nouvelles explorations sont nécessaires pour pouvoir comprendre ces résultats. Dans ce chapitre, ces nouvelles explorations prennent la forme d'enquêtes visuelles durant lesquelles nous essayons de comprendre comment le modèle performe mieux dans un domaine géographique plutôt qu'un autre.

Ces explorations visuelles consistent en une visualisation d'instances de données et de leur prédictions, et une interprétation de la raison qui a poussé le modèle à prédire justement ou injustement une classe spécifique. C'est une méthode d'exploration et d'analyse basée sur les instances, qui est éprouvée dans la littérature[78]. La visualisation des images et prédictions pour chacune des instances d'une classe permettra à l'opérateur de fonder des hypothèses sur les facteurs influençant la prédiction d'une classe en particulier, ces hypothèses pouvant par la suite être testées. Cette méthode possède tout de même ses propres faiblesses, comme les biais internes des opérateurs, ou le fait de se reposer sur des classes prédéfinies qui sont elle-même biaisées[107]. En définitive, une synthèse comparative de plusieurs analyses réalisées par des opérateurs de cultures et niveaux sociaux différents serait idéale, mais de telles conditions n'ont pu être réunies pour cette expérimentation. Nous nous reposerons donc sur une unique analyse pour la suite de l'expérimentation.

L'outil utilisé pour la visualisation des données sera le même que l'outil précédemment développé pour l'évaluation manuelle des performances des modèles de vision. Cet outil permet la visualisation des données, des annotations et prédictions liées à cette donnée, et le parcours de l'ensemble des données liées à une classe par ces annotations ou prédictions. Il est donc adapté à la tâche d'exploration visuelle des données et d'interprétation des facteurs influençant les prédictions d'un système de vision.

6.1.2 Sélection des classes et observations

La visualisation de résultats de prédictions sur des données de manière aléatoire peut permettre d'obtenir une vision d'ensemble du fonctionnement du modèle, mais risque de perdre l'opérateur dans une diversité trop grande et un temps de traitement trop long. Par soucis de simplicité et de concision, nous sélectionnons des classes particulières pour une analyse en profondeur des mécanismes du modèle sur les données relatives à ces classes. La sélection de ces classes se basera sur leur intérêt particulier pour la compréhension de ce qui pousse le modèle CSRA à mieux performer dans une zone géographique plutôt qu'une autre. Les classes sélectionnées seront donc celles qui présentent le plus de divergence entre performance sur données occidentales et sur données non occidentales.

Les reports de résultats en chapitre précédent présentent les classes "train" et "bus" comme étant les plus performante pour des données occidentales pour la première, la seconde pour des données non occidentales, pour les prédictions de CSRA sur COCOWU. Comprendre pourquoi ces classes penchent vers l'un ou l'autre des contextes permettra de formuler des hypothèses quant au phénomène

responsable de ces tendances. Ce sont donc ces dernières qui sont sélectionnées pour une analyse visuelle.

6.1.2.1 La classe bus (meilleure performance sur des données hors occident)

68 images sont annotées ou prédites avec la classe "bus" en utilisant le modèle CSRA sur COCOWU. En observant les images prédites avec la classe "bus", on note que le modèle prédit souvent cette classe en présence de gros véhicules sur un axe routier - comme des SUV, des camions, ou des trains. Certaines variations dans ces classes influent sur le modèle, les grosses voitures étant plus souvent annotées comme des bus que les petites voitures. Comme la classe "car" est la plus représentée dans les classes associées au transport, ce phénomène peut influencer les performances de la classe "bus". L'analyse de toutes les images contenant des voitures et une distinction entre petites et grosses voitures par un opérateur permet d'obtenir le ratio de grosses et petites voitures dans les données occidentales et non occidentales, reportés en tableau 13. La définition de ce qu'est une petite ou grosse voiture est laissé à la discrétion de l'opérateur réalisant cette opération.

Table 13: Nombre de grosses et petites voitures dans les données occidentales (ED) et non occidentales (HD) de la base de données COCOWU. Les grosses voitures ont tendance à influencer le modèle CSRA à prédire la classe "bus".

	ED	HD
Grosses voitures	93	84
Petites voitures	85	179

La répartition entre petites et grosses voitures n'est pas uniforme entre données occidentales et non occidentales, les grosses voitures étant statistiquement plus présentes dans les données occidentales (52% des voitures) que dans les données non occidentales (32% des voitures). Si la prédiction de la classe "bus" est influencée par ces grosses voitures, provoquant ainsi des erreurs, le modèle CSRA a plus de chance de mal prédire la classe "bus" dans les données ED que dans les données HD. Ce phénomène mène à une meilleure performance du modèle pour la classe "bus" sur des données hors occident, et plaide en faveur de facteurs extérieurs au biais occidental pour l'explication des variations des performances du modèle sur la classe bus entre les deux contextes occidentaux et non occidentaux. D'autres phénomènes peuvent aussi être à l'oeuvre, mais ce résultat souligne la présence de contextes et d'influences cachés, résonnant avec les stratifications cachées mentionnées par Oakden-Rayner et al. (2020). Nous appellerons dans la suite ces biais cachés des biais concomitants, identifiant des phénomènes influençant les prédictions des modèles et portant atteinte à la caractérisation d'autres biais car plus influents que ceux sélectionnés.

6.1.2.2 La classe train (meilleure performance sur des données occidentales)

Les 148 images annotées dans COCOWU ou prédites par CSRA avec la classe "train" sont étudiées de la même manière pour produire une interprétation de la tendance du modèle à mieux performer pour cette classe sur les données occidentales. Une analyse visuelle de ces données et prédictions révèle une forte corrélation entre la présence de rails dans l'image et la prédiction de la classe "train" par le modèle CSRA. Dans l'ensemble de COCOWU, la moitié des mauvaises prédictions liées à cette classe sont dues à des images comportant des rails mais pas de train. Ces images sont bien plus fréquentes dans les données hors occident (15 images sur 90) que dans les données occidentales (3 images sur 58). Ceci mène à une meilleur prédiction de la classe "train" sur les données occidentales que sur les données non occidentales. Une des raisons pour laquelle la classe "train" est plus performantes sur les données occidentales n'est pas du à une variation du contexte géographique, mais à une répartition d'un certain type de données de manière inégale. Ce phénomène a plus d'importance que le biais occidental dans les données pour l'évaluation des prédictions de la classe "train". Nous identifions alors un nouveau biais concomitant portant atteinte à la caractérisation du biais occidental.

6.2 Discussions

6.2.1 Interprétation des performances

Les observations réalisées mettent en lumière une situation commune lors de déploiement de modèles sur le terrain : l'apparition de biais concomitants et leur influence sur les performances attendues d'un modèle. Ces biais sont difficiles à prévoir en amont, et peuvent avoir un impact important sur les mesures et métriques utilisées lors du déploiement du modèle. Dans notre cas, ils ont plus d'influence sur les performances du modèle que le biais occidental, rendant caduque une possible caractérisation de l'occidentalité du modèle. Ces biais concomitants sont à l'origine des variations de performance pour les deux classes étudiées, et permettent aussi d'expliquer l'absence de variation de performance entre donnée occidentales et non occidentales dans la plupart des classes : le modèle est sensible à d'autres biais et répartition que celle réalisée dans la séparation des données en deux domaines géographiques distincts. Une analyse exhaustive de ces biais concomitants est hors du champ de cette exploration, le phénomène étant suffisamment illustré sur les deux classes explorées.

L'explication attendue des variations de performance entre données occidentales et données non occidentales est la variation de contexte entre ces données, intégrant des paramètres sociaux, économique et géographiques. Une exploration des prédictions des modèles sur ces données indique d'autres sources d'explications, combinant mécanismes internes du modèle et répartition de données. Les mécanismes internes du modèle identifiés sont des proximités de définition de classes (proximité entre grosses voitures et bus) et l'association entre classe et contexte (association entre rails et train). Les répartitions de données identifiées sont des biais de représentation issus de la collecte des données (statistiquement plus de grosses voitures en occident, statistiquement moins de rails vides en occident). Ces biais de représentation peuvent être interprétés comme des biais géographiques, mais ont des effets imprévus sur les performances des modèles testés.

Ce qui est observé grâce à cette analyse visuelle, c'est qu'il y a bel et bien un biais géographique, mais que ce dernier intervient au niveau des données des bases de données, et non au niveau des mécanismes internes du modèle. Si cette analyse met à jour des biais dans le modèle, ce sont des biais liant des concepts entre eux, et non pas des biais occidentaux. Les différences de performance entre données occidentales et données non occidentales ne s'expliquent pas par les biais occidentaux du modèle, mais par la combinaison de biais géographique non maîtrisés dans les bases de données et de mécanismes internes du modèle étudié. Le biais géographique intervient dans les données autrement qu'à travers les notions d'occident ou de pays, dans la répartition de concepts auxquels les modèles sont sensibles.

6.2.2 De la complexité du biais

Ces résultats montrent que la complexité des biais est d'une importance majeure. Les concepts vagues comme le "biais géographique" ou le "caractère occidental" encapsulent des biais de plus faible complexité, comme le "taux de petites ou grosses voitures" et le "taux de rails sans trains". Les performances d'un modèle sont influencés par ces biais de faible complexité, empêchant la caractérisation de biais plus complexes, car ces biais de faibles complexité partitionnent chacun les données en des domaines qui leur sont spécifiques. Formellement parlant, cela se rapproche de la décomposition d'un contexte C en un ensemble de sous-contextes $\{C_1, ..., C_n\}$, contextes auxquels on peut lier des biais. Le contexte occidental d'une donnée encapsule ainsi un ensemble de sous-contextes qu'il est difficile de cerner en amont et d'identifier en aval sans opérations temporellement coûteuses. Dans le cas où les biais liés à ces sous-contextes partitionnent les données en des domaines compatibles avec le contexte encapsulant, on pourra caractériser ce contexte plus large. Dans le cas contraire, la caractérisation d'un biais lié au contexte encapsulant sera rendue complexe par les biais concomitants liés aux sous-contextes, influençant les performances sur une partition des données différentes de celle réalisée par le biais plus large.

Il est possible d'utiliser le protocole STI en décomposant ces biais de forte complexité en leurs composants de faible complexité pour les caractériser. Cette opération nécessite néanmoins des opéra-

tions temporellement prenantes et l'investissement d'opérateurs humains, eux-même non exempts de biais. Un arbitrage entre complexité d'un biais et capacité à caractériser ce biais est nécessaire, et peut faire l'objet de futures expérimentations.

6.2.3 A propos du biais géographique

Le biais géographique est un concept vaste, encapsulant des contextes historiques, culturels, sociaux, économiques, techniques. Nous discutons ici des aspects que nous avons mis à jour dans nos travaux et des problématiques rencontrées.

6.2.3.1 Un problème de définition

Le principal problème lié au biais géographique tient à la multiplicité de sa définition, aux multiples manières dont il survient dans un système. Il peut intervenir dans la source des données, dans l'échantillonnage de ces dernières, dans la construction la tâche à réaliser, dans le processus d'apprentissage, d'évaluation, de déploiement... La source de ces interventions est également multiple, comme le rappelle l'arbre de sources en chapitre 4. Nous devons reconnaître que le biais géographique est au sommet d'une hiérarchie de biais moins complexes. Si on peut le caractériser lors de génération artificielles, l'isolation d'un biais complexe dans des données réelles est une tâche ardue, et sa caractérisation demeure donc inaccessible.

6.2.3.2 Mitiger le biais géographique

Comment mitiger le biais géographique lorsque la caractérisation de ce dernier reste hors de portée ? Atwood et al. (2020) offre une solution à travers la compétition "The Inclusive Image Competition", mesurant la capacité de modèles à généraliser sur des données inclusives. Cette stratégie de mitigation assume que les bases de données possèdent des biais géographiques, et c'est aux modèles de trouver des stratégies pour s'adapter aux nouveaux contextes géographiques. Ces travaux se concentrent sur une partie du développement des systèmes, se montrant efficace pour pallier les biais internes des modèles, et reflétant le point de vue des auteurs que se débarrasser du biais géographique dans les données est une tâche impossible.

C'est une bonne illustration de l'importance des hypothèses et conjectures lors de la construction d'une stratégie de mitigation. Ils identifient une source pour le biais identifié (le manque d'inclusivité des bases de données d'images), un type de biais (le biais d'apprentissage) et une métrique d'impact (performances des modèles développés sur un jeu de données caché) afin de caractériser le biais appris par les modèles. Cette stratégie ne prend pas en compte les autres manières dont le biais géographique intervient dans le développement des systèmes, mais assume leur présence dans n'importe quel système, et ce qui est isolé n'est alors pas le biais géographique mais la capacité des modèles à s'adapter à des contextes inclusifs. En s'inspirant de leurs travaux, les futures initiatives visant à mitiger un biais doivent s'employer à définir un cadre spécifique pour ce biais, en définissant source, type et impact, pour le définir entièrement.

6.2.3.3 Implications

Comment les expérimentations du chapitre précédent et les explorations visuelles réalisées jusqu'ici permettent-elles de modifier la compréhension du biais géographique ? Nous avons montré qu'il n'était pas trivial de démontrer le biais occidental d'un modèle, malgré l'utilisation de bases de données plus inclusives que les banques d'images classiques. Le biais occidental est une entité plus complexe que suggéré par les précédentes publications académiques, avec des hiérarchies de contextes sous-jacents dont les effets sont impossibles à prévoir.

Les biais sous-jacents identifiés dans les explorations ne sont pourtant pas exempt d'un caractère géographique. Pour la classe "bus", c'est la répartition entre petites et grosses voitures qui impacte les performances du modèle. Cette répartition a une composante géographique, les pays de l'occident

ayant une plus grand proportion de grande voitures dues à un contexte économique et industriel plus favorable à ces dernières. Pour la classe "train", c'est le taux de photos de rails sans train qui influence les performances du modèle. On peut conjecturer que ces photos sont plus fréquentes dans les pays du Sud que dans les pays du Nord, puisque les trains sont plus présents dans les pays occidentaux, et donc les rails plus fréquentées dans ces derniers. Cette explication est néanmoins bancale, car il y a aussi le revers d'une plus grande présence de train en occident : une plus grande présence de rails, et donc une plus grande probabilité d'avoir des rails dans des photos. Encore une fois, c'est un facteur géographique qui joue sur la répartition de ces données. Mais ces facteurs géographiques ne sont pas ceux attendus, et interagissent de manière surprenante avec les mécanismes internes du modèle. Le décalage attendu entre données occidentales et données non occidentales était celui provenant d'un glissement de conception (différences dans l'apparence), de contexte (différence dans l'environnement) ou de définition (différence dans la manière de désigner une chose) des classes sélectionnées. En théorie, ces glissements impactent les performances du modèles à cause des mécanismes internes (les poids et paramètres sélectionnés) appris sur des données occidentales. Nos explorations montrent qu'il en est autrement, et que les mécanismes internes du modèle CSRA utilisé sont impactés par des facteurs géographiques imprévus.

Les observations réalisées sont spécifiques au modèle CSRA sur la base de données COCOWU, même si les résultats pour les classes de transport sont similaires sur la base de données DSD. De plus amples expérimentations, impliquant d'autres modèles et bases de données, sont nécessaires pour pouvoir étendre nos observations au domaine de la vision par ordinateur.

6.3 L'évaluation du caractère occidental des systèmes d'IA

Nos expérimentations ont jusqu'ici montré qu'il était difficile de caractériser l'occidentalité du modèle CSRA à l'aide de la base de données COCOWU. Cette difficulté s'étend-elle à travers différents modèles et bases de données ? Nous proposons de réduire les contraintes sur la sélection de modèles et de bases de données pour explorer de manière plus générale l'occidentalité dans les systèmes de vision.

6.3.1 Expérience

Nous appliquons ici la méthodologie STI pour la caractérisation du caractère occidental de modèles de vision par ordinateur. Nous étendons l'expérience précédente grâce à une levée sur certaines contraintes pour la sélection de modèles, et en conséquence utilisons uniquement des évaluations automatiques. Les hypothèses sur lesquelles cette expérience se repose sont les mêmes que celles présentées dans le chapitre précédent.

6.3.1.1 Bases et jeux de données

Les mêmes bases de données candidates sont considérées pour cette expérience. Parmi ces cinq bases de données, nous sélectionnons pour cette expérience COCOWU pour sa couverture totale du globe, DSD pour sa qualité de métadonnées, et GYFCC en tant que représentant des bases de données à large volume. Nous couvrons ainsi plusieurs types de bases de données, avec des conceptions et des philosophies différentes. Il est néanmoins nécessaire pour chacune de ces bases de données d'être séparées en jeu de données occidental et jeu de données non occidental. Cette séparation est effectuée en triant les données par pays ou zone géographique de provenance et est précisée dans le tableau 8. Les données occidentales seront dans toute l'expérience considérées comme les données ED ("En Distribution"), et les données hors occident considérées HD ("En Distribution"). Cette décision fait écho au constat d'une domination sans équivoque des données occidentales dans les bases de données d'images utilisées pour entraîner les modèles des systèmes de vision.

