UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
Mathematics and Computer Science

Cycle XXXVI

TITLE OF THE Ph.D. THESIS

Unfairness Assessment, Explanation and Mitigation

in Machine Learning Models for Personalization

Scientific Disciplinary Sector(s)

INF/01

Ph.D. Student:	Giacomo Medda
Supervisor:	Prof. Gianni Fenu
Co-Supervisor:	Dott. Mirko Marras, Dott. Ludovico Boratto

Final exam. Academic Year 2022/2023
Thesis defence: February 2024 Session

Statement of Authorship

I declare that this thesis entitled "Unfairness Assessment, Explanation and Mitigation in Machine Learning Models for Personalization" and the work presented in it are my own. I confirm that:

- this work was done while in candidature for this PhD degree;
- when I consulted the work published by others, this is always clearly attributed;
- when I quoted the work of others, the source is always given;
- I have acknowledged all main sources of help;
- with the exception of the above references, this thesis is entirely my own work;
- appropriate ethics guidelines were followed to conduct this research;
- for work done jointly with others, my contribution is clearly specified.

Abstract

The last decade has been pervaded by the automatic applications leveraging *Artificial Intelligence* technologies. Novel systems have been adopted to automatically solve relevant tasks, from scanning passengers during border controls to suggesting the groceries to buy to fill the fridge. One of the most captivating applications of *Artificial Intelligence* is represented by *voice assistants*, like Alexa. They enable people to use their voice to perform simple tasks, such as setting an alarm or saving an appointment in an online calendar. Due to their worldwide usage, voice assistants are required to aid a diverse range of individuals encompassing various cultures, languages, accents, and preferences. It is then crucial for these systems to function fairly across different groups of people to ensure reliability and provide assistance without being influenced by sensitive attributes that may vary among them.

This thesis deals with the design, implementation, and evaluation of *Artificial Intelligence* models that are optimized to operate fairly in the context of voice assistant systems. Assessing the level of performance of existing fairness-aware solutions is an essential step towards comprehending how much effort should be put to provide fair and reliable technologies. The contributions result in extensive analyses of existing methods to counteract unfairness, and in novel techniques to mitigate and explain unfairness that capitalize on *Data Balancing*, *Counterfactuality*, and *Graph Neural Networks Explainability*. The proposed solutions aim to support system designers and decision makers over several fairness requirements. Specifically, over methodologies to evaluate fairness of models outcomes, techniques aimed to improve users' trustworthiness by mitigating unfairness, and strategies that generate explanations of the potential causes behind the estimated unfairness. Through our studies, we explore opportunities and challenges introduced by the latest advancements in *Fair Artificial Intelligence*, a relevant and timely topic in literature.

Supported by extensive experiments, our findings illustrate the feasibility of designing *Artificial Intelligence* solutions for the mitigation and explanation of unfairness issues in the models adopted in voice assistants. Our results provide guidelines on fairness evaluation, and design of methods to counteract unfairness concerning the voice assistant scenario. Researchers can use our findings to follow a schematic protocol for fairness assessment, to discover the data aspects affecting the model fairness, and to mitigate the outcomes unfairness, among others. We expect that this thesis can support the adoption of fairness-aware solutions in the voice assistant pipeline, from the voice authentication to the requested task resolution.

Dissemination

The research that contributed to the content part of this Ph.D. thesis has resulted from *8* papers fully published in national and international journals, conference and workshop proceedings. I would sincerely thank my co-authors for their precious contribution, and such a gratitude would be demonstrated with the adoption of the scientific 'We' throughout the thesis.

I firstly make it clear my contribution. I envisioned the research presented in this thesis and completed the majority of the work. I designed the approaches and chose the research directions. I collected the datasets and was responsible for data analysis. Moreover, I was in charge of the implementation of the related scripts. Finally, I wrote the papers for submission, managed the peer-review pipeline, and subsequently revised them. I collaborated closely with the listed co-authors throughout all stages. Co-authors provided feedback on the approaches, offered technical support, discussed techniques, and contributed to the preparation of the submitted work. Furthermore, I was in charge of the presentation of *6* papers at conferences and workshops.

Exception is made for papers numbered as (ii), (v), and (vi) as I and Dr. Giacomo Meloni equally contributed. Similarly, I and Dr. Vivek Kumar put comparable effort for the paper numbered as (v).
The detailed references to the produced papers are provided below. In most of the references I am not the first author, due to an Italian formality of listing the authors' names in alphabetical order.

Peer-reviewed Publications in Journals

i. Boratto, L., Fenu, G., Marras, M., & **Medda, G.** (2023). *Practical perspectives of consumer fairness in recommendation.* In: Information Processing & Management, 60(2), 103208, Elsevier, Q1. Talk (Conference associated with the journal publication). https://doi.org/10.1016/j.ipm.2022.103208

ii. Fenu, G., Marras, M., **Medda, G.**, & Meloni, G. (2023). *Causal reasoning for algorithmic fairness in voice controlled cyber-physical systems.* In: Pattern Recognition Letters, 168, 131-137, Elsevier, Q1. https://doi.org/10.1016/j.patrec.2023.03.014