6.3.1.2 Sélection de modèles

Nous suivons pour la sélection de modèles l'initiative de DeVries et al. (2019) et considérons des systèmes de vision par ordinateur prêts à l'emploi. Ils testent ainsi les systèmes de Microsoft Azure[1], Clarifai[2], Google Cloud Vision[3], Amazon Rekognition[4] et IBM Watson[5], en plus d'un système à l'état de l'art entraîné sur des données publiques. Nous nous inspirons de ces travaux et testons les modèles d'Amazon, Google et Microsoft dans des conditions multi-classe, et conservons CSRA[177] entraîné sur la base de données MS COCO en tant que modèle à l'état de l'art. Ces modèles et systèmes automatiques permettent de répondre à la tâche identifiée tout en permettant une comparaison à l'existant et aux précédents travaux. Nous levons ainsi la contrainte financière concernant les API pour la sélection de modèles candidats, bien que ces contraintes s'appliquent à notre situation et nous empêcheront certaines des computations de l'expérience.

6.3.1.3 Protocole STI et métrique

Nous conservons les mêmes choix de source, de type et d'impact de biais pour cette expérience, en considérant le caractère occidental d'un modèle comme un sous-produit du biais géographique des modèles. Le protocole STI est donc déjà mis en place, et il ne reste qu'à déterminer comment générer la métrique d'impact choisie.

La métrique qui sera utilisée pour témoigner du biais occidental sera, comme dans les expériences précédentes, la différence de performance du modèle entre données occidentales et données hors occident. La performance peut être mesurée selon plusieurs métriques de performances : précision top-1, précision top-5, Hamming Loss. La précision top-1 et la Hamming Loss étant liées en classification multi-classe (précision = 1 - Hamming Loss), nous présenterons les résultats pour les métriques de précision top-1 et de précision top-5.

Les métriques de performances sélectionnées font l'hypothèse d'un ensemble de classes partagés par les modèles et les bases de données sur lesquelles ces modèles sont testés, c'est à dire $Y_S = Y_M$, avec Y_M l'espace de sortie du modèle et Y_S l'espace des annotations de la base de donnée. Comme ces expériences sont basés sur l'évaluation de systèmes de vision par ordinateur sans l'utilisation d'entraînement supplémentaire via apprentissage par transfert ou peaufinage des modèles, cette hypothèse n'est pas vérifiée. Dans la plupart des cas de l'expérience, nous avons une différence d'ensemble de classes, c'est à dire $Y_S \neq Y_M$. Des approches alternatives pour l'évaluation de la performance d'un modèle sur les données doivent donc être considérées. Deux de ces approches sont présentées ci-dessous, une approche par intersection d'ensemble de classes et une approche par correspondance floue manuelle.

L'intersection des ensembles de classes correspond à l'abandon de toute classe qui n'est pas dans les deux ensembles de classes distincts. L'approche par correspondance floue manuelle consiste à manuellement lier les ensembles de classes, en faisant un parallèle entre les classes des modèles et les classes des bases de données, cherchant pour chaque classe les correspondances possibles et abandonnant la classe si aucune correspondance n'est trouvée. Une autre stratégie pourrait automatiser ce processus de recherche de correspondance entre différents ensembles de classes, par exemple à l'aide d'un adaptateur linéaire. Diverses implémentations pour cette stratégie ont vu le jour, mais aucune n'a produit de résultats satisfaisants, et cette dernière approche est donc laissée en tant que considération pour des travaux futurs. Les deux autres approches sont utilisées et comparées dans les expériences implémentées.

L'intersection d'un ensemble de classes entre un modèle M et une base de données S est une simple intersection entre les classes des deux entités : $Y_{it} = Y_M \cap Y_S$. Dans le cas de la correspondance floue manuelle, l'ensemble de classes utilisé est $Y_{fz} = \mathcal{F}_{match}(Y_M, Y_S)$ où $\mathcal{F}_{match} : Y_M, Y_S \rightarrow Y_S$ est

[1] https://azure.microsoft.com/
[2] https://www.clarifai.com/
[3] https://cloud.google.com/vision
[4] https://aws.amazon.com/rekognition/
[5] https://www.ibm.com/watson

une fonction qui fait correspondre à chaque classe du modèle une ou plusieurs classes de la base de données, ou l'ensemble vide. Dans le cas où l'ensemble vide est attribué à une classe, on abandonne cette classe dans le processus d'évaluation de la performance du modèle. Chaque couple de modèle et base de données a sa propre fonction de correspondance. Ces fonctions de correspondances ont été réalisées à la main, à l'aide des données d'entraînement quand ces dernières étaient disponibles. Quand aucune donnée d'entraînement n'était disponible, la correspondance dépend uniquement de la similarité entre les classes, estimée par les expérimentateurs. Certains systèmes sélectionnés pour cette expérience ne fournissent pas l'ensemble de classes utilisé par les modèles implémentés. Dans ce cas, pour chacune des bases de données sélectionnées, un ensemble de classes est considéré comme la liste des classes uniques prédites par le modèle sur les données de la base de donnée considérée. Dans ce cas, Y_M peut donc varier en fonction de la base de donnée utilisée pour tester le modèle.

6.3.1.4 Détails d'implémentation

Cette expérimentation présente une implémentation du protocole STI sur des données réelles. Trois bases de données ont été sélectionnées pour tester la méthodologie STI : DSD, COCOWU et GYFCC. L'exploitation de ces bases de données est plus ou moins complexe ; les données de DSD sont disponibles en téléchargement, tandis que COCOWU ne comprend pas directement des images mais des liens vers des images hébergées sur Flickr. Quant à GYFCC, le téléchargement des données est complexe du à la taille de la base de donnée. Mais les métadonnées comprennent assez d'information pour retrouver les adresses d'hébergement des images de la base de donnée, permettant l'accès direct aux données de manière individuelle.

Les systèmes de vision par ordinateur d'Amazon, de Google et de Microsoft sont accessibles via API. Parmi ces systèmes, plusieurs modèles sont disponibles - versions, tâches, taille des modèles, plusieurs paramètres varient pour chacun de ces modèles. Par chacune des API, nous sélectionnons les modèles qui réalisent la tâche de classification multi-classe d'objets et de concepts de la vie commune. Le modèle à l'état de l'art en 2022 utilisé est CSRA[177], est utilisé avec des poids pré-entraînés sur la base de données MS COCO.

Pour chaque combinaison de base de donnée et modèle, les performances des modèles sont générées par batch de données pour chacune des métriques sélectionnées. Ces performances sont sauvegardées sous format csv pour générer par la suite les métriques et rapports de performance.

L'ensemble de l'implémentation est réalisé en python. Par soucis de compatibilité avec les différentes API, bases de données et modèles sélectionnés, la gestion des calculs en local est effectué sous la bibliothèque PyTorch. Les résultats des systèmes de vision proposés par Google, Amazon et Microsoft sont récupérés entre Juin 2022 et septembre 2023.

Le coût d'utilisation de ces API est variable pour chaque base de donnée et dépend du volume de données pour chacune. Une estimation du prix pour chacune de ces bases de données est disponible en tableau 14. Ce coût est dissuasif pour les bases de données GYFCC (plus de 1000$ par système) et DSD (autour de 50$ par système). Les systèmes de vision des grands groupes seront donc uniquement utilisés sur les données de COCOWU (environ 10$ par système) pour cause de contraintes budgétaires.

Table 14: Estimation des coûts d'utilisation des diverses API sur les bases de données sélectionnées pour l'expérience.

API Datasets	DSD	COCOWU	GYFCC
Amazon	38$	9$	1118$
Google	57$	13$	1720$
Microsoft	38$	9$	1095$

Ces contraintes ne sont pas observées pour le modèle à l'état de l'art dont les computations sont en local, et ce dernier pourra donc être décliné sur l'ensemble des bases de données.

6.3.2 Résultats

Les résultats de cette expérience sont présentés dans les tableaux 15, 16, 17 et 18. Ces tableaux reportent les performances des modèles sélectionnés sur les bases de données DSD, COCOWU et GYFCC, en utilisant les différents ensembles de classes d'intersection (indication "it") et d'association manuelle (indication "mn"). Ces tableaux reportent en plus de la différence entre performance sur données occidentales (considérées ED) et non occidentales (considérées HD).

Table 15: Évaluation en précision moyenne des performances des modèles sélectionnés sur les bases de données DSD, COCOWU et GYFCC, avec les ensembles de classes d'intersection. × reporte une situation où les évaluations n'ont pu être menées à cause de coûts de computation trop élevés.

Base de donnée	DSD			COCOWU			GYFCC		
Jeu de données	ED_{it}	HD_{it}	$diff_{it}$	ED_{it}	HD_{it}	$diff_{it}$	ED_{it}	HD_{it}	$diff_{it}$
CSRA	0.9754	0.9798	−0.0044	0.9903	0.9906	−0.0003	0.9873	0.9875	−0.0002
AmazonRekog	×	×	×	0.9892	0.9890	0.0001	×	×	×
GoogleCV	×	×	×	**0.9944**	**0.9946**	−0.0002	×	×	×
Microsoft	×	×	×	0.9907	0.9911	−0.0004	×	×	×

Table 16: Évaluation en précision moyenne des performances des modèles sélectionnés sur les bases de données DSD, COCOWU et GYFCC, avec les ensembles de classes par association manuelle. × reporte une situation où les évaluations n'ont pu être menées à cause de coûts de computation trop élevés.

Base de données	DSD			COCOWU			GYFCC		
Jeu de données	ED_{mn}	HD_{mn}	$diff_{mn}$	ED_{mn}	HD_{mn}	$diff_{mn}$	ED_{mn}	HD_{mn}	$diff_{mn}$
CSRA	0.9602	0.9679	-0.0077	0.9903	0.9906	-0.0003	0.9873	0.9874	-0.0001
AmazonRekog	×	×	×	0.9917	0.9916	0.0001	×	×	×
GoogleCV	×	×	×	**0.9920**	**0.9919**	0.0001	×	×	×
Microsoft	×	×	×	0.9915	**0.9919**	-0.0004	×	×	×

Les valeurs négatives dans les colonnes "diff" des tableaux témoignent d'une meilleure performance sur les données non occidentales que sur les données occidentales. Ce sont donc des instances contraires au résultat attendu, qui est une meilleure performance des modèles sur les données occidentales. Les différences de performances sont faibles pour chacune des instances, pratiquement négligeables.

6.3.3 Discussion

Les résultats présentent les même tendances que ceux de l'expérience précédente : les tendances observées ne sont pas celles attendues. Les résultats ne permettent pas de conclure sur l'impact du biais occidental des modèles, et semblent souligner qu'un tel impact n'existe pas.

Les variations d'un ensemble de classes à un autre n'apportent pas de changement important et les différences de performance dues à ces variations, tout chose égale par ailleurs, sont négligeables. La variation de la précision top-1 moyenne à la précision top-5 moyenne en tant que métrique de performance des modèles ne permet pas non plus de modifier grandement les tendances, et si certains scores de différence passent de négatif et positif ou inversement, les valeurs de ces différences sont négligeables. La plupart des valeurs de différences étant négatives, on pourrait d'ailleurs conclure que les modèles étudiés possèdent un caractère non occidental, c'est à dire qu'ils se comportent mieux sur des données possédant un contexte non occidental. Ces résultats sont alignés avec ceux de l'expérience précédente, et la consistance de ce dernier porte à confusion.

Une exploration visuelle de l'ensemble des prédictions combinant modèles et bases de données utilisées n'est pas envisageable, à cause d'un coût temporel et humain élevé. Comme nous observons

Table 17: Évaluation en précision top-5 des performances des modèles sélectionnés sur les bases de données DSD, COCOWU et GYFCC, avec les ensembles de classes d'intersection. × reporte une situation où les évaluations n'ont pu être menées à cause de coûts de computation trop élevés.

Base de données		DSD			COCOWU			GYFCC	
Jeu de données	ED_{it}	HD_{it}	$diff_{it}$	ED_{it}	HD_{it}	$diff_{it}$	ED_{it}	HD_{it}	$diff_{it}$
CSRA	0.9887	0.9739	0.0148	**0.9338**	**0.9416**	-0.0078	0.9654	0.9682	-0.0028
AmazonRekog	×	×	×	0.9042	0.9208	-0.0166	×	×	×
GoogleCV	×	×	×	0.8934	0.8955	-0.0021	×	×	×
Microsoft	×	×	×	0.8209	0.8332	-0.0123	×	×	×

Table 18: Évaluation en précision top-5 des performances des modèles sélectionnés sur les bases de données DSD, COCOWU et GYFCC, avec les ensembles de classes par association manuelle. × reporte une situation où les évaluations n'ont pu être menées à cause de coûts de computation trop élevés.

Base de données		DSD			COCOWU			GYFCC	
Jeu de données	ED_{mn}	HD_{mn}	$diff_{mn}$	ED_{mn}	HD_{mn}	$diff_{mn}$	ED_{mn}	HD_{mn}	$diff_{mn}$
CSRA	0.9568	0.9169	0.0399	**0.9338**	**0.9416**	-0.0078	0.9225	0.9339	-0.0114
AmazonRekog	×	×	×	0.9047	0.9217	-0.0170	×	×	×
GoogleCV	×	×	×	0.8405	0.8378	0.0027	×	×	×
Microsoft	×	×	×	0.8252	0.8389	-0.0137	×	×	×

les mêmes phénomènes pour les différentes combinaisons de modèles et de bases de données que celui précédemment observé et exploré plus en détail pour le modèle CSRA sur COCOWU, nous faisons l'hypothèse que des biais concomitants sont encore à l'oeuvre ici et nuisent à l'évaluation du caractère occidental des API utilisées. Ces derniers ne sont alors pas spécifiques à la situation de l'expérience précédente, ou au modèle CSRA et la base de donnée COCOWU, mais sont plutôt la règle dans les systèmes de vision. Cette hypothèse, si elle est vérifiée, nuit lourdement à la conception de protocoles ou benchmarks permettant d'évaluer en amont la capacité d'un modèle à être déployé dans une zone géographique particulière.

Si l'exploration visuelle des données pour identifier les biais concomitants est une tâche qui paraît ici titanesque, nous pouvons néanmoins proposer une visualisation avec une granularité plus fine, en observant les performances des modèles pour chaque pays ou zone géographique représentée dans les données. Une observation visuelle de ces données pourrait permettre de déterminer certaines tendances et d'expliquer les résultats obtenus dans cette expérience. Nous proposons une telle expérience dans la section suivante.

6.4 Visualisation des performances par domaine géographique

Dans cette section, nous proposons d'explorer visuellement les différentes combinaisons de modèles et bases de données de l'expérience précédente, à l'aide d'une visualisation de performances par pays. Cette visualisation se rapproche de la carte de performance proposée par DeVries et al. (2019) et permettra peut-être de découvrir des tendances géographiques qui pourront faire naître des hypothèses pour expliquer les résultats obtenus précédemment.

6.4.1 Expériences

Pour chaque combinaison de modèle et base de donnée de l'expérience précédente, nous allons produire une carte du monde permettant de visualiser les performances par pays du modèle sélectionné sur la base de donnée sélectionnée.

Nous reprenons la construction de l'expérience précédente : même implémentation du protocole STI, même sélection de bases de données, de modèles, mêmes computations.

6.4.1.1 Implémentation

Il n'est pas nécessaire pour produire ces visualisations de refaire des calculs, les résultats des prédictions des calculs de l'expérience précédente étant sauvegardées. Nous utilisons donc les mêmes computations pour la visualisation des performances par pays. Nous suivons donc les mêmes contraintes computationnelles et limites au niveau des expériences réalisées.

Ces visualisations sont générées en python à l'aide de plotly[6], une bibliothèque pour la génération de graphes et cartes.

6.4.2 Résultats

Les cartes présentées ci-dessous présentent les performances obtenues par les modèles sous forme de carte de performance par pays pour chacune des bases de données où les computations ont pu être réalisées.