Peer-reviewed Publications in International Conference Proceedings

iii. Boratto, L., Fabbri, F., Fenu, G., Marras, M., & **Medda, G.** (2023). *Counterfactual Graph Augmentation for Consumer Unfairness Mitigation in Recommender Systems*. In: Proceedings of the 32nd ACM International Conference on Information & Knowledge Management (CIKM 2023), NA, ACM, A. Poster. https://doi.org/10.1145/3583780.3615165

iv. Boratto, L., Fenu, G., Marras, M., & **Medda, G.** (2023). *Consumer Fairness in Recommender Systems: Contextualizing Definitions and Mitigations*. In: Proceedings of the Advances in Information Retrieval - 44th European Conference on IR Research (ECIR 2022), 552-566, Springer, A. Talk. https://doi.org/10.1007/978-3-030-99736-6_37

v. Fenu, G., Marras, M., **Medda, G.**, & Meloni, G. (2021). *Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition*. In: Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), 1892-1896, ISCA, A. Talk. https://doi.org/10.21437/Interspeech.2021-1857

vi. Fenu, G., Marras, M., **Medda, G.**, & Meloni, G. (2020). *Improving Fairness in Speaker Recognition*. In: Proceedings of the European Symposium on Software Engineering (ESSE 2020), 129-136, ACM, Unranked. Best Presentation Award. Talk. https://doi.org/10.1145/3393822.3432325

Peer-reviewed Publications in International Workshop Proceedings

vii. Kumar, V., **Medda, G.**, Reforgiato, D., Riboni, D., Helaoui, R, & Fenu, G. (2023). *How Do You Feel? Information Retrieval in Psychotherapy and Fair Ranking Assessment*. In: Proceedings of the Advances in Bias and Fairness in Information Retrieval - 4th International Workshop (BIAS 2023), 119-133, Springer, Unranked. Talk. https://doi.org/10.1007/978-3-031-37249-0_10

Peer-reviewed Publications in National Workshop Proceedings

viii. Boratto, L., Fenu, G., Marras, M., & **Medda, G.** (2023). *Consumer Fairness Benchmark in Recommendation*. In: Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), 60-65, CEUR-WS, Unranked. Talk. https://ceur-ws.org/Vol-3448/paper-27.pdf

Acknowledgements

I want to spend a few words to thank the people that supported me during my PhD journey, inside and outside of the research environment.

First of all, I thank with immense gratitude my supervisors, Prof. *Gianni Fenu*, Dr. *Mirko Marras*, and Dr. *Ludovico Boratto*. Not everyone had the pleasure to be advised by more than one supervisor, especially three brilliant minds, who supported me throughout my research career. Each one contributed in enriching me as a person and as a researcher at different levels, and I could not thank you enough for this experience. Heartfelt thanks to Dr. *Francesco Fabbri*, who have supported and guided me since my abroad internship at EURECAT.

I also thank the research group that hosted me, and all the people I have met and I have shared these adventures with. A particular acknowledgement is required for my colleagues, with whom I share funny memories inside and outside of the research environment. A warm thanks also to my friends for all the moments of lightheartedness.

Finally, I am extremely thankful to the people who significantly supported me outside of the work. To my family, who have supported me throughout this journey and have cared to provide me with anything I could have needed. To my chosen brother, *Giacomo*, your help was invaluable, as a colleague since M.Sc. study and as a true friend. To my girlfriend, *Pamela*, for your continuous support in brightening my low moments and for your determination in encouraging me to pursue ambitious objectives. Thank you for making me feel that my research works are real successes.

Contents

Statement of Authorship **I**

Abstract **II**

Dissemination **IV**

Acknowledgements **VII**

List of Figures **XIII**

List of Tables **XV**

1 Introduction **1**
1.1 Motivation and Open Issues 1
1.2 Contributions 2
1.3 Dissertation Structure 3

2 Background **5**
2.1 Deep Learning 5
2.1.1 Feed Forward Neural Networks 6
2.1.2 Convolutional Neural Networks 6
2.1.3 Graph Neural Networks 6
2.1.4 Residual Neural Networks 7
2.1.5 Recurrent Neural Networks 7
2.2 Algorithmic Fairness 8
2.2.1 Historical Background 8
2.2.2 Artificial Intelligence Regulations 8
2.2.3 Fairness Notions 9
2.3 Algorithmic Methods 11
2.3.1 Methods for Speaker Verification 11
2.3.2 Methods for Recommendation 13
2.3.3 Methods for Algorithmic Fairness Analysis 15

3 Fairness in Speaker Verification **19**
3.1 Introduction . . . 19
3.2 Problem Formulation . . . 21
3.2.1 Speaker Recognition Task . . . 21
3.2.2 Task Optimization Targets . . . 22
3.3 Techniques for Unfairness Assessment and Mitigation . . . 25
3.3.1 Methodology . . . 26
3.3.2 Experimental Setup . . . 28
3.3.3 Results . . . 30
3.4 Counterfactual Reasoning for Unfairness Explanation . . . 35
3.4.1 Methodology . . . 36
3.4.2 Experimental Setup . . . 40
3.4.3 Results . . . 41
3.5 Findings and Discussion . . . 45

4 Fairness in Recommendation **47**
4.1 Introduction . . . 47
4.2 Problem Formulation . . . 50
4.2.1 Model-based Recommendation Task . . . 50
4.2.2 Graph-based Recommendation Task . . . 51
4.2.3 Task Optimization Targets . . . 52
4.3 Techniques for Unfairness Assessment and Mitigation . . . 57
4.3.1 Methodology . . . 58
4.3.2 Results . . . 65
4.4 Unfairness Explanation via Graph Perturbation . . . 80
4.4.1 Methodology . . . 80
4.4.2 Experimental Setup . . . 85
4.4.3 Results . . . 90
4.5 Unfairness Mitigation via Graph Augmentation . . . 96
4.5.1 Methodology . . . 97
4.5.2 Experimental Setup . . . 98
4.5.3 Results . . . 99
4.6 Findings and Discussion . . . 101