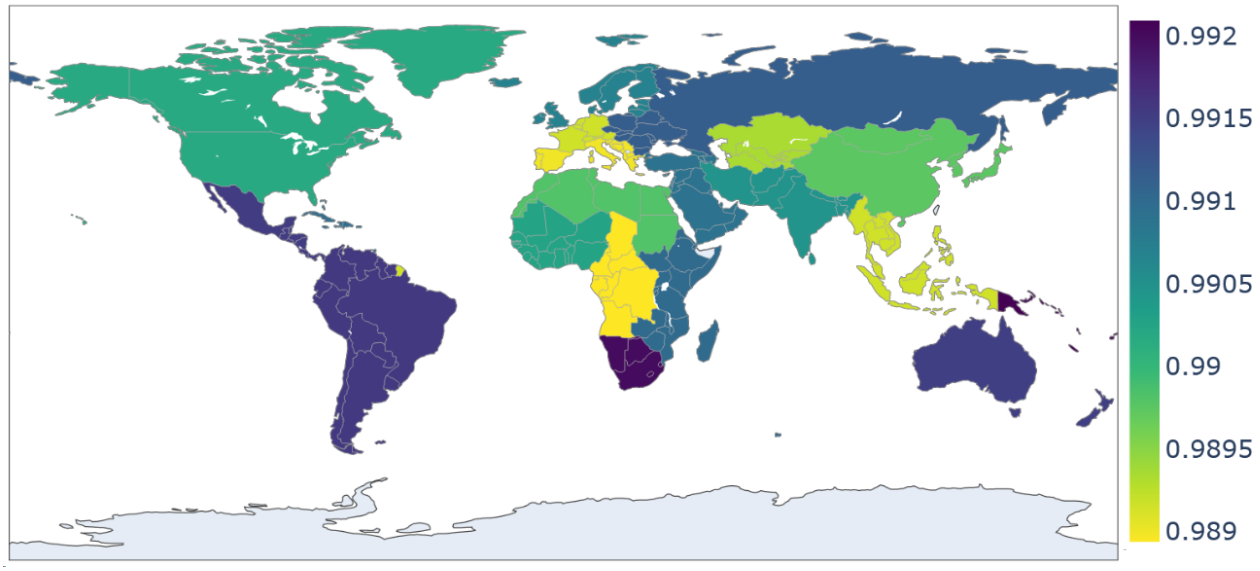


Figure 44: Précision top-1 par zone géographique du modèle CSRA sur la base de données CO-COWU. Le modèle et la base de données possédant le même ensemble de classes, il n'est pas nécessaire de sélectionner un ensemble de classes particulier.

[6]https://plotly.com/

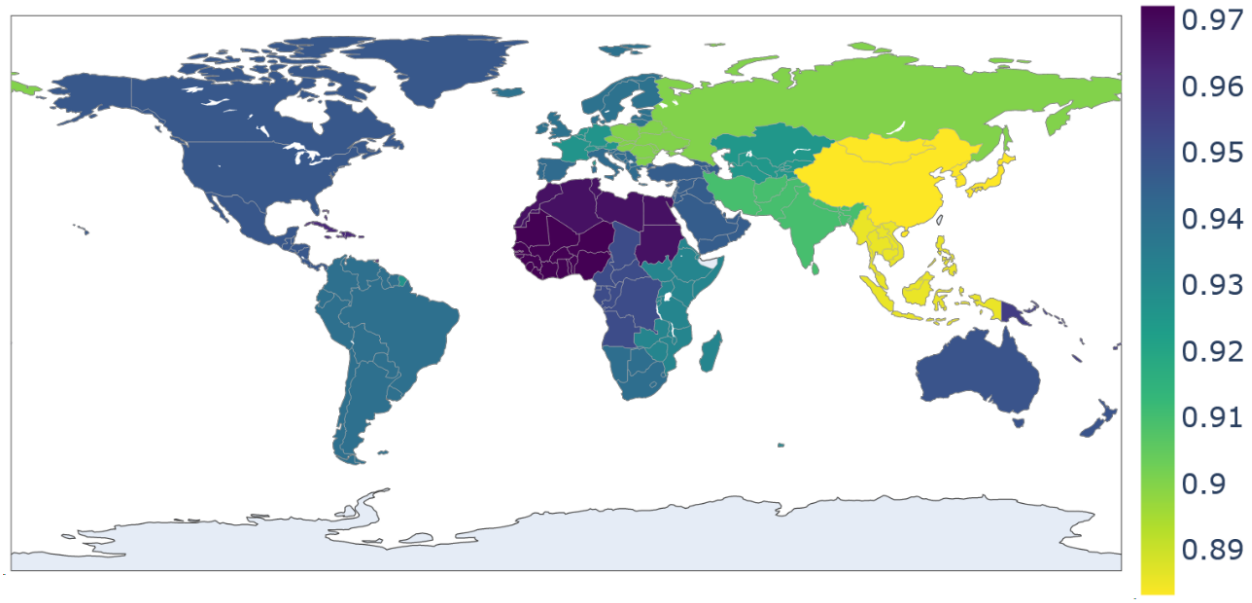


Figure 45: Précision top-5 par zone géographique du modèle CSRA sur la base de données COCOWU. Le modèle et la base de données possédant le même ensemble de classes, il n'est pas nécessaire de sélectionner un ensemble de classes particulier.

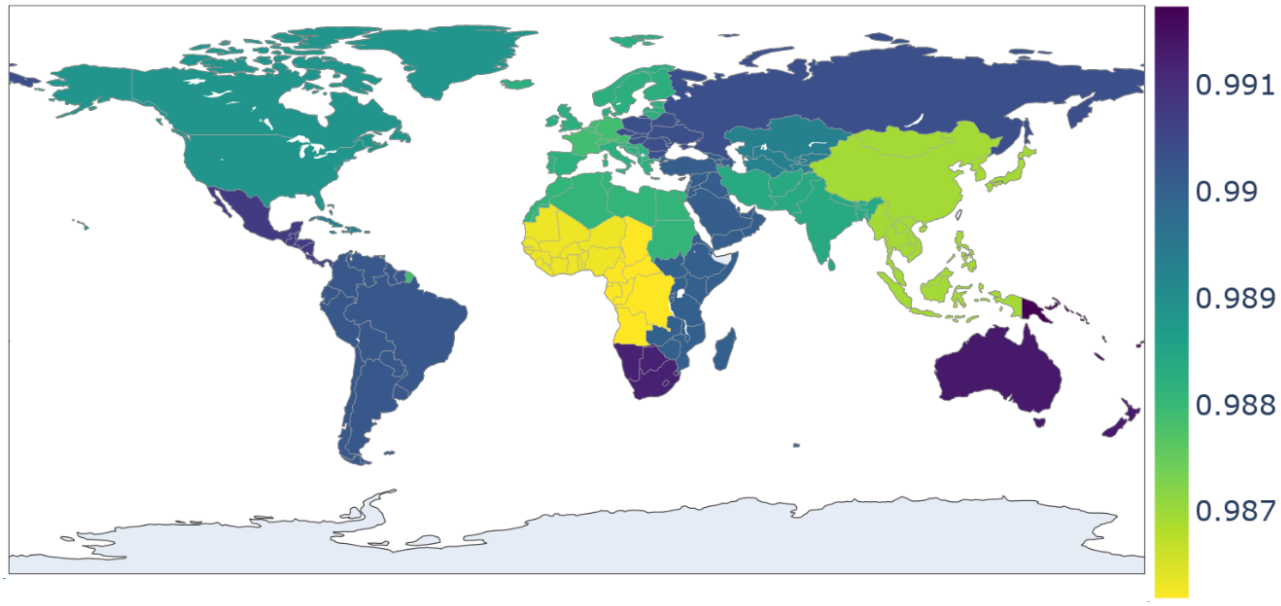


Figure 46: Précision top-1 par zone géographique du modèle d'Amazon sur la base de données COCOWU avec l'ensemble de classes d'intersection.

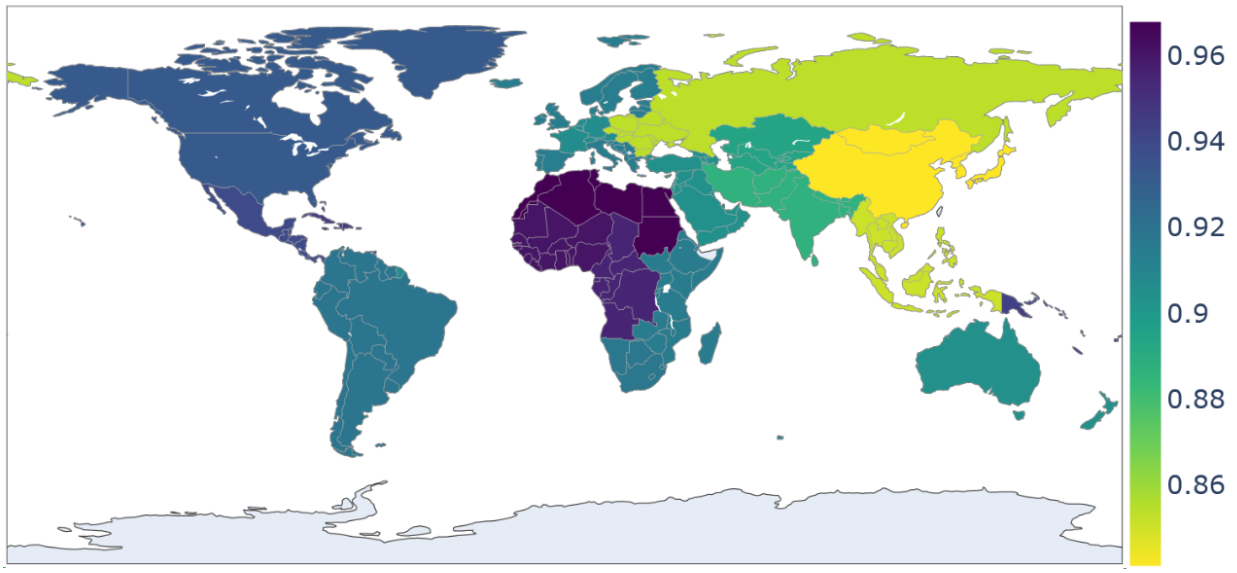


Figure 47: Précision top-5 par zone géographique du modèle d'Amazon sur la base de données COCOWU avec l'ensemble de classes d'intersection.

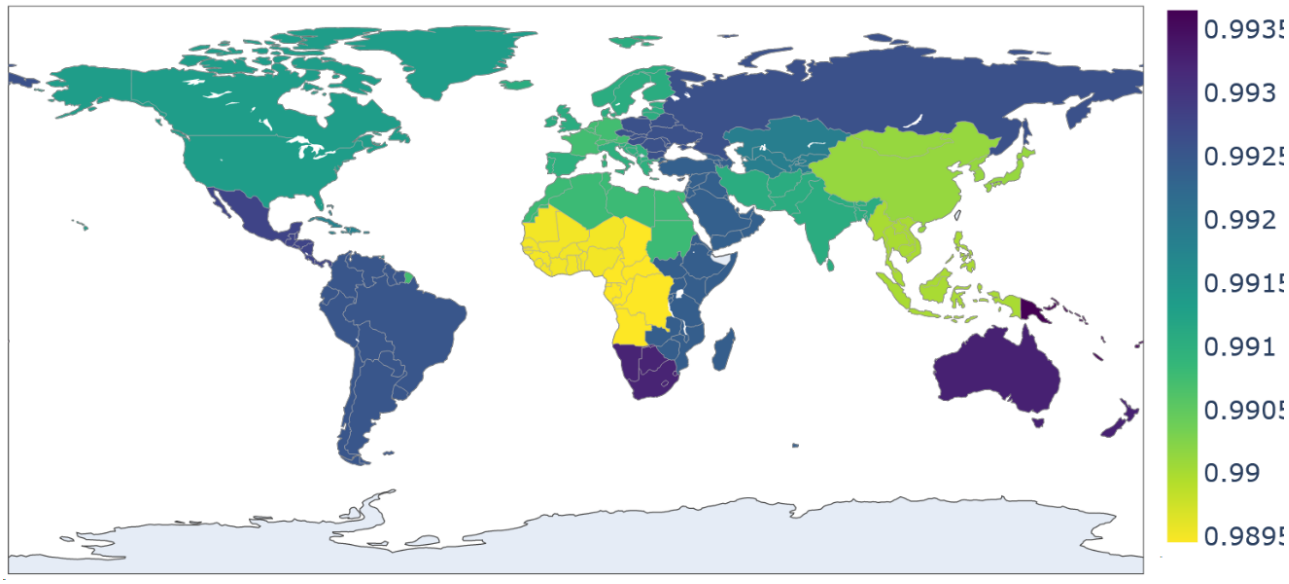


Figure 48: Précision top-1 par zone géographique du modèle d'Amazon sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

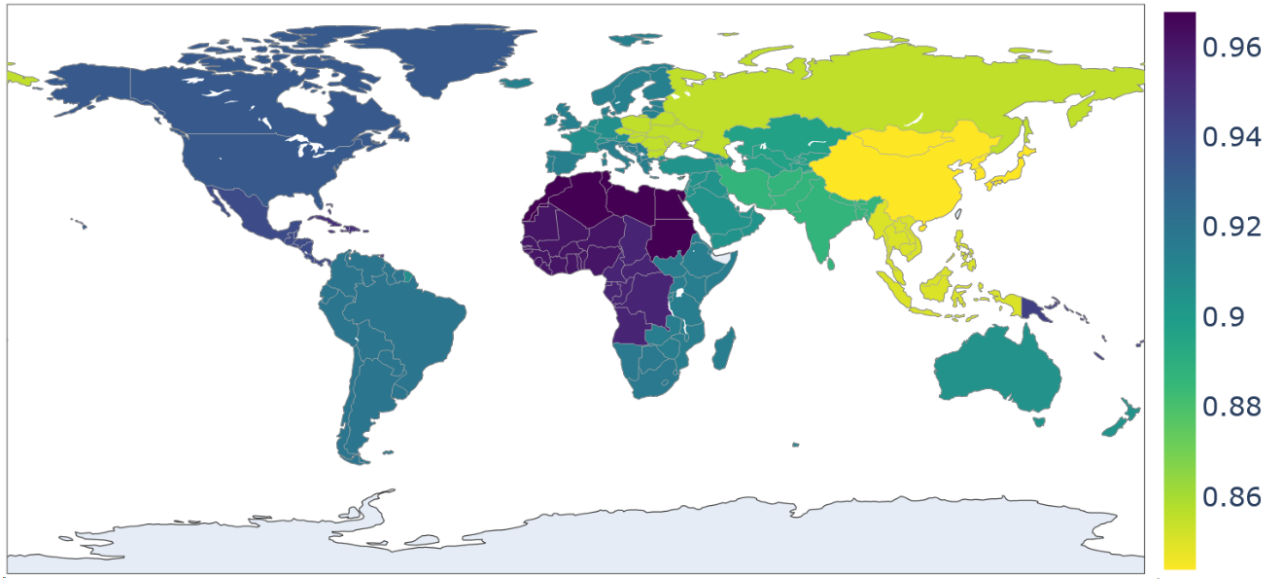


Figure 49: Précision top-5 par zone géographique du modèle d'Amazon sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

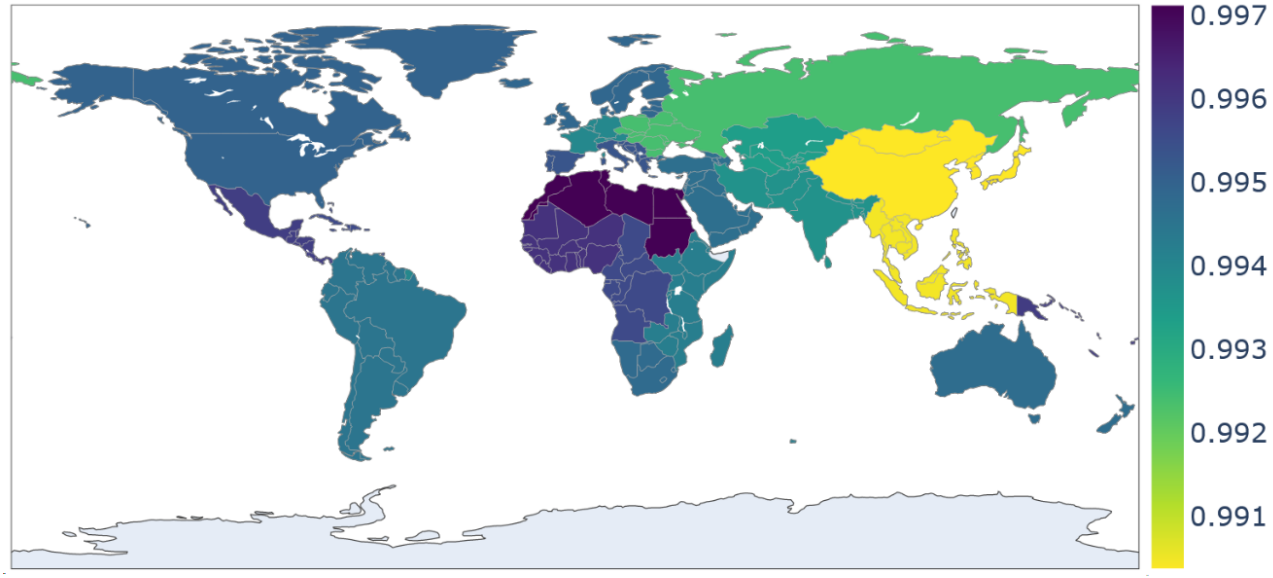


Figure 50: Précision top-1 par zone géographique du modèle de Google sur la base de données COCOWU avec l'ensemble de classes d'intersection.