5 Conclusions **105**
5.1 Contributions Summary . . . 105
5.2 Limitations and Open Issues . . . 106
5.3 Future Works . . . 106

A Fairness in Therapeutic Counseling **109**
A.1 Introduction . . . 109
A.2 Literature Review . . . 110
A.3 Methodology . . . 110

A.4 Experimental Settings 111
A.5 Results 113
A.6 Conclusions and Future Works 117

B Generalized Explainer of Global Issues **119**
B.1 Proposed Method 119
B.2 Experimental Evaluation 121

Bibliography **123**

List of Figures

1.1 Fairness in machine learning models for personalization. 2

2.1 AI Risks in the AI Act. 9
2.2 Categorization of recommender systems. 14

3.1 *Enrollment* and *Verification* process in speaker recognition. 21
3.2 Fairness-aware framework pipeline in speaker recognition. 26
3.3 Fairness estimates of speaker recognition models under NB and UB. . 32
3.4 X-Vector: impact of decision threshold on the trade-off between fairness, security, and usability. 33
3.5 ResNets: impact of decision threshold on the trade-off between fairness, security, and usability. 34
3.6 Exploratory analysis pipeline for voice characteristics impact. 36
3.7 Pearson correlation between explanatory variables. 42
3.8 Permutation feature importance of voice characteristics. 43
3.9 Counterfactual analysis of protected class flipping. 44

4.1 Overview of the research on fairness in recommendation. 49
4.2 Paper collection and reproducibility study pipeline. 58
4.3 Consistency property evaluation under CES. 73
4.4 Data Robustness property evaluation. 74
4.5 Trade-off property evaluation. 77
4.6 `GNNUERS` perturbation process. 81
4.7 `GNNUERS` fairness evaluation across users' subgroups for age. 90
4.8 `GNNUERS` fairness evaluation across users' subgroups for gender. 91
4.9 `GNNUERS` deleted edges distribution over age quartiles on NGCF. . . . 94
4.10 `GNNUERS` deleted edges distribution over gender quartiles on NGCF. . 94
4.11 Relative difference in NDCG/ΔNDCG before and after augmentation. 101

A.1 NDCG distribution for rankers in therapeutical counseling. 114
A.2 Pairwise difference in NDCG for rankers in therapeutical counseling. . 115
A.3 NDCG for any queried topic (column) and any ranker (row) in therapeutical counseling. 116

B.1 ΔNDCG over epochs of `GENIUS-RS`$^{+}$ and `GENIUS-RS`$^{-}$. 120

List of Tables

3.1 EER performance of speaker recognition models. 30
3.2 Disparity Score for age, gender under NB and UB training sets. . . . 31
3.3 Same age or same gender performance of speaker recognition models. 40

4.1 Reproducible mitigation procedures for consumer fairness. 60
4.2 Reproducibility study datasets with consumers' sensitive attributes. . 61
4.3 Consistency property evaluation in Top-k recommendation for gender. 66
4.4 Consistency property evaluation in Top-k recommendation for age. . 67
4.5 Consistency property evaluation in Rating prediction for gender. . . . 67
4.6 Consistency property evaluation in Rating prediction for age. 68
4.7 Consistency property evaluation under F1 Score. 72
4.8 Consistency property evaluation under ϵ-fairness. 72
4.9 Scalability property evaluation. 76
4.10 Transferability property evaluation on Ekstrand et al.'s method. . . . 79
4.11 Transferability property evaluation on Li et al.'s method. 79
4.12 Datasets for evaluation of `GNNUERS`. 87
4.13 NDCG relative change after applying `GNNUERS`. 93
4.14 Mitigation performance of the sampling policies for graph augmentation. 99
4.15 NDCG and ΔNDCG after applying the graph augmentation. 102
4.16 Summary categorization of reproduced mitigation procedures. 103

A.1 Distribution of topics over each *MI quality* class. 113

Chapter 1

Introduction

1.1 Motivation and Open Issues

Artificial intelligence (AI) systems have been pervading an immense amount of activities, jobs, and services to ease the human involvement. Automatic systems powered by AI are currently employed in several scenarios, for instance to guarantee security (e.g., border controls scanning, identity verification for bank transactions) [138], to process or generate text data (e.g., language translation, question answering) [95], to support mental health issues (e.g., chatbots) [3], or to improve user satisfaction in e-commerce and streaming platforms (e.g., recommender systems) [128].

The increment of responsibility provided to automatic decision-making systems raise issues about side consequences related to the usage of AI. Ongoing efforts by researchers have been focused on improving AI algorithms in terms of trustworthiness [147], ethics [35], fairness [107], and explainability [170]. Such aspects have been analyzed in some of the most popular tools leveraging AI, among which *voice assistants* like Alexa [72, 131] and Google Home. A concerted effort has been particularly made to analyze and mitigate *unfairness* in voice-based assistants [131] and *speaker recognition systems* in general [118]. However, current studies only questioned whether the recognition task performed by such tools was fair.

Voice assistants provide several services, from planning the shopping cart to suggesting activities or movies to watch. Indeed, some of the voice assistants operations involve additional AI systems, e.g., *recommender systems* [128], which are devoted to suggest movies, songs, and so on, personalized on the basis of a user's personal preferences. It follows that accounting for the fair decision-making of voice assistants in their speaker recognition task is not enough, but also the personalization for the consumers should be generated by a fair process [150].