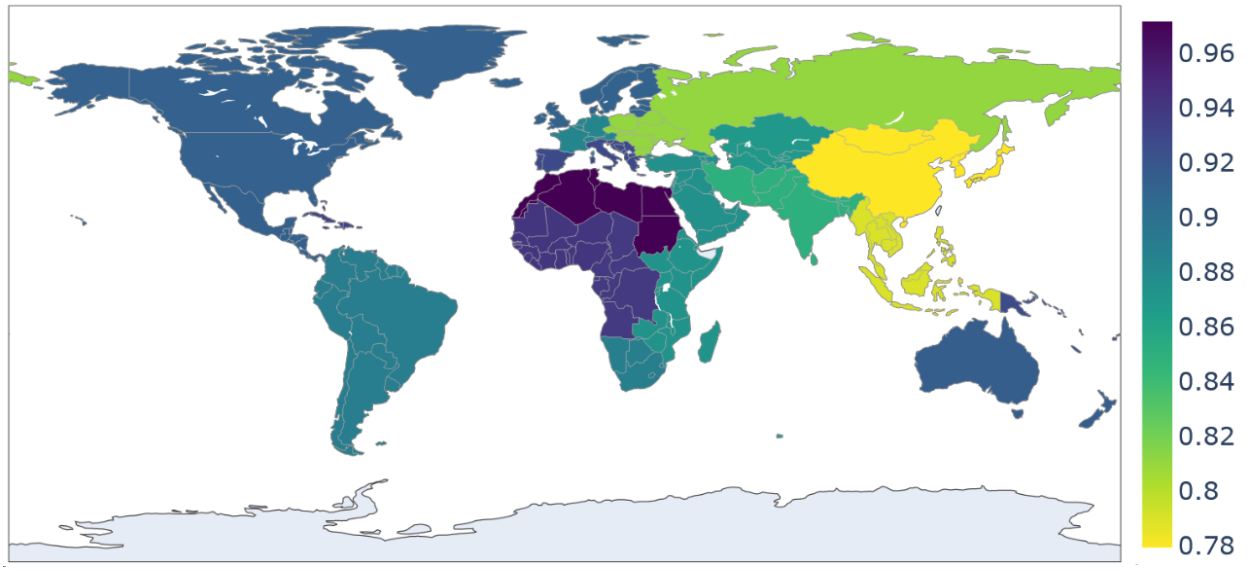


Figure 51: Précision top-5 par zone géographique du modèle de Google sur la base de données COCOWU avec l'ensemble de classes d'intersection.

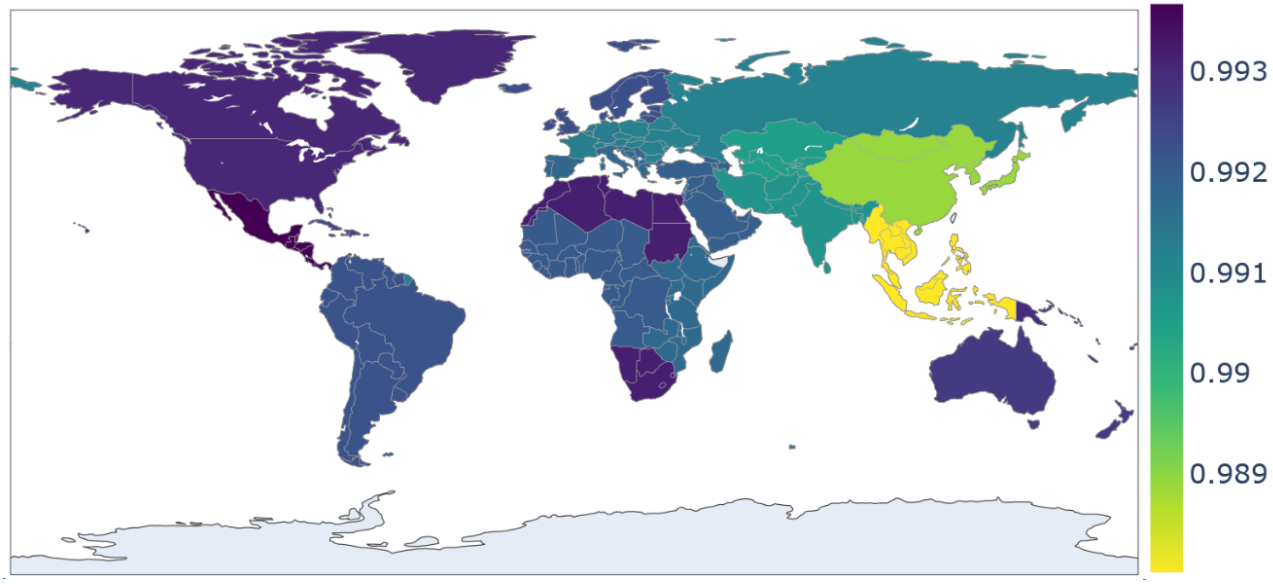


Figure 52: Précision top-1 par zone géographique du modèle de Google sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

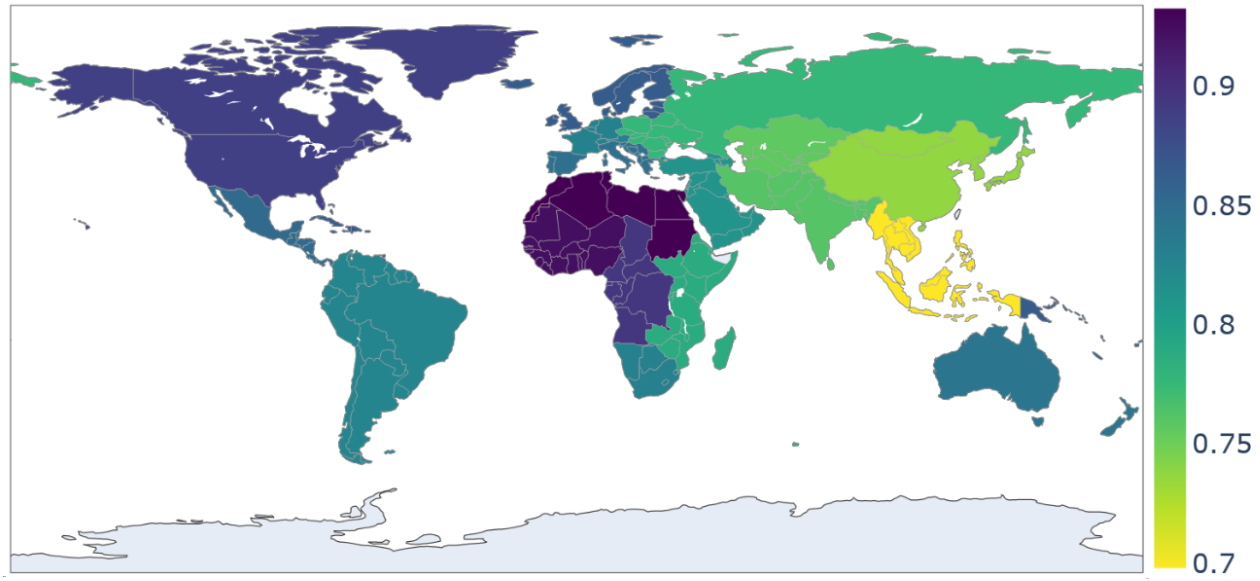


Figure 53: Précision top-5 par zone géographique du modèle de Google sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

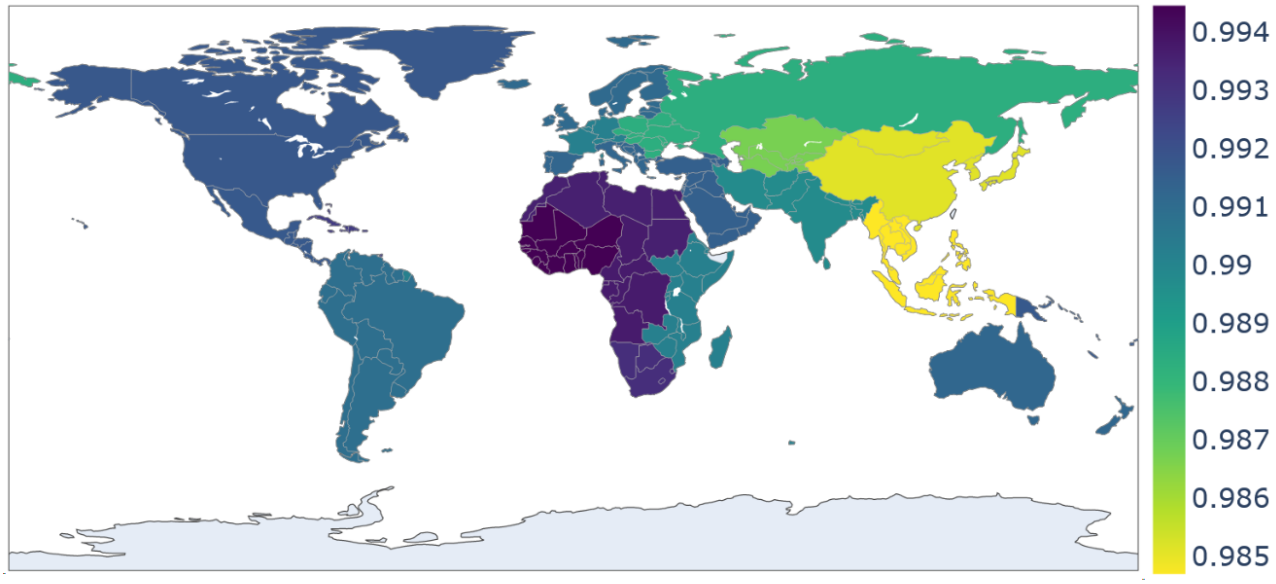


Figure 54: Précision top-1 par zone géographique du modèle de Microsoft sur la base de données COCOWU avec l'ensemble de classes d'intersection.

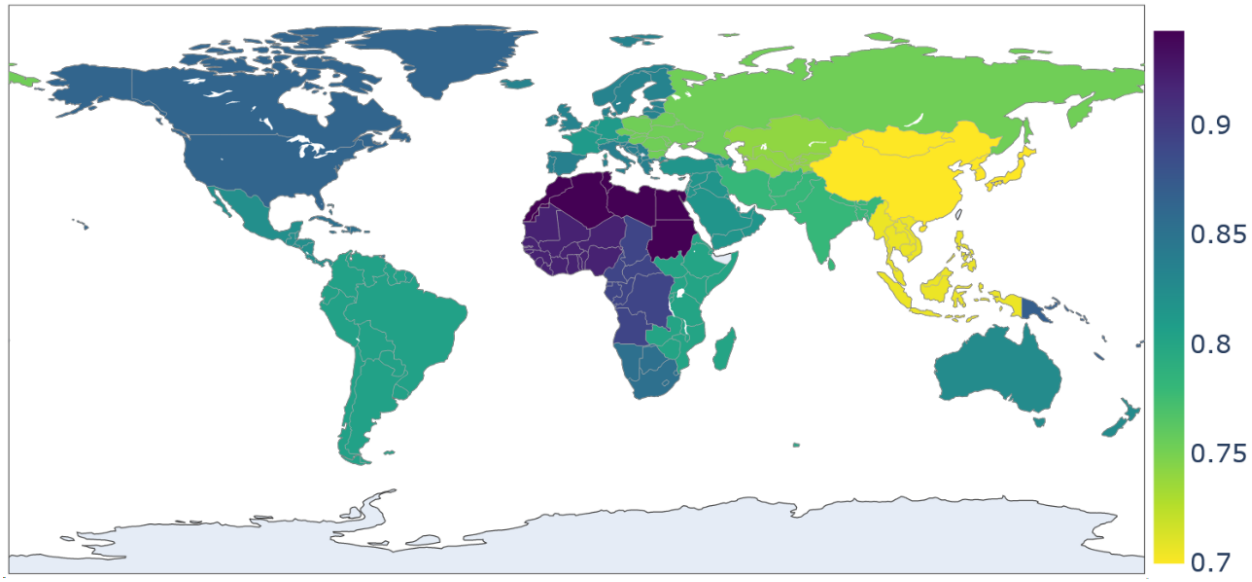


Figure 55: Précision top-5 par zone géographique du modèle de Microsoft sur la base de données COCOWU avec l'ensemble de classes d'intersection.

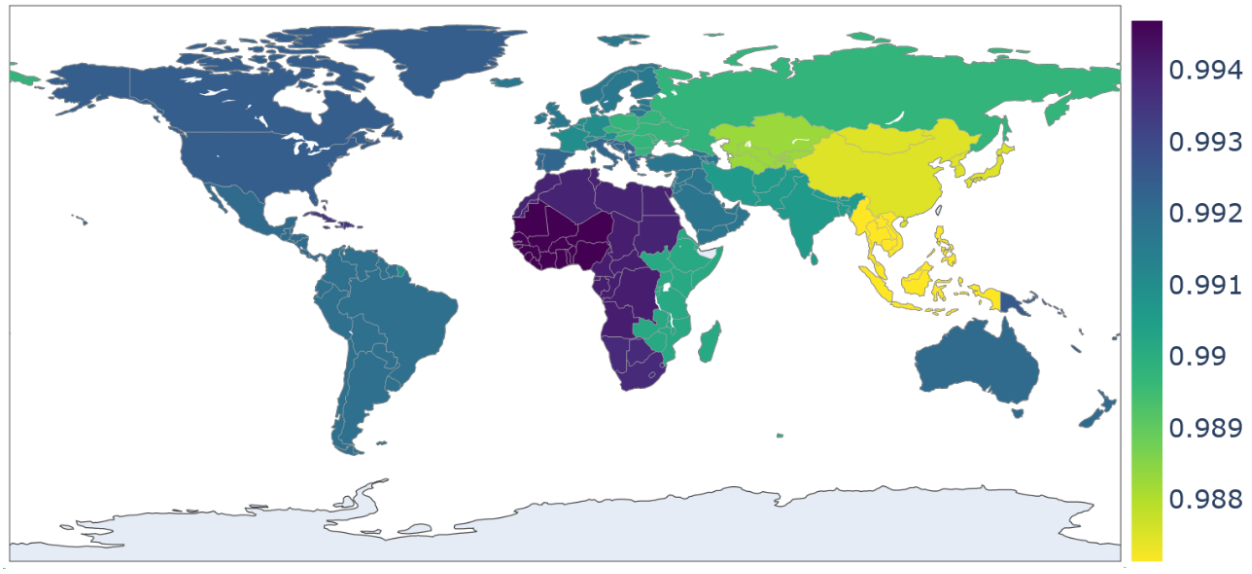


Figure 56: Précision top-1 par zone géographique du modèle de Microsoft sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

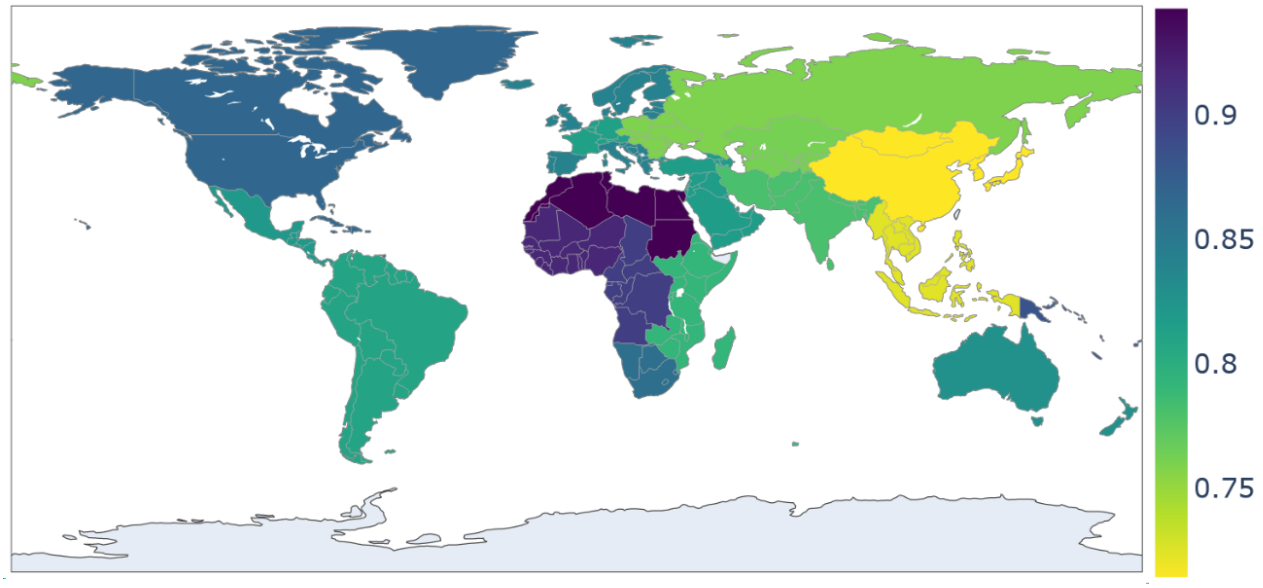


Figure 57: Précision top-5 par zone géographique du modèle de Microsoft sur la base de données COCOWU avec l'ensemble de classes par association manuelle.

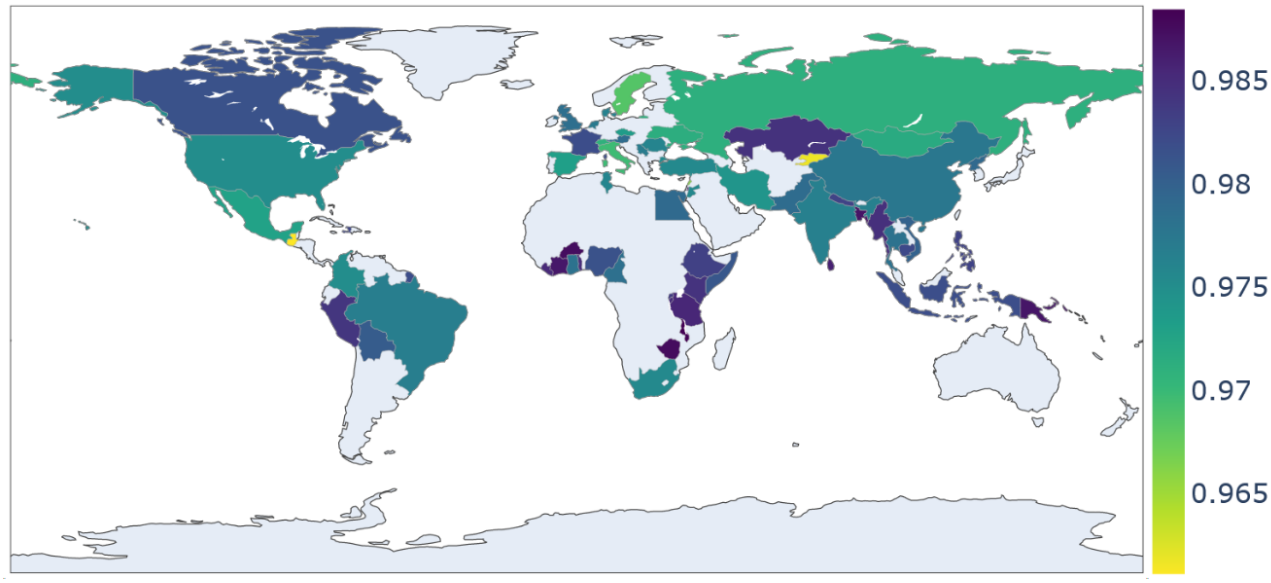


Figure 58: Précision top-1 par pays du modèle CSRA sur la base de données DSD avec l'ensemble de classes d'intersection

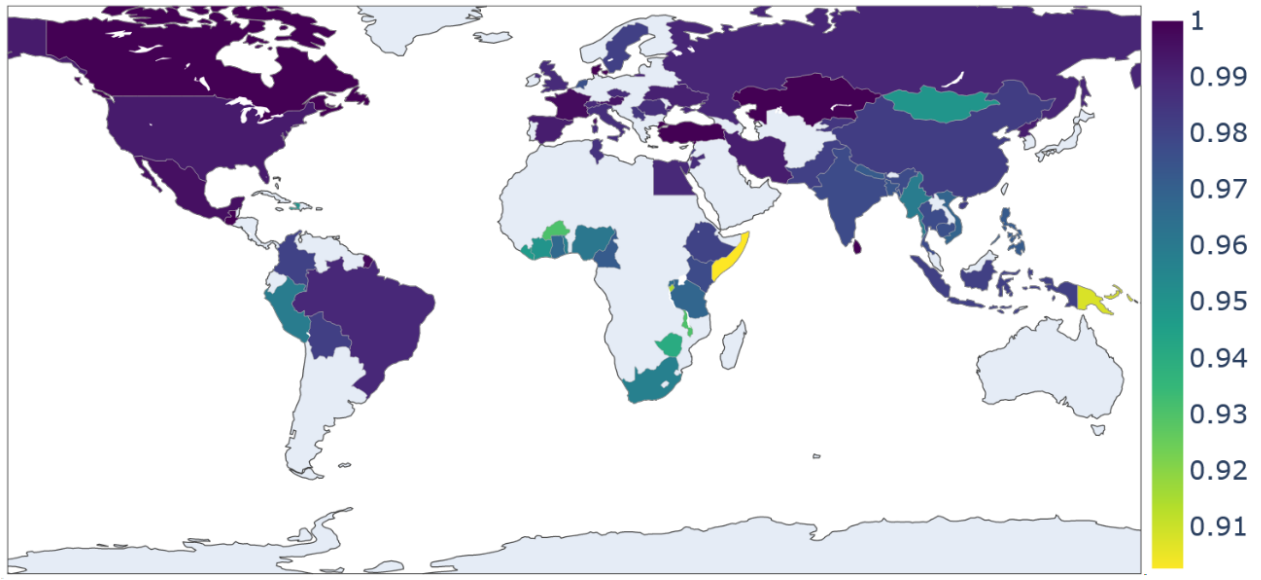


Figure 59: Précision top-5 par pays du modèle CSRA sur la base de données DSD avec l'ensemble de classes d'intersection

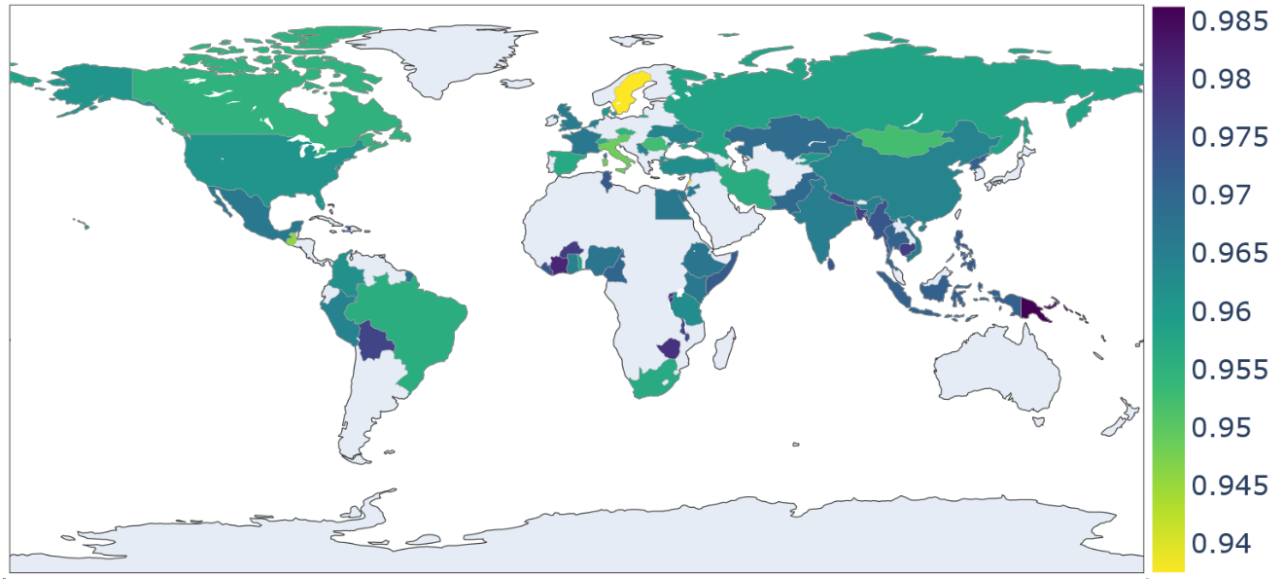


Figure 60: Précision top-1 par pays du modèle CSRA sur la base de données DSD avec l'ensemble de classes par association manuelle

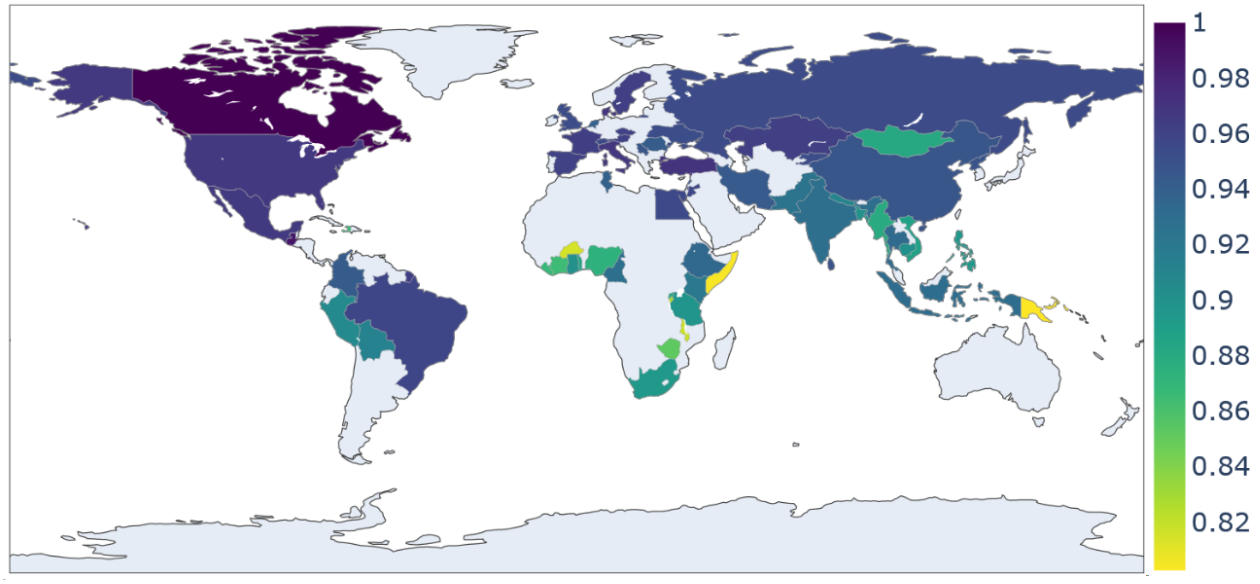


Figure 61: Précision top-5 par pays du modèle CSRA sur la base de données DSD avec l'ensemble de classes par association manuelle

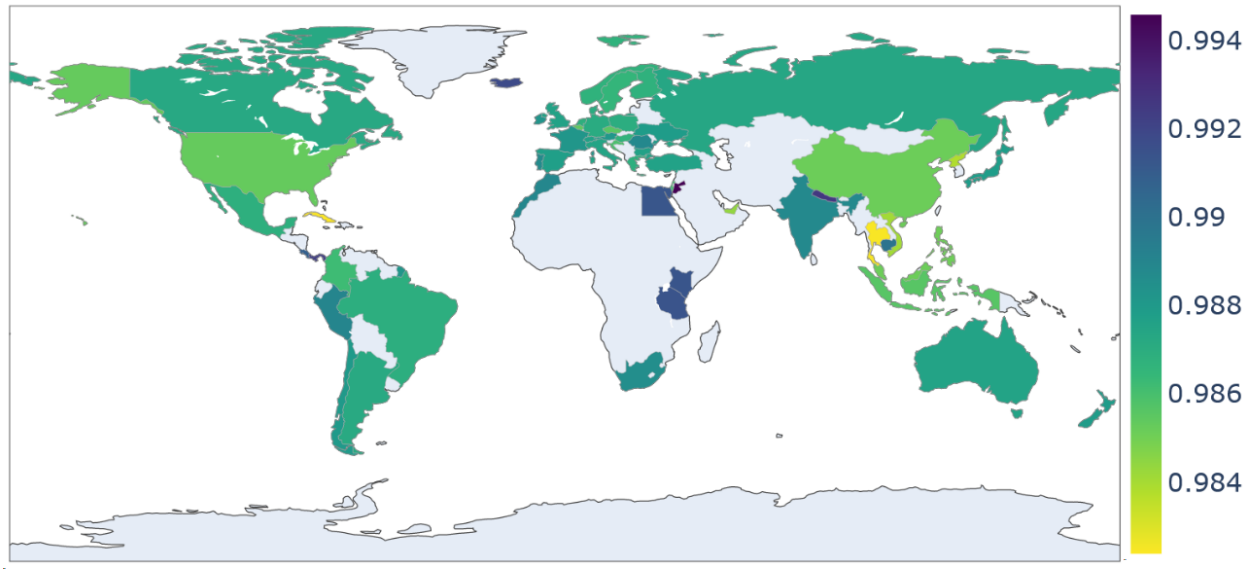


Figure 62: Précision top-1 par pays du modèle CSRA sur la base de données GYFCC avec l'ensemble de classes d'intersection

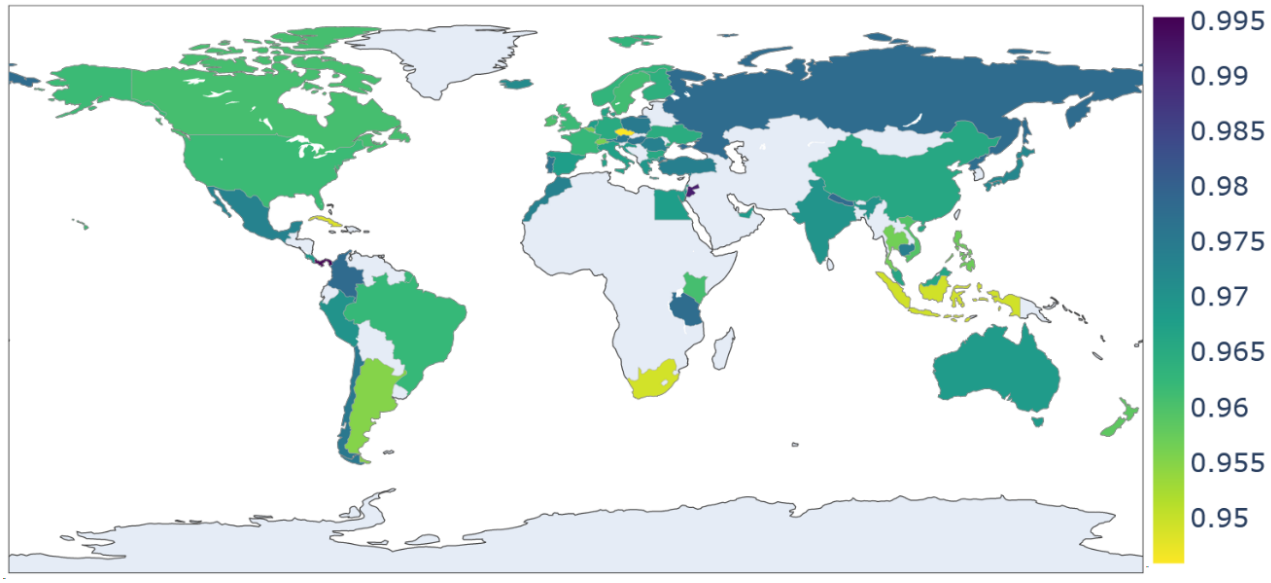


Figure 63: Précision top-5 par pays du modèle CSRA sur la base de données GYFCC avec l'ensemble de classes d'intersection