The literature in recommendation has been active in assessing [47], mitigating [92], and explaining [59] unfairness issues in the outcomes of the employed models. However, there are still challenges related to the reliability and consistency of the methods devised to improve and better understand fairness in recommender systems. Additionally, the community has not yet explored the fairness require-

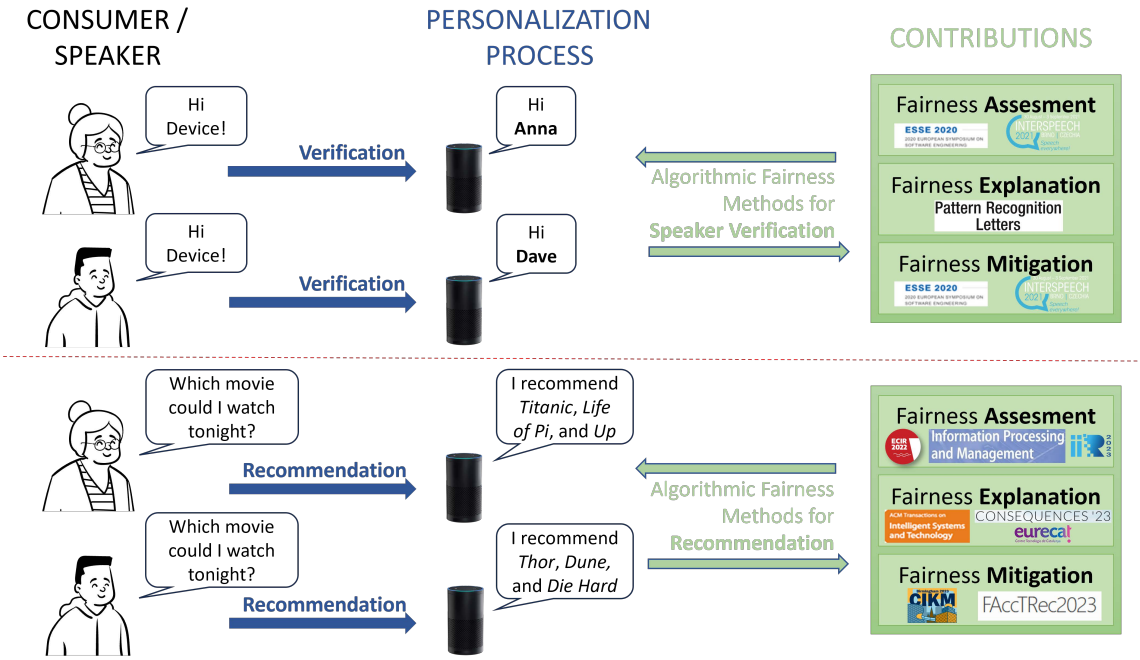


Figure 1.1: Personalization context empowered with this thesis contributions in the speaker verification and recommendation processes.

ments to be adopted in an end-to-end voice assistant task of recognizing speakers and suggesting personalized content based on their request.

1.2 Contributions

In this thesis, the fairness concerns involved in a voice assistant for the multiple tasks of speaker recognition and personalized recommendation are studied by individually analyzing each task. The Figure 1.1 depicts the pipeline of the considered multiple tasks and the diverse fairness requirements planned in our study. The focus is on contextualizing the existing works that counteract unfairness issues in speaker recognition and recommendation, devising novel methods to improve fairness, and also exploring strategies to better comprehend the causes behind AI unfair decisions. The proposed contributions consist of a comprehensive analysis of the performance of existing unfairness mitigation methods, an extended protocol to evaluate fairness-aware algorithms, and a framework to both mitigate unfairness and identify the data subset that causes a system to generate unfair outcomes. This series of works aims to support the process of fairly verifying the identity of a speaker and fairly provide personalized recommendations to improve the user satisfaction and trustworthiness towards such unified technologies.

Going more into detail, we provide (i) a framework to examine the performance in unfairness mitigation of existing approaches, which have been devised for recom-

mender systems or adopted to speaker recognition from other domains, (ii) a set of practical perspectives that forms a unified evaluation protocol to verify an algorithm fairness from different viewpoints and scenarios, (iii) two frameworks powered by explainability strategies, one identifies the peculiar voice features negatively affecting the most a speaker recognition system, the other extracts a subset of user-item interactions that can be used to both explain and mitigate unfairness in recommender systems.

Overall, this thesis provides guidelines, insights, limitations, and future directions for AI algorithms adopted to counteract unfairness in speaker recognition and recommender systems. The content of this manuscript could support researchers into understanding the various fairness notions used in the literature, how they are operationalized to study the issue from different viewpoints, and how fairness should be evaluated on AI systems outcomes.

1.3 Dissertation Structure

The remainder of this thesis is organized as follows: *Chapter 2* provides a brief introduction to deep AI architectures, fairness notions and regulations, and the literature methods devised to solve the tasks underlying this thesis overall contributions.

Chapter 3 illustrates our frameworks to assess, mitigate and explain unfairness in high-end speaker recognition models. The main framework leverages multiple fairness notions to analyze the outcomes bias from different viewpoints and a data balancing technique as a mitigation tool. An extension of the framework leverages surrogate models to estimate the influence of speech covariates on the accuracy of speaker recognition models. Such influence highlights specific voice characteristics that can be used as an explanation of aspects discovered in the outcomes, e.g., unfairness in recognition accuracy. This work has been partially studied jointly with dr. *Giacomo Meloni*, an independent researcher, and published on the "*European Symposium on Software Engineering*" (ESSE) [55] and the "*The International Speech Communication Association*" (INTERSPEECH) conference proceedings [53]. The framework component aimed to estimate the influence of speech covariates has been described in a paper published on the "*Pattern Recognition Letters*" journal [54].