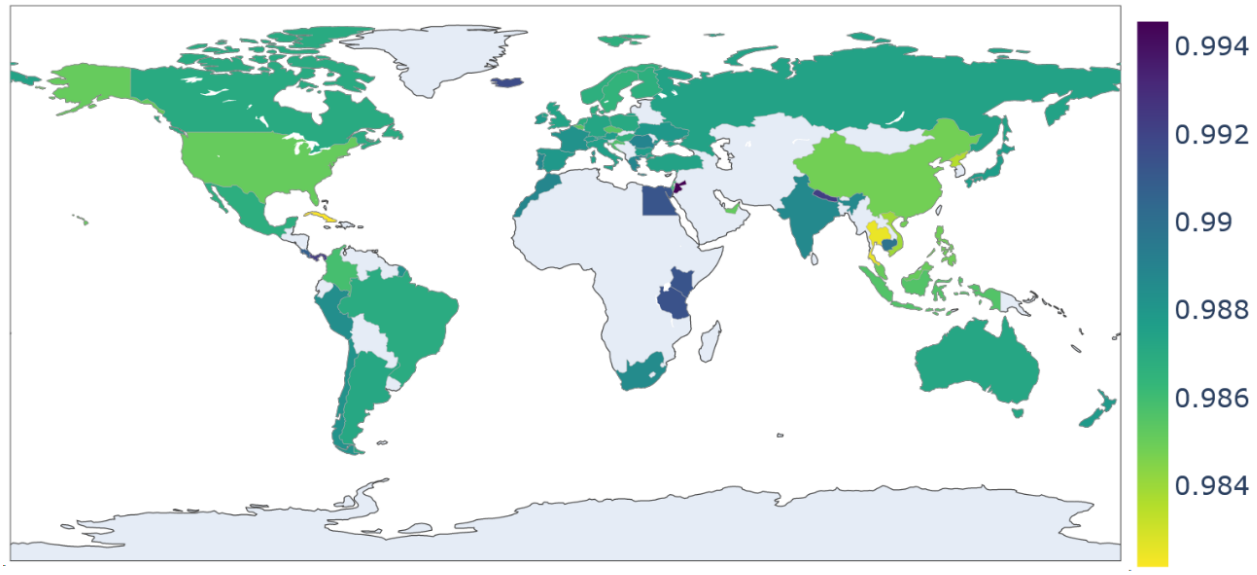


Figure 64: Précision top-1 par pays du modèle CSRA sur la base de données GYFCC avec l'ensemble de classes par association manuelle

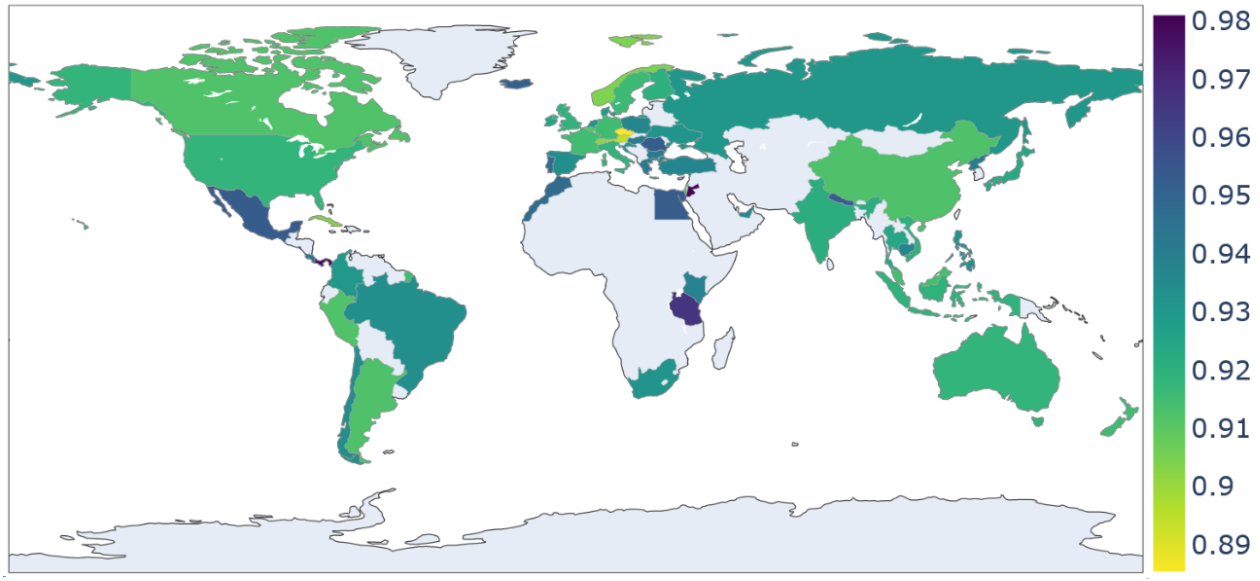


Figure 65: Précision top-5 par pays du modèle CSRA sur la base de données GYFCC avec l'ensemble de classes par association manuelle

Ces cartes présentent l'impact du biais géographique sur les différents systèmes de vision testés selon le protocole STI. Si on ne s'attend pas à ce que chaque modèle implémenté dans ces systèmes présente exactement les mêmes résultats, certaines tendances, appuyées par les travaux précédents, sont attendues : une meilleur performance globalement dans les pays du Nord que dans les pays du Sud pour chacun de ces systèmes.

A modèle fixé, la comparaison de ces cartes permet de constater l'impact d'une variation de l'ensemble de classes ou de la métrique de performance du modèle sur le témoignage du biais géographique. A métrique ou ensemble de classé fixé, la comparaison de ces cartes permet d'évaluer la pertinence de la métrique ou de l'ensemble de classes pour l'évaluation du biais géographique selon le protocole STI. Ces analyses sont réalisées dans la section suivante.

6.4.3 Synthèse des résultats

Une première observation des cartes de la section précédente est sans appel : les tendances observées ne sont pas celles attendues. La plupart des cartes présentent des performances soit mixtes dans les différents pays du monde ou zones géographiques (comme par exemple la carte 44), soit meilleures globalement dans les pays du Sud (par exemple en figure 58). Deux cartes seulement présentent clairement la tendance attendue en terme de variation de performances entre Nord et Sud : celles présentant les résultats du modèle CSRA sur la base de données DSD et la métrique de précision top-5 (cartes 59 et 61). Cette première observation remet en question la méthodologie utilisée, et encourage d'autres observations plus fines.

A modèle fixé, les résultats ne sont pas forcément consistants, même sur une même base de données, témoignant de l'importance des choix de métrique de performance ou d'ensemble. Les résultats soulignent que ces deux paramètres peuvent avoir de l'importance. Si les modèles se montrent très performants de manière générale, atteignant au dessus de 70% de précision au minimum, cela est en partie du à la construction des métriques de précision top-1 et top-5, qui sont lissées par le nombre important de classes considérées, puisque chaque vrai négatif dans chaque prédiction est compté.

A modèle et métrique fixés, on peut observer une certaine consistance à travers les jeux de données, même si cette consistance n'est pas uniformément vérifiée. Le choix d'un ensemble de classes particulier a de l'impact sur la performance d'un modèle sur un pays ou une zone géographique, les

performances moyennes observées diminuant en passant de l'intersection à l'association manuelle pour les ensembles de classes. Mais les tendances entre performances Nord et Sud ne sont pas particulièrement impactées, dénotant dans notre cas une absence d'impact sur le témoignage du biais géographique.

A modèle et ensemble de classes fixés, on retrouve plus d'inconsistance. Ainsi, les tendances des performances du modèle CSRA varient sur COCOWU de la précision top-1 à la précision top-5 sur les cartes 44 et 45. C'est le même phénomène pour les modèles d'Amazon, de Google et de Microsoft sur la base de donnée COCOWU, dont les tendances s'accordent avec le choix de la métrique.

Ces observations nous fournissent une analyse intéressante, soulignant l'importance du choix de la métrique de performance, et présentant le faible impact du choix de l'ensemble de classes dans notre cas précis. Mais c'est la tendance générale observée sur les cartes qui frappe le plus ces analyses, puisque ces cartes ne témoignent pas d'une diminution des performances des systèmes de vision par ordinateur dans les pays les moins représentés ; et appuient même pour certaines cartes une hypothèse contraire. Au final, il est impossible d'observer le biais géographique des systèmes sélectionnés avec la méthodologie employée. Ces observations appuient la présence des biais concomitants observés lors de l'enquête visuelle, et soulignent la complexité de la caractérisation du biais géographique.

6.5 Conclusion

L'exploration visuelle des classes "bus" et "train" de COCOWU avec les prédictions de CSRA a permis de mettre au jour la présence de biais concomitants, influençant les prédictions du modèle sur d'autres axes que l'opposition occident / hors occident proposé. Ces biais ne sont pas exempts de composante géographique, mais interviennent directement dans les bases de données en tant que biais historique ou de représentation. De fait, ils ne sont pas les biais que l'on cherche à mesurer, et perturbent la caractérisation souhaitée par l'application du protocole STI. Le constat des résultats contradictoire n'est pas restreint au modèle CSRA, mais s'étend aux API de Google, Microsoft et Amazon. Le constat n'est pas non plus restreint à COCOWU puisqu'il est aussi réalisé pour DSD et GYFCC. La diversité des approches dans les conception de ces bases de données et des modèles utilisés plaide pour un phénomène qui touche tous les systèmes de vision. L'utilisation de cartes pour essayer d'observer le phénomène de variation de performance géographique à une granularité plus fine rencontre encore le même problème, soulignant son importance et la présence uniforme de biais concomitants dans les bases de données d'images, même inclusives.

Il est nécessaire de comprendre pourquoi nos résultats sont si mitigés quand les expériences de DeVries et al. (2019) donnent au contraire une tendance claire. Nous faisons l'hypothèse que le passage de la classification mono-classe à la classification multi-classe est une source importante de cette disparité. Notamment, la classification mono-classe souffre moins du problème du glissement d'environnement, puisque seule la classe annotée sur une image est pénalisée par une mauvaise prédiction. Les biais concomitants ont donc plus d'influence sur les métriques en classification multi-classe qu'en classification mono-classe. Ce phénomène rend difficile la caractérisation du biais géographique dans des données images. De meilleurs protocoles sont nécessaire si l'on veut être capables de témoigner en amont de la capacité d'un modèle à être déployé dans une zone géographique particulière. Ces travaux exploratoires permettent de mieux cerner la problématique et d'avancer vers cette caractérisation en mettent en avant les défis et les écueils dans le domaine du glissement géographique.

Chapter 7

Conclusion et perspectives

Chapter Summary

7.1 Conclusion

Notre travail porte sur l'évaluation du biais géographique dans les systèmes de vision. L'objectif est de pouvoir permettre à n'importe quel acteur d'estimer la capacité d'un modèle à performer dans un contexte géographique donné. Les précédentes contributions dans ce domaine ont amené les conclusions suivantes :

- Les bases de données d'images manquent d'inclusivité et ne représentent pas une bonne partie des localités et des cultures.
- Les modèles à l'état de l'art en vision par ordinateur ne sont pas capables de généraliser sur ces données dans le cadre de la classification mono-classe.
- Les systèmes de vision par ordinateur sont conceptualisés avec des objectifs occidentaux, et manquent de pertinence hors de leur contexte.

Notre première contribution[17] présentée en partie 2.5.5 met en avant les approches en co-construction pour le développement de modèles en collaboration avec tous les acteurs concernés par une problématique, limitant les effets du troisième point. Dans ces travaux, une construction itérative incluant les acteurs concernés par les problématiques halieutiques a permis le développement d'un modèle incluant les connaissances et contributions de chacun, assurant la pertinence et l'inclusivité du modèle.

Une seconde contribution de cette thèse est l'implémentation d'une adaptation de méthodologie de génération de base de données pour rendre cette dernière inclusive[18]. Cet effort contribue à souligner le manque d'inclusivité des bases de données d'images, et propose des solutions pour pallier ce manque, à travers deux adaptations particulières :

- L'ajout de termes géographiques dans les requêtes aux plateformes d'images en ligne, forçant la présence de données représentatives dans les données collectées.
- L'ajout de métadonnées géographiques aux images collectées, permettant de caractériser les performances par zone géographique et soutenant ainsi les efforts d'évaluation du biais géographique.

Cette contribution a été accompagnée de la publication de la base de données COCO World URLs[16], comprenant des URLs pointant vers des images sur la plateforme Flickr et réparties par zone géographique. 400 images issues de ces URLs ont été annotées par nos soins pour chacune des 23 zones géographiques utilisées, et ces annotations sont partagées sous forme de tableur au sein de la base de données.

Nous examinons dans la suite de nos travaux les performances des systèmes de vision à travers un modèle à l'état de l'art et des API pour la classification multi-classe, et tentons de caractériser le biais occidental de ces solutions. Notre contribution[19] souligne les défis rencontrés dans cette caractérisation, à nommer :

- La qualité des données et des métadonnées des bases de données contenant des informations géographiques.
- La problématique des différents ensembles de classes lors de l'emploi de modèles fixés sans ré-entraînement dans différentes situations
- Le choix de métriques de performances appropriées.

Ces travaux mettent en lumière la complexité du biais géographique et la complexité de sa caractérisation. Ce constat est celui qui nous pousse dans notre thèse à développer une méthodologie pour l'étude des biais qui prend en compte cette complexité. Le protocole introduit en section 3 permet la prise en compte de cette complexité, en intégrant une triple dimension aux biais.

Le protocole STI décompose un biais en trois composantes : Source, Type, Impact. Cette décomposition fait état des multiples formes que prennent les biais et de la manière dont ces derniers sont décrits, identifiés, et mitigés dans la littérature. Cette décomposition n'est pas triviale ; elle demande d'identifier pour un biais spécifique l'ensemble des sources possibles de ce biais, les différentes manières dont il intervient dans un système, et les différentes manière de reporter son impact. Pour chaque biais identifié, ces efforts peuvent néanmoins être réemployés, discutés, contestés. Ce protocole apporte plusieurs bénéfices à l'étude et la caractérisation de biais :

- Une vue panoramique du biais, offrant une meilleure compréhension des enjeux et vues autour de ce dernier. Le protocole STI aide à la prise de conscience que les biais ne sont pas uniquement des artefacts techniques.
- Il incite à la formalisation d'un cadre de travail particulier pour la mitigation d'un biais, en identifiant précisément les enjeux et limites du cadre proposé. Ceci permet aussi d'identifier les situations dans lesquelles une stratégie de mitigation peut être ou ne pas être efficace.
- Il permet de distinguer entre elles les différentes approches proposées dans la littérature en définissant de manière compréhensive ces dernières.

Son caractère évolutif et participatif en fait un outil idéal pour le suivi des initiatives de la communauté scientifique dans le domaine de la lutte contre les impacts négatifs suivant les déploiements de systèmes algorithmiques.

Cette méthodologie est mise à l'épreuve sur des données synthétiques en section 4. Puisqu'aucune base de donnée synthétique d'images ne contient des variations géographiques, nous créons une telle base de données, appelée GeoMNIST, s'inspirant des modifications de la base de données MNIST réalisées par RotatedMNIST et ColoredMNIST. GeoMNIST propose d'associer à chaque donnée de MNIST un pays, et d'effectuer des transformations sur cette donnée en fonction du pays qui lui est associée. Une base de donnée GeoMNIST est ainsi définie par la répartition des données dans les différents pays du monde. Trois bases de données sont ainsi crées, appelées GeoMNIST-A, GeoMNIST-B et GeoMNIST-C, respectant respectivement les répartitions de population dans le monde, une répartition uniforme, et la répartition des base de données d'images tel qu'Imagenet ou MS COCO. Ces trois bases de données sont figées à l'aide de tableurs qui associent à chaque donnée non seulement un pays, mais fixe aussi les transformations associées à la donnée, afin d'assurer la reproductibilité des

résultats de nos travaux. Le protocole STI est ensuite déployé sur ces bases de données GeoMNIST, dans le but de caractériser le biais d'évaluation de modèles étant entraînés sur ces différentes bases de données. Cette expérimentation met en lumière les multiples aspects à prendre en compte lors de la caractérisation d'un biais, et les choix à réaliser lors de l'implémentation du protocole STI. Le biais géographique des modèles entraînés sur les différentes bases de données GeoMNIST est caractérisé et décortiqué, démontrant l'intérêt de la méthodologie proposée.

Le même protocole est mis à l'épreuve sur des données réelles dans le cadre de la caractérisation du biais occidental d'un modèle à l'état de l'art pour la classification multi-classe d'objets et de concepts de la vie de tous les jours en section 5. L'objectif est de pouvoir évaluer à quel point les systèmes de vision par ordinateur peuvent être déployés sans adaptation dans les différents contextes géographiques mondiaux. Nous procédons à des évaluations automatiques et manuelles, pour nous affranchir des biais d'évaluation inhérent au choix d'une pratique en particulier. Ces comparaisons permettent dans le cas de nos expérimentations de mettre à jour des problèmes liées aux données et aux annotations. Les résultats obtenus se montrent surprenants, et argumentent en faveur de l'absence de biais occidental dans le modèle CSRA sur les données de COCO World URLs. Une analyse plus fine des prédictions de CSRA sur les classes de transport montre une hétérogénéité dans les performances du modèle relativement au caractère occidental ou non des données. Ces résultats allant à l'encontre des théories et des précédents travaux sur le biais géographiques, des analyses plus sérieuses sont nécessaires pour mettre à jour les phénomènes à l'origine de ces résultats.