Chapter 4 depicts a research process aimed to first contextualize existing algorithms for unfairness mitigation in recommendation under a common evaluation protocol, and then identify a comprehensive list of practical perspectives that should be met to consider an unfairness mitigation technique as reliable in practice. Such research process extends towards a novel method that is not only able to mitigate unfairness in recommendation, but it is also driven by explainability techniques to being more interpretable and to highlight the possible causes of unfairness issues in recommender systems. The first work on contextualization of unfairness mitigation algorithms and identification of practical perspectives has been partially studied jointly with prof. *Ludovico Boratto* from the *University of Cagliari* (Italy), and

published on the "*European Conference on Information Retrieval*" (ECIR) conference proceedings [18] and the "*Information Processing & Management*" (IPM) journal [19]. The novel method to mitigate and explain unfairness in recommendation has been partially studied jointly with prof. *Ludovico Boratto* from the *University of Cagliari* (Italy) and dr. *Francesco Fabbri* from *Spotify*, and published on the "*International Conference on Information Knowledge & Management*" (CIKM) conference proceedings [16] and submitted to the "*Transactions on Intelligent Systems and Technology*" journal.

Chapter 5 discusses the implications of our research with particular focus on its limitations, but also on open issues related to our work and potential advancements.

Additional content is provided on two appendices, related to two works regarding aspects not directly related to the main topic of fairness advancements in recommendation and speaker recognition. *Appendix A* presents a work aimed to examine the adoption of information retrieval models to support the autonomous browsing of information related to mental health issues. Such a framework would enable patients to get more familiar with their own diseases for a therapeutical purpose, and an unfairness evaluation is also performed due to the sensitive information associated with personal health issues. This topic has been partially studied with dr. *Vivek Kumar* from the *University of Cagliari* (Italy), and published on the "*International Workshop on Algorithmic Bias in Search and Recommendation*" (BIAS) workshop proceedings [86]. *Appendix B* regards an extension of the explanation framework devised for unfairness in recommendation, such that it is generalized to explain several aspects beyond fairness in recommendation, e.g., coverage, diversity, model instability, or novelty. This topic has been partially studied with prof. *Ludovico Boratto* from the *University of Cagliari* (Italy) and dr. *Francesco Fabbri* from *Spotify*, and submitted for publication to the last year edition of the "*European Conference on Information Retrieval*" (ECIR).

Chapter 2

Background

This chapter provides essential context around deep learning, algorithmic fairness background, and methods for speaker verification, recommendation, and algorithmic fairness studied and used in this thesis.

2.1 Deep Learning

AI is a thriving field with many practical applications that aims to create machines capable of simulating human-like intelligence and decision-making processes. It encompasses a wide range of techniques, algorithms, and methodologies to solve complex problems and learn from data.

Deep Learning (DL) is a subfield of AI that focuses on training artificial neural networks with multiple layers to perform tasks such as image recognition, natural language processing, and speech recognition. It has revolutionized many AI applications due to its ability to automatically learn hierarchical representations from data. Indeed, we can identify two other sub-levels between DL and AI, i.e. *Machine Learning* and *Representation Learning*, where the former is broader than the latter. Representation learning is a fundamental concept within deep learning, where the neural networks learn to automatically extract meaningful features and representations from raw data. DL extends representation learning by introducing representations that are expressed in terms of other, simpler representations [64]. Machine Learning (ML) is a broader field that encompasses deep learning, representation learning, but also other algorithms and techniques that extract patterns from raw data. ML algorithms, as well as the ones in the mentioned sub-fields, can be classified as supervised, unsupervised, or reinforcement learning, depending on the type of training data and the learning approach.

At its core, AI represents the broader goal of creating intelligent machines, and DL specializes in training deep neural networks for complex tasks, where the data is processed along several mathematical functions, i.e. *layers*, that learn a new representation. Each layer can be thought of as the state of the computer's memory after executing another set of instructions in parallel [64]. Networks with greater

depth can execute more instructions in sequence. Here we revise some of the most common, but also recent neural networks used to solve tasks contemplated in this thesis contributions.

2.1.1 Feed Forward Neural Networks

Deep feedforward networks, also known as *feedforward neural networks* (FFNNs) or *multilayer perceptrons* (MLPs), are fundamental models in deep learning. The main objective of these networks is to approximate a specific function f^* [64], which describes a phenomenon that can be learnt by specific tasks, e.g., classification. The network is composed of layers of functions, forming a directed acyclic graph, and there are no feedback connections in the model, i.e. the information flows through a uni-lateral direction. Feedforward networks play a crucial role in machine learning applications and in many commercial systems. During training, the network learns to approximate the desired output $f^*(x)$ by processing training data examples through a chain of layers until the final layer, denoted as *output layer*, is reached. The term "deep learning" comes from the network's depth, which refers to the number of layers in the model. The inner layers between the first and the last one are called *hidden layers*, because their output is not directly related to the training data, but their usage is established by the network itself. These networks are denoted as "neural" because they are inspired by neuroscience, with each hidden layer representing a vector-valued function akin to neurons.

2.1.2 Convolutional Neural Networks

Convolutional Neural Networs (CNNs) are a family of FFNNs suited for data with a grid-like topology, such as time-series and image data. The name derives from the mathematical operation performed by such networks, i.e. *convolution*. Convolutions are applied on the grid-like data by means of small filters, called *kernels*, that slide across the grid and extract local patterns, such as edges and low-level features [64]. CNNs typically include pooling layers that process the convoluted data through a down-sampling operation to represent the data with low-dimensional feature vectors. After multiple convolutional and pooling layers, a sequence of *fully connected* layers, also denoted as *dense* layers, takes in input the output features for several downstream tasks. CNNs have especially become a fundamental tool in various computer vision applications thanks to their ability to learn relevant features from spatial information, such as images and videos.

2.1.3 Graph Neural Networks

Graph Neural Networks (GNNs) are a family of networks suited for graphs, characterized by a non-Euclidean data structure. A graph is a data structure representing a set of entities, denoted as *nodes*, and their relationships, denoted as *edges*. GNNs

present some similarities with CNNs given that the Euclidean grid-like data structure used by CNNs can be regarded as instances of graphs [175]. The main idea behind GNNs is to propagate information across the graph to learn representations for each node that capture both the node's own features and the information from its neighboring nodes. This propagation of information is typically done through message passing mechanisms, where each node aggregates information from its neighbors and updates its representation accordingly. We can distinguish among several types of GNNs based on the underlying implementation and aggregation mechanism: graph convolutional networks (GCNs), graph attention networks (GATs), graph sample and aggregate (GraphSAGE), gated graph neural networks, heterogenous information networks (HINs), and so on.

2.1.4 Residual Neural Networks

Residual Neural Networks, commonly known as *ResNets*, are a type of deep CNN architecture introduced by [70]. ResNets are specifically designed to address the vanishing gradient problem that occurs in very deep neural networks. The vanishing gradient problem arises when gradients become very small as they propagate backward through many layers during the training process. ResNets tackle this issue by using a novel *shortcut connection* or *skip-connection* that allows the network to learn residual mappings instead of directly learning the desired output. In simpler terms, instead of learning the mapping from the input to the output, ResNets learn the difference between the desired output and the input (the residual). The network then learns to predict this residual and adds it back to the original input, effectively making it easier for the network to learn the identity mapping.

2.1.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a family of networks suited to deal with sequential data. RNNs leverage the concept of *parameter sharing* to learn generalizable patterns across diverse sequences [64]. Without parameter sharing, RNNs could not generalize due to separate parameters targeting each value of the time index. Each member of the output is a function of the previous members of the output, utilizing the same update rule applied to the previous outputs. Hence, computational graphs of RNNs encompass feedback and cycles, reflecting the influence of the present value of a variable on its own value at a future time step. Extensions of RNNs, such as *Long Short-Term Memory* (LSTM) and *Gated Recurrent Unit* (GRU), introduce *self-loops* to accumulate information over a long duration, with focus on the sequence context. Additionally, several gates, controlled by other hidden units, are introduced to control the way the internal state is managed, by also leveraging a *forgetting* factor to discard information of previous states.

2.2 Algorithmic Fairness

Algorithmic fairness is an essential property that must be considered and analyzed in modern machine learning systems. Indeed, the algorithms that drive such systems are vulnerable to biases exactly as humans, resulting in unfair decisions towards an individual or demographic group. As [107] states, "*fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics*". In this section, we describe real-world example of algorithmic unfairness, governmental regulations proposed in USA and EU, and several fairness notions that define what is and what is not fair.

2.2.1 Historical Background

Due to the prolific spread of AI and machine learning in different real-world applications, safety and fairness constraints have become a significant issue. Inherent biases exist in modern AI applications, affecting our daily lives when interacting with chatbots, employment matching, flight routing, automated legal aid for immigration systems, and advertising placement algorithms [107]. A canonical example is represented by COMPAS, a software used by courts in the USA to decide whether to release or keep in prison an offender. An investigation uncovered COMPAS being unfair towards African-American [107], given that the software reported a higher likelihood of predicting African-American offenders to be at a higher risk of recommitting a crime compared with Caucasian offenders. Another demonstration of bias present in decision-making systems is an algorithm devised to deliver advertisements promoting jobs in STEM (Science, Technology, Engineering and Math) fields [90]. Such advertisements should have been delivered by the system in a gender-neutral way, but an investigation reported the algorithm being discriminatory. Indeed, it considered younger women to be a valuable subgroup and more expensive to show advertisements to, given that less women compared with men saw the advertisement.

2.2.2 Artificial Intelligence Regulations

Due to the aforementioned issues and events raising insecurities about the reliability of AI applications, governments have been examining and introducing novel regulations on such systems. In USA, regulating AI is in its early days [83], and no one knows how such a law about AI will look like. Moreover, USA remains far behind Europe, where lawmakers are preparing to enact an AI law [2]. In earlier stages, European laws regarding unfairness issues were already present, regulating the definition and characteristics of *sensitive* attribute and *protected* groups. Explicit mentions are given in Art. 21 of the EU Charter of Fundamental Rights, Art. 14 of European Convention on Human Rights, and Art. 18-25 of the Treaty on the Functioning of the European Union. The novel *Artificial Intelligence Act* [2, 57] will