Ces analyses sont menées en section 6. Pour les réaliser, nous utilisons un outil pour la visualisation et l'analyse des prédictions de modèles sur des données images. Par soucis de concision, nous sélectionnons alors parmi les classes de transport, les classes "bus" et "train", possédant respectivement les tendances les plus contraires et en accord avec celles attendues en terme de performance à caractère occidental pour le modèle CSRA. Les images associées à ces classes sont analysées une par une, et l'opérateur interprète les concepts qui influencent les prédictions du modèle CSRA relatives à ces classes. Nous interprétons de ces analyses que les prédictions du modèle sont liés à une combinaison de phénomènes : la sensibilité des prédictions à des proximités entre concepts, et la variation de la répartition de ces concepts entre données occidentales et non occidentales. Cette expérience nous permet de démontrer que les prédictions des modèles sont sensibles à d'autres phénomènes que le biais géographiques, moins complexes, plus restreints. Ces différents phénomènes s'entremêlent, ayant des impacts tantôt dans le sens du biais géographique, tantôt dans le sens contraire, tantôt sur un axe tout à fait différents. Nous appelons ces phénomènes biais concomitants, et argumentons que lors du déploiement de modèles sur des données réelles, ces biais concomitants empêchent la caractérisation du biais géographique.

En voulant caractériser le biais géographique des systèmes de vision disponibles et à l'état de l'art, nous avons constaté les nombreux défis qui s'opposent aujourd'hui à cette caractérisation. Si cette thèse ne parvient pas à répondre à l'exercice, elle propose de nouveaux outils pour avancer dans cette direction, et ouvre la voie à de nouvelles réflexions et à des travaux participatifs qui ont vocation à prolonger la recherche académique pour la compréhension et la caractérisation du biais géographique dans les bases de données d'images et systèmes de vision.

7.2 Perspectives

Ce travail ouvre la voie à une large gamme de potentielles nouvelles questions de recherche, qui n'ont pu être explorées dans le temps de la thèse. Ces questions se concentrent sur les thèmes suivants :

- La caractérisation d'un biais. Si la lutte contre les impacts des biais sont l'objet de nombreux travaux, la compréhension effective de ce qu'est un biais et du cadre dans lequel il évolue est essentiel pour que cette lutte soit pertinente.

- La génération de bases de données inclusives. C'est une question qui gagne en importance et qui inclue des problématiques qui sortent de la sphère technique.

- Les biais concomitants. Si la notion n'est pas nouvelle, la détermination de leur impact et le rôle qu'ils jouent dans nos expériences nécessite de plus amples recherches.
- le développement et l'utilisation d'IA dans les pays du Sud. Nous nous heurtons aux limites des capacités technologiques et financières dans la recherche au Sud, et questionnons la pertinence de certains approches dans le contexte des pays en développement.

La caractérisation des biais est un élément essentiel de la lutte contre les impacts nocifs constatés lors du déploiement de modèles d'IA. Pour le moment, la lutte s'est concentrée sur les notions de justice algorithmique et d'adaptation ou de généralisation de domaines. Mais certaines initiatives soulignent que ces stratégies sont parfois peu effective et peinent à se montrer probantes hors laboratoires. Pour des applications concrètes, il est nécessaire de comprendre dans quel cadre un biais peut avoir un impact. La caractérisation du biais permet cette identification du cadre et de la portée du biais, pour proposer des stratégies de lutte plus effectives. Le protocole STI proposé décompose un biais en trois éléments : sources, types et impacts, qui regroupent différentes approches pour cerner les caractéristiques d'un biais. Cette décomposition peut encore aller plus loin en intégrant d'autres paramètres, comme son niveau d'abstraction ou sa nature anthropomorphique. L'intégration de nouvelles composantes peut être l'objet de futures recherches. Par manque de temps, le protocole proposé n'a pas été employé autrement que sur le biais géographique, et son application à d'autres applications pourraient permettre de le mettre plus à l'épreuve, et d'entamer des constructions collaboratives autour de divers biais. Une application à la justice algorithmique paraît une suite logique pour le développement et l'adoption du protocole STI. De telles implémentations pourront faire l'objet de travaux futurs.

De même, les bases de données synthétiques GeoMNIST conceptualisées et implémentées durant cette thèse ne sont utilisées que dans le cadre de l'implémentation du protocole STI. Ces dernières pourraient être utilisées dans le cadre de tests pour des stratégies d'adaptation ou de généralisation de domaine géographique, ou dans le cadre de la justice algorithmique. De nouveaux benchmarks pourraient être développés à partir de la méthodologie utilisée pour générer les bases de données GeoMNIST, permettant de tester des stratégies d'adaptation sur des bases de données d'images à large volume, ce qui est difficile à obtenir en situation réelle pour des images comportant des métadonnées géographiques. Il est néanmoins notable que les bases de données synthétiques ont du mal à refléter la complexité des bases de données d'images réelles. La génération de bases de données d'image réelles inclusives est une tâche qui requiert la prise en compte de nombreux paramètres, et de comprendre comment chaque processus intègre des biais. Une revue compréhensive des différentes initiatives et des biais qui sont identifiés dans ces initiatives pourrait permettre de proposer une vue d'ensemble pour la génération de données inclusives. Plus de recherche sur les pratiques autour des bases de données d'images sont nécessaires pour comprendre comment une image code une multitude d'informations et de concepts complexes.

Cette complexité est illustrée par les biais concomitants identifiés dans le dernier chapitre. La distinction d'un biais avec ses composantes introduit de nombreuses questions : comment mesurer l'impact d'une composante par rapport à un biais englobant ? Peut-on isoler une composante parmi les autres ? Est-il possible de tester la présence d'une composante en particulier ? L'introduction de cette nouvelle granularité intensifie la complexité de la caractérisation d'un biais. Si cela souligne les nombreuses formes que peut prendre un biais, tenter d'intégrer trop de complexité dans un protocole peut rendre ce dernier impraticable, et donc inutile. Le juste milieu entre complexité et utilité d'un protocole de caractérisation de biais est une question importante pour la suite. Le biais géographique est une bonne illustration de ce phénomène : un protocole simple permet de le caractériser dans des cas simples, mais est incapable d'atteindre ses objectifs dans un cas plus complexe. Si nos travaux permettent d'illustrer comment le biais géographique intervient de manière inopinée dans des évaluations de prédictions de modèles, ils ne permettent pas de caractériser le biais géographique dans les systèmes de vision, seulement de préciser la complexité de la tâche. Il est nécessaire de poursuivre les efforts réalisés dans cette thèse pour permettre la caractérisation du biais géographique dans les

systèmes de vision, et ainsi de pouvoir déterminer les capacités de ces systèmes à être déployés dans diverses géographies.

Enfin, nous nous heurtons durant cette thèse à de nombreuses problématiques systémiques au Sud pour le développement de l'IA : le manque de moyens, les capacités de calcul réduites, le défi de l'accès aux données. Surmonter ces problèmes pour pouvoir proposer une recherche de qualité est complexe, et demande des ressources et une énergie colossale, sans commune mesure avec les conditions de travail dans les laboratoires du Nord. Cette situation crée un schisme géographique dans l'accès à la capacité de développement de modèles d'IA, qui creuse les inégalités déjà observées dans les modèles prédictifs. De fait, nous avons rencontré des impossibilités de financement de certaines de nos expériences, et des lenteurs dans les expérimentation réalisées dues à des capacités de calcul accessibles réduites. De telles observations mènent à un questionnement de la pertinence des techniques reposant sur des systèmes énergivores, financièrement coûteux et reposent sur la disponibilité de masses de données dans des contextes inappropriés. Quel avenir l'IA a-t-il dans les pays du Sud, où les conditions ne sont pas réunies pour un développement local ? Ne risque-t-on pas de promouvoir des systèmes conçus par les pays du Nord, avec leurs propres considérations, et donc peu appropriés aux problématiques de terrain ? La légitimité de la technologie est une question peu explorée mais qui mérite une attention particulière, si l'on veut éviter que les déploiements d'IA se résument à des processus aliénants et délétères.

7.3 Mot de la fin

Dans cette thèse, nous contribuons à l'évaluation du biais géographique dans les systèmes de vision, en soulignant la complexité de cette tâche et en identifiant les points bloquants dans les méthodologies actuelles. Nous montrons comment des évidences comme "les modèles marchent moins bien dans les pays du Sud" peuvent être mis en défaut par des évaluations rapides, et combien il est important de construire des chaînes de traitement capables de caractériser le glissement géographique dans les modèles d'IA. Nos travaux permettent d'identifier quels mécanismes empêchent cette caractérisation et mettent en avant des outils dont la communauté scientifique peut se saisir pour avancer vers une telle caractérisation.

La difficulté principale de la recherche menée a été le caractère surprenant des résultats obtenus, qui ont transformé ce qui devait être une simple transposition méthodologique en véritable épopée pour la compréhension du phénomène du glissement géographique, de ses sources, et de sa nature. Cette recherche est au finale source d'un apport scientifique dont nous sommes fiers, et d'une satisfaction certaine puisqu'elle a permis d'explorer un sujet encore trop peu traité à ce jour. Les nombreuses perspectives ouvertes par nos travaux sont une grande source de motivation pour les années à venir.

Bibliography

Abdalla, M. and Abdalla, M. (2021). The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 287–297, Virtual Event USA. ACM.

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., and Sadagopan, S. (2021). Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 329–341.

Abuhamad, G. and Rheault, C. (2020). Like a Researcher Stating Broader Impact For the Very First Time. In *Navigating the Broader Impacts of AI Research Workshop*.

Agarwal, S. (2021a). Trade-Offs between Fairness and Interpretability in Machine Learning. In *Workshop on AI for Social Good, IJCAI 2021*.

Agarwal, S. (2021b). Trade-Offs between Fairness and Privacy in Machine Learning. In *Workshop on AI for Social Good, IJCAI 2021*.

Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., and Brundage, M. (2021). Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications.

Alemdar, H., Leroy, V., Prost-Boucle, A., and Petrot, F. (2017). Ternary neural networks for resource-efficient AI applications. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2547–2554, Anchorage, AK, USA. IEEE.

Alonso, C. (2020). *Will the AI Revolution Cause a Great Divergence?* International Monetary Fund, S.l.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260, Virtual Event Canada. ACM.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2020). Invariant Risk Minimization.

Atwood, J., Halpern, Y., Baljekar, P., Breck, E., Sculley, D., Ostyakov, P., Nikolenko, S. I., Ivanov, I., Solovyev, R., Wang, W., and Skalic, M. (2020). The Inclusive Images Competition. In Escalera, S. and Herbrich, R., editors, *The NeurIPS '18 Competition*, pages 155–186. Springer International Publishing, Cham.

Avila, R. (2022). Against Digital Colonialism. *Platforming Equality: policy challenges for the digital economy*, 1:13.

Babenko, B., Hersh, J., Newhouse, D., Ramakrishnan, A., and Swartz, T. (2017). Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images,

With an Application in Mexico. In *Workshop on Machine Learning for the Developing World, NeurIPS 2017*.

Balayn, A., Kulynych, B., and Guerses, S. (2021). Exploring Data Pipelines through the Process Lens: A Reference Model for Computer Vision. In *Beyond Fair Computer Vision Workshop*.

Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W., and Hildebrandt, M. (2019). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In *Being Profiled*, pages 84–89. Amsterdam University Press.

Bayet, T. (2023). COCO-style geographically unbiased image dataset for computer vision applications.

Bayet, T., Brochier, T., Cambier, C., Bah, A., Denis, C., Thiam, N., and Zucker, J.-D. (2021). A Machine Learning approach to improve the monitoring of Sustainable Development Goals : A case study in Senegalese artisanal fisheries. In *CNIA 2021 : Conférence Nationale En Intelligence Artificielle,*, page 8, Bordeaux, France.

Bayet, T., Denis, C., Bah, A., and Zucker, J.-D. (2022). Distribution Shift nested in Web Scraping : Adapting MS COCO for Inclusive Data. In *ICML Workshop on Principles of Distribution Shift 2022*, Baltimore, United States.

Bayet, T., Denis, C., Bah, A., and Zucker, J.-D. (2023). Les défis du glissement de contexte géographique. In *Plate-Forme d'Intelligence Artificielle 2023, Journée Résilience et IA*, Strasbourg, France.

Beery, S. (2021). Scaling Biodiversity Monitoring for the Data Age. *XRDS: Crossroads, The ACM Magazine for Students*, 27(4):14–18.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

Benjamin, G. (2021). Mapping CV as an Assemblage of (Unfair) Sociotechnical Relations. In *Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision, CVPR 2021*, page 5.

Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., and Shee, A. (2019). TOWARDS THE STANDARDIZATION OF DATA LICENSES. In *AI for Social Good Workshop, ICLR 2019*.

Besson, D. (2021). En attendant les robots. Enquête sur le travail du clic. Antonio Casilli, Seuil, Paris (Collection La couleur des idées), 2019. *Management international*, 25(5):223.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and van den Oord, A. (2020). Are we done with ImageNet? In *Workshop - ImageNet: Past, Present, and Future, NeurIPS 2021*.

Bezuidenhout, L., Kelly, A. H., Leonelli, S., and Rappert, B. (2017). '$100 Is Not Much To You': Open Science and neglected accessibilities for scientific research in Africa. *Critical Public Health*, 27(1):39–49.

Birhane, A. (2020). Algorithmic Colonization of Africa. *SCRIPTed*, 17(2):389–409.

Birhane, A. and Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546, Waikoloa, HI, USA. IEEE.

Blumenstock, J. (2018). Don't forget people in the use of big data for development. *Nature*, 561(7722):170–172.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the Opportunities and Risks of Foundation Models.

Bordia, S. and Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Brandt, J., Ertel, J., Spore, J., and Stolle, F. (2023). Wall-to-wall mapping of tree extent in the tropics with Sentinel-1 and Sentinel-2. *Remote Sensing of Environment*, 292:113574.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Buolamwini, J. and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Busto, P. P. and Gall, J. (2017). Open Set Domain Adaptation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 754–763, Venice. IEEE.

Caton, S. and Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38.

Chavez, H., Albornoz, M. B., and Martín, F. (2022). 'Big data' Research: A Bibliometric Analysis of the Scopus Database, 2009–2019. *Journal of Scientometric Research*, 11(1):64–78.

Chen, H., Lindshield, S., Ndiaye, P. I., Ndiaye, Y. H., Pruetz, J. D., and Reibman, A. R. (2023a). Applying Few-Shot Learning for In-the-Wild Camera-Trap Species Classification. *AI*, 4(3):574–597.

Chen, Z., Wu, M., Chan, A., Li, X., and Ong, Y.-S. (2023b). Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges [Review Article]. *IEEE Computational Intelligence Magazine*, 18(2):60–77.

Choi, B. and Kamalu, J. (2021). Crowd-Sourced Road Quality Mapping in the Developing World. In *Workshop on AI for Social Good, IJCAI 2021*, page 5.

Cichos, F., Gustavsson, K., Mehlig, B., and Volpe, G. (2020). Machine learning for active matter. *Nature Machine Intelligence*, 2(2):94–103.

Clark, W. C. and Harley, A. G. (2020). An Integrative Framework for Sustainability Science. In *Sustainability Science: A Guide for Researchers*. PubPub, 1 edition.

Coleman, D. (2019). Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws. *Michigan Journal of Race & Law*, 1(24.2):417.

Cvitkovic, M. (2019). Some requests for machine learning research from the east african tech scene. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 37–40, Accra Ghana. ACM.

Delisle, L., Kalaitzis, A., Majewski, K., de Berker, A., Marin, M., and Cornebise, J. (2019). A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. In *Workshop on AI for Social Good, NeurIPS 2019*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

DeVries, T., Misra, I., Wang, C., and van der Maaten, L. (2019). Does Object Recognition Work for Everyone? In *CVPR Workshop on Computer Vision for Global Challenges (CV4GC), 2019*.

Dibia, V. (2018). COCO-Africa: A Curation Tool and Dataset of Common Objects in the Context of Africa. In *2nd Black in AI Workshop, NeurIPS 2018*.

Dubey, A., Ramanathan, V., Pentland, A., and Mahajan, D. (2021). Adaptive Methods for Real-World Domain Generalization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14335–14344, Nashville, TN, USA. IEEE.