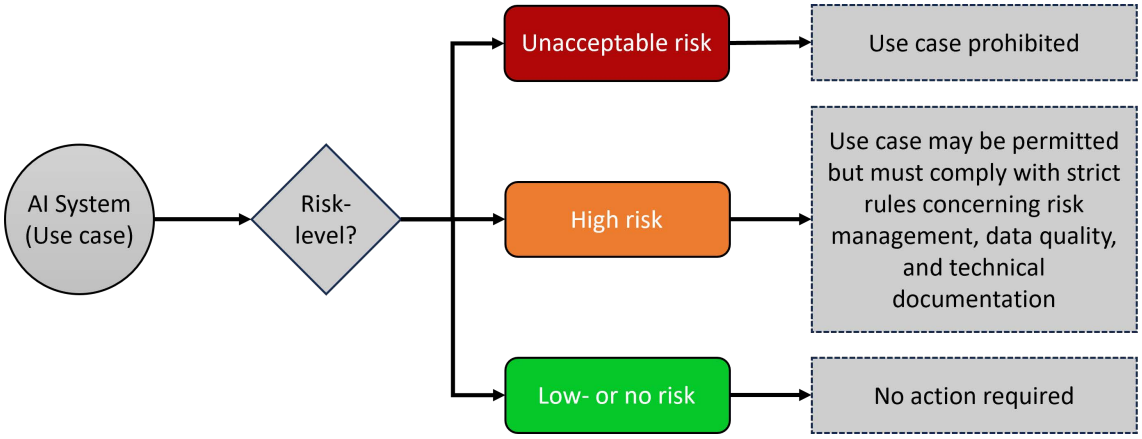


Figure 2.1: Risk categories for AI use cases under the Artificial Intelligence Act.

put new restrictions on AI applications on the basis of a categorization of the risks, depicted in Figure 2.1.

In particular:

- **prohibited practices** (unacceptable risk) include real-time biometric systems, social scoring algorithms that could lead to unfavourable treatment of individuals, and manipulative systems aimed to distort individuals' behaviors to cause physical or psychological harm.
- **high-risk AI systems** include biometric identification, employment and worker management, law enforcement, border control management, education and vocational training, and so on
- **low-risk AI systems** include systems whose operation neither depend or use personal data nor make predictions that could lead to unfavourable outcome or affect any individual directly or indirectly

2.2.3 Fairness Notions

We capitalize on [107] to describe some of the most operationalized fairness notions in the AI areas studied in this thesis. In particular, we will focus only on four fairness notions: *demographic parity*, *equal opportunity*, *equalized odds*, and *counterfactual fairness*. For further information and a complete list of all the fairness notions, we suggest the reader to consult [107].

Demographic Parity

Also denoted as statistical parity, *demographic parity* is a fairness notion that is satisfied if the likelihood of a positive outcome is the same regardless of the demographic group a person belongs to. Let $\hat{Y}$ be a predictor, under the simplest scenario

of a binary sensitive attribute that leads to the demographic groups set $A = \{0, 1\}$, *demographic parity* is formally defined as:

$$P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1). \tag{2.1}$$

In practice, this definition is typically operationalized by measuring the difference in performance experienced by an AI system across demographic groups, e.g., the difference in speaker verification accuracy across males and females.

Equal Opportunity

Confining our discussion to classification tasks, *equal opportunity* is related to the *true positive* term, which refers to a positive outcome correctly assigned to a person in the positive class. Indeed, equal opportunity is satisfied if the likelihood that a person in the positive class is assigned to a positive outcome is the same regardless of the demographic group a person belongs to. Extending the notation used for demographic parity with $Y = \{0, 1\}$, which denotes the negative (0) and the positive (1) class a person is associated with, equal opportunity can be formally defined as:

$$P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1). \tag{2.2}$$

In other words, equal opportunity is satisfied if the predictor report equal *true positive rates* across demographic groups.

Equalized Odds

The notion of *equalized odds* does not only take into account the true positives as equal opportunity, but also the *false positives*, which refer to the positive outcomes incorrectly assigned to the people in the negative class. Indeed, equalized odds is satisfied if both the probability of a person in the negative class being incorrectly assigned a positive outcome and the probability of a person in the positive class being correctly assigned a positive outcome are the same across demographic groups. Formally:

$$P(\hat{Y} = 1 \mid A = 0, Y = y) = P(\hat{Y} = 1 \mid A = 1, Y = y). \tag{2.3}$$

In other words, equalized odds is satisfied if the predictor reports equal *true positive rates* and equal *false positive rates* across demographic groups.

Counterfactual Fairness

Before delving into the next type of unfairness notion, we first define what *counterfactual reasoning* means by relying on [88, 11]. The term *counterfactual* denotes the opposite of the term *factual*. The latter is used to describe an event actually occurred in the real-world or a property actually established. An event is then denoted

as factual if it is occurred in the *actual world*. Conversely, an event is denoted as counterfactual if it is occurred in a *counterfactual world*. Counterfactuality studies address questions about the effect of hypothetical actions or *interventions*, and, once such effect is understood, we can ask what action plausibly *caused* an event [11]. In simpler words, counterfactual reasoning studies the event that would occur in the counterfactual world with respect to event occurred in the actual world when a property, condition or attribute is modified. For instance, asking what would happen if you turned left with your car instead of right at an intersection is an example of counterfactual thinking, because you are assuming a distorted world (the counterfactual one) where the event changed and you turned left, but you actually turned right (in the actual world).

[88] introduced the fairness notion of *counterfactual fairness*, which is satisfied if the likelihood of a predictor outcome is the same regardless we assign a person to a demographic group a or a', e.g., male or female. The outcome of the predictor should then not be causally dependent on a sensitive attribute, i.e. on the demographic group the person under consideration belongs to. Counterfactual fairness can be formally defined as:

$$P(\hat{Y}_{A\leftarrow a} = y \mid A = a) = P(\hat{Y}_{A\leftarrow a'} \mid A = a). \tag{2.4}$$

for all y and for any a' attainable by A. Hence, A should not be a cause of $\hat{Y}$ in any individual instance.