Dutta, A. and Zisserman, A. (2019). The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2276–2279, Nice France. ACM.

Duvenhage, B. (2019). Short Text Language Identification for Under Resourced Languages. In *Workshop on Machine Learning for the Developing World, NeurIPS 2019*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, Cambridge Massachusetts. ACM.

Efremova, N., West, D., and Zausaev, D. (2019). AI-Based Evaluation of the SDGs: The Case of Crop Detection With Earth Observation Data. *SSRN Electronic Journal*.

Elsahar, H. and Gallé, M. (2019). To Annotate or Not? Predicting Performance Drop under Domain Shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338.

Famularo, J., Hensellek, B., and Walsh, P. (2021). Data Stewardship: A Letter to Computer Vision from Cultural Heritage Studies. In *Beyond Fairness: Towards a Just, Equitable and Accountable Computer Vision, CVPR 2021*, page 11.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. (2021). A Brief Review of Domain Adaptation. In Stahlbock, R., Weiss, G. M., Abou-Nasr, M., Yang, C.-Y., Arabnia, H. R., and Deligiannidis, L., editors, *Advances in Data Science and Information Engineering*, pages 877–894. Springer International Publishing, Cham.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney NSW Australia. ACM.

Frajberg, D., Fraternali, P., and Torres, R. N. (2018). Crowdsourcing mountain images for water conservation. In *Workshop on Artificial Intelligence for Wildlife Conservation, IJCAI-ECAI 2018*, page 6.

Fu, A., Hosseini, M. S., and Plataniotis, K. N. (2021). Reconsidering CO2 emissions from Computer Vision. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2311–2317, Nashville, TN, USA. IEEE.

Gajane, P. and Pechenizkiy, M. (2018). On Formalizing Fairness in Prediction with Machine Learning. In *FATML 2018*.

Gargiulo, F., Fontaine, S., Dubois, M., and Tubaro, P. (2023). A meso-scale cartography of the AI ecosystem. *Quantitative Science Studies*, 4(3):574–593.

Gaviria Rojas, W., Diamos, S., Kini, K., Kanter, D., Janapa Reddi, V., and Coleman, C. (2022). The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 12979–12990. Curran Associates, Inc.

Goh, G., †, N. C., †, C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3):e30.

Gonzalez, J., Bhowmick, D., Beltran, C., Sankaran, K., and Bengio, Y. (2019). Applying Knowledge Transfer for Water Body Segmentation in Peru. In *Workshop on Machine Learning for the Developing World, NeurIPS 2019*, page 5.

Gratiot, N., Klein, J., Challet, M., Dangles, O., Janicot, S., Candelas, M., Sarret, G., Panthou, G., Hingray, B., Champollion, N., Montillaud, J., Bellemain, P., Marc, O., Bationo, C.-S., Monnier, L., Laffont, L., Foujols, M.-A., Riffault, V., Tinel, L., Mignot, E., Philippon, N., Dezetter, A., Caron, A., Piton, G., Verney-Carron, A., Delaballe, A., Bardet, N., Nozay-Maurice, F., Loison, A.-S., Delbart, F., Anquetin, S., Immel, F., Baehr, C., Malbet, F., Berni, C., Delattre, L., Echevin, V., Petitdidier, E., Aumont, O., Michau, F., Bijon, N., Vidal, J.-P., Pinel, S., Biabiany, O., Grevesse, C., Mimeau, L., Biarnès, A., Récapet, C., Costes-Thiré, M., Poupaud, M., Barret, M., Bonnin, M., Mournetas, V., Tourancheau, B., Goldman, B., Bonnet, M. P., and Michaud Soret, I. (2023). A transition support system to build decarbonization scenarios in the academic community. *PLOS Sustainability and Transformation*, 2(4):e0000049.

Gulrajani, I. and Lopez-Paz, D. (2021). In Search of Lost Domain Generalization. In *International Conference on Learning Representations*.

Gupta, A., Dollár, P., and Girshick, R. (2019). LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. (2024). Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1):99–120.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE.

Hecht, B., Wilcox, L., Bigham, J. P., Schöning, J., Hoque, E., Ernst, J., Bisk, Y., De Russis, L., Yarosh, L., Anjum, B., Contractor, D., and Wu, C. (2021). It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process.

Helber, P., Gram-Hansen, B., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V., and Bilinski, P. (2019). Generating Material Maps to Map Informal Settlements. In *Workshop on Machine Learning for the Developing World, NeurIPS 2018*.

Hestness, J., Ardalani, N., and Diamos, G. (2019). Beyond human-level accuracy: Computational challenges in deep learning. In *Proceedings of the 24th Symposium on Principles and Practice of Parallel Programming*, pages 1–14, Washington District of Columbia. ACM.

Hohman, F., Kahng, M., Pienta, R., and Chau, D. H. (2019). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. (2020). Characterising Bias in Compressed Models.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

Hryniewska, W., Czarnecki, P., Wiśniewski, J., Bombiński, P., and Biecek, P. (2021). Prevention is better than cure: A case study of the abnormalities detection in the chest. In *Workshop Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision, CVPR 2021*.

Huynh, B. Q. and Basu, S. (2020). Forecasting Internally Displaced Population Migration Patterns in Syria and Yemen. *Disaster Medicine and Public Health Preparedness*, 14(3):302–307.

Islam, M. T., Fariha, A., Meliou, A., and Salimi, B. (2022). Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification. In *Proceedings of the 2022 International Conference on Management of Data*, pages 232–246, Philadelphia PA USA. ACM.

Jain, G., Parsheera, S., and Project, C. (2021). 1.4 billion missing pieces? Auditing the accuracy of facial processing tools on Indian faces. In *Ethical Considerations in Creative Applications of Computer Vision*, page 6.

Jo, E. S. and Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, Barcelona Spain. ACM.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kalluri, T., Xu, W., and Chandraker, M. (2023). GeoNet: Benchmarking Unsupervised Adaptation across Geographies. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15368–15379, Vancouver, BC, Canada. IEEE.

Kaneko, A., Kennedy, T., Mei, L., Sintek, C., Burke, M., Ermon, S., and Lobell, D. (2019). Deep Learning For Crop Yield Prediction in Africa. In *AI for Social Good Workshop, ICML 2019*, page 5.

Kates, R. W. (2011). What kind of a science is sustainability science? *Proceedings of the National Academy of Sciences*, 108(49):19449–19450.

Katzman, J., Wang, A., Scheuerman, M., Blodgett, S. L., Laird, K., Wallach, H., and Barocas, S. (2023). Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. In *Special Track on AI for Social Impact, AAAI 2023*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Kshirsagar, V., Wieczorek, J., Ramanathan, S., and Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach. In *Workshop on Machine Learning for the Developing World, NeurIPS 2017*.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. (2020). The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Lambrecht, A. and Tucker, C. E. (2016). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*.

Leavy, S., O'Sullivan, B., and Siapera, E. (2020). Data, Power and Bias in Artificial Intelligence. In *AI for Social Good Workshop, IJCAI 2020*, page 5.

Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., and Procaccia, A. D. (2019). WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35.

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. (2017). WebVision Database: Visual Learning and Understanding from Web Data.

Li, Z., Ren, K., Jiang, X., Shen, Y., Zhang, H., and Li, D. (2023). SIMPLE: Specialized Model-Sample Matching for Domain Generalization. In *The Eleventh International Conference on Learning Representations*.

Li Deng (2012). The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

Li Fei-Fei, Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7):100779.

Ligozat, A.-L., Lefevre, J., Bugeau, A., and Combaz, J. (2022). Unraveling the Hidden Environmental Impacts of AI Solutions for Environment Life Cycle Assessment of AI Solutions. *Sustainability*, 14(9):5172.

Ligozat, A.-L. and Luccioni, S. (2021). A Practical Guide to Quantifying Carbon Emissions for Machine Learning researchers and practitioners. Technical report, MILA ; LISN.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing, Cham.

Liu, F., Bugliarello, E., Ponti, E. M., Reddy, S., Collier, N., and Elliott, D. (2021a). Visually Grounded Reasoning across Languages and Cultures. In *EMNLP 2021 : Conference on Empirical Methods in Natural Language Processing*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada. IEEE.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Miceli, M. and Posada, J. (2021). Wisdom for the Crowd: Discoursive Power in Annotation Instructions for Computer Vision. In *Workshop: Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision, CVPR 2021*, page 4.

Mitchell, T. M. (2013). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York, nachdr. edition.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507.

Mohamed, S., Png, M.-T., and Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4):659–684.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, 26(4):2141–2168.

Nachmany, Y. and Alemohammad, H. (2018). Generating a Training Dataset for Land Cover Classification to Advance Global Development. In *Workshop on Machine Learning for the Developing World, NeurIPS 2018*.

Nguyen, K. A., Seeboonruang, U., and Chen, W. (2023). Projected Climate Change Effects on Global Vegetation Growth: A Machine Learning Approach. *Environments*, 10(12):204.

Nishant, R., Kennedy, M., and Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management*, 53:102104.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York.

Norström, A. V., Cvitanovic, C., Löf, M. F., West, S., Wyborn, C., Balvanera, P., Bednarek, A. T., Bennett, E. M., Biggs, R., de Bremond, A., Campbell, B. M., Canadell, J. G., Carpenter, S. R., Folke, C., Fulton, E. A., Gaffney, O., Gelcich, S., Jouffray, J.-B., Leach, M., Le Tissier, M., Martín-López, B., Louder, E., Loutre, M.-F., Meadow, A. M., Nagendra, H., Payne, D., Peterson, G. D., Reyers, B., Scholes, R., Speranza, C. I., Spierenburg, M., Stafford-Smith, M., Tengö, M., van der Hel, S., van Putten, I., and Österblom, H. (2020). Principles for knowledge co-production in sustainability research. *Nature Sustainability*, 3(3):182–190.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1356.

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, Toronto Ontario Canada. ACM.

Okolo, C. T. (2020). AI in the "Real World": Examining the Impact of AI Deployment in Low-Resource Contexts. In *Navigating the Broader Impacts of AI Research Workshop, NeurIPS 2020*.

Omeiza, D. (2019). A Step Towards Exposing Bias in Trained Deep Convolutional Neural Network Models. In *Workshop on Machine Learning for the Developing World*.

Oneto, L. and Chiappa, S. (2020). Fairness in Machine Learning. In Oneto, L., Navarin, N., Sperduti, A., and Anguita, D., editors, *Recent Trends in Learning From Data*, volume 896, pages 155–196. Springer International Publishing, Cham.

Oyewola, D. O., Dada, E. G., Misra, S., and Damaševičius, R. (2022). A Novel Data Augmentation Convolutional Neural Network for Detecting Malaria Parasite in Blood Smear Images. *Applied Artificial Intelligence*, pages 1–22.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Papadopoulos, D. P., Uijlings, J. R. R., Keller, F., and Ferrari, V. (2017). Extreme Clicking for Efficient Object Annotation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4940–4949, Venice. IEEE.

Parcollet, T. and Ravanelli, M. (2021). The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. In *Interspeech 2021*, pages 4583–4587. ISCA.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon Emissions and Large Neural Network Training.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. (2019). Moment Matching for Multi-Source Domain Adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1406–1415, Seoul, Korea (South). IEEE.

Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D., and Ermon, S. (2017). Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning. In *Workshop on Machine Learning for the Developing World, NeurIPS 2017*.

Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., and Weikum, G. (2018). On Measuring Bias in Online Information. *ACM SIGMOD Record*, 46(4):16–21.

Prabhu, V., Selvaraju, R. R., Hoffman, J., and Naik, N. (2022). Can domain adaptation make object recognition work for everyone? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3980–3987, New Orleans, LA, USA. IEEE.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38 Th International Conference on Machine Learning*.

Reddyhoff, D. (2022). Dependency, Data and Decolonisation: A Framework for Decolonial Thinking in Collaborative AI Research. In *Resistance AI Workshop, NeurIPS 2022*, page 5.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., Creutzig, F., Chayes, J., and Bengio, Y. (2023). Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2):1–96.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sagawa, S., Koh, P. W., Lee, T., Gao, I., Xie, S. M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., Beery, S., David, E., Stavness, I., Guo, W., Leskovec, J., Saenko, K., Hashimoto, T., Levine, S., Finn, C., and Liang, P. (2022). Extending the WILDS Benchmark for Unsupervised Adaptation. In *International Conference on Learning Representations*. OpenReview.

Saito, K., Yamamoto, S., Ushiku, Y., and Harada, T. (2018). Open Set Domain Adaptation by Backpropagation. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, volume 11209, pages 156–171. Springer International Publishing, Cham.

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., and Sheaves, M. (2020). A Realistic Fish-Habitat Dataset to Evaluate Algorithms for Underwater Visual Analysis. *Scientific Reports*, 10(1):14671.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., and Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 315–328, Virtual Event Canada. ACM.

Sambasivan, N., Arnesen, E., Hutchinson, B., and Prabhakaran, V. (2020). Non-portability of Algorithmic Fairness in India. In *Navigating the Broader Impacts of AI Research Workshop, NeurIPS 2020*.

Santurkar, S., Tsipras, D., and Madry, A. (2020). BREEDS: Benchmarks for Subpopulation Shift. In *International Conference on Learning Representations*.

Sequeira, L. N., Moreschi, B., and Santos, V. A. A. D. (2021). WHICH HUMAN FACES CAN AN AI GENERATE? LACK OF DIVERSITY IN THIS PERSON DOES NOT EXIST. *AoIR Selected Papers of Internet Research*.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *Workshop on Machine Learning for the Developing World, NeurIPS 2017*.

Shen, H., Deng, W. H., Chattopadhyay, A., Wu, Z. S., Wang, X., and Zhu, H. (2021). Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–861, Virtual Event Canada. ACM.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations 2015*.

Stark, L. (2019). GPT-3 output as human rights violation collaboration, and resistance via GeDi class-conditional generative discrimination. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):50–55.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Suresh, H. and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA. ACM.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML 2019*.

Tessier, C. (2021). Éthique et IA: analyse et discussion. In *PFIA 2021 : Plateforme Française en Intelligence Artificielle 2021*.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Tolkova, I., Chu, B., Hedman, M., Kahl, S., and Klinck, H. (2021). Parsing Birdsong with Deep Audio Embeddings. In *Artificial Intelligence for Social Good (AI4SG) Workshop*.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, Colorado Springs, CO, USA. IEEE.

Trivedi, A., Mukherjee, S., Tse, E., Ewing, A., and Ferres, J. L. (2019). Risks of Using Non-verified Open Data: A case study on using Machine Learning techniques for predicting Pregnancy Outcomes in India. In *Workshop on Machine Learning for the Developing World*.

Tsiskaridze, R., Reinhold, K., and Jarvis, M. (2023). Innovating HRM Recruitment: A Comprehensive Review Of AI Deployment. *Marketing and Management of Innovations*, 14(4):239–254.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., Van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., Van Horn, G., Crofoot, M. C., Stewart, C. V., and Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):792.

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden. ACM.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1):233.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O. (2022). REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision*, 130(7):1790–1810.

Wang, H., Ge, S., Xing, E. P., and Lipton, Z. C. (2019). Learning Robust Global Representations by Penalizing Local Predictive Power. In *NeurIPS 2019*.

Wang, J., Nadarajah, S., Wang, J., and Ravikumar, A. (2020). A Machine Learning Approach to Methane Emissions Mitigation in the Oil and Gas Industry. Preprint, Environmental Engineering.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Washington, A. L. and Kuo, R. (2020). Whose side are ethics codes on?: Power, responsibility and the social good. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 230–240, Barcelona Spain. ACM.

Whittlestone, J. and Ovadya, A. (2019). The tension between openness and prudence in responsible AI research. In *Joint Workshop on AI for Social Good, NeurIPS 2019*.

Wilson, B., Hoffman, J., and Morgenstern, J. (2019). Predictive Inequity in Object Detection. In *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision, CVPR 2019*.

World Bank Group (2017). *The Growing Role of Minerals and Metals for a Low Carbon Future*. World Bank, Washington, DC.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges and opportunities. In Marculescu, D., Chi, Y., and Wu, C., editors, *Proceedings of Machine Learning and Systems, PMLS 2022*, volume 4, pages 795–813.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, Barcelona Spain. ACM.

Ying, X., Li, X., and Chuah, M. C. (2021). Weakly-supervised Object Representation Learning for Few-shot Semantic Segmentation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1496–1505, Waikoloa, HI, USA. IEEE.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Zhu, K. and Wu, J. (2021). Residual Attention: A Simple but Effective Method for Multi-Label Recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 184–193.

Zhuang, J., Chen, Z., Wei, P., Li, G., and Lin, L. (2022). Open Set Domain Adaptation By Novel Class Discovery.