2.3 Algorithmic Methods

2.3.1 Methods for Speaker Verification

Speaker recognition is implemented via two main tasks: *identification* aims to detect the speaker's identity within a gallery of candidate speakers; *verification* aims to confirm the identity of the claimed speaker and operates in an open-set regime based on a gallery of enrolled speech samples. *Automatic Speaker Verification* corresponds to the verification task, so as to verifying an uncertain voice sample belongs to the speaker under consideration. Such systems are typically employed to secure the access to a private area or device, e.g., a phone or a bank account, by impeding speakers different from the owner to gain access.

Following [78], the speaker verification task can be outlined in two stages: *feature extraction* (also denoted as *front-end*) and *feature matching* (also denoted as *back-end*). The former is responsible to transform the digital signals of a vocal sample into a different representation, such as feature vectors or numerical descriptors. The latter corresponds to the verification stage, where an unclassified audio sample is converted into a feature vector, then compared with the vocal fingerprint of the genuine speaker to check they where generated by the same individual. Some modern

architectures perform both the front-end and back-end task, whose technique is denoted as *end-to-end speaker recognition*.

Speaker modeling has been recently dominated by deep neural networks [33] (DNNs) which significantly outperform classic solutions like GMM-UBM [126] or I-Vectors [38]. DNNs are typically pre-trained for the identification task, but are then adapted to open-set verification by discarding the classification head and extracting an intermediate representation, referred to as a *speaker embedding*. The embeddings of the query and enrolled samples are compared to confirm the speaker's identity.

Countless deep neural architectures have been proposed for speaker modeling. Some of the most prominent differences among the existing architectures involve the *input acoustic representation*, the *backbone network*, and the *temporal pooling strategy*. Directly using waveforms to learn a representation is possible [125], but it is much more common to use a hand-crafted 2D representation (e.g., spectrograms or filterbanks). The latter enables the adaptation of successful backbones from computer vision, e.g., VGG (Visual Geometry Group) [114] or *ResNet* (residual networks) [151, 161]. Recurrent [144], pooling [161], or *time delay neural networks* [133] can be then used to deal with the time dimension typical of the vocal input. Usually, trainable pooling layers achieve better results than simple pooling operators (e.g., average pooling [161] or statistical pooling [133]). Some of the most successful learned designs include the family of *VLAD* (Vector of Locally Aggregated Descriptor) models. *NetVLAD* [159] assigns each frame-level descriptor to a cluster and computes residuals to encode the output features. Its variant *GhostVLAD* [159] excludes some of the original NetVLAD clusters from the final concatenation, such that undesirable speech sections are down-weighted.

Being the front-end the main operation of speaker recognition systems, we will describe in detail some of the algorithms devised for such task, and also a modern end-to-end architecture.

Gaussian Mixture Model

The *Gaussian Mixture Model* (GMM) [78] is a probabilistic model that can be thought as a generalization of k-means clustering, where each cluster relates to a Gaussian distribution. GMM assumes datasets are formed by a mixture of Gaussian distribution with uncertain variables, and the combining factors related to each cluster and Gaussian distribution are probabilities. GMMs have been the most common probability functions in text-independent speaker recognition where continuous features are used. GMMs are trained from a set of acoustic features extracted from the speech data of each speaker. Specifically, each speaker is modeled using a separate GMM, which captures the statistical distribution of the acoustic features for each speaker and estimates the parameters that best fit such distribution.

X-Vector

X-vectors is the name given to the segment level speaker embeddings, generated by the corresponding method and used to discriminate between speakers [78]. The approach is developed on deep neural network embeddings, where the network is based on a time-delay architecture and the feature matching is performed by a separated classifier like principal linear discriminant analysis. The complete process leverages a time delay to extract short term temporal frame-level context to then use a statistic pooling layer to aggregate over the input segment, computing mean and standard deviation. The final speaker embeddings are indeed denoted as x-vectors.

ResNet

The family of *ResNets* architectures, e.g., ResNet-34, ResNet-50, is based on the residual neural networks, previously described in Section 2.1.4. In particular, ResNets are a mixture of CNN and residual blocks, where the number specified in their name, e.g., 34 in ResNet-34, stands for the number of layers of the architecture. This type of algorithms is leveraged for automatic speaker verification by learning from a different representation of the vocal fingerprint of a speaker. Indeed, audio signals can be transformed in a 2D representation, e.g., *spectrograms*, which can represent in a single format the intensity and the spectrum of frequencies of a vocal sample as it varies with time.

2.3.2 Methods for Recommendation

There has been an increasing effort in recommendation literature to devise novel methods to solve the recommendation task. The goal of a preference model is to predict whether or to what extent a user $u \in U$ shows interest in an item $i \in I$, e.g., a song, a movie, or a job candidate. The main categorization for recommendation systems is *collaborative filtering*, *content-based filtering*, and *hybrid systems* [128]. We focus on collaborative filtering techniques, which learn the interaction preference of each user from the knowledge of other users based, for instance, on the similarity of their interaction histories, i.e. the items they interacted with. Collaborative filtering techniques can be subsequently categorized in *memory-based* and *model-based*. The latter includes several systems, based on different architecture types, e.g., *matrix factorization (MF)*, *deep learning-based*. The Figure 2.2 depicts the recommender systems categorization that has just been introduced.

The researchers in recommendation have gradually shifted their attention towards deep-learning based systems. Such systems leverage powerful and cutting-edge architectures, such as GNNs [85, 145, 94, 156, 166], transformers [137], and diffusion models [148]. Due to the extensive research conducted with the first architecture type, the literature nowadays refers to collaborative filtering performed with GNNs as graph collaborative filtering (GCF). Recommender systems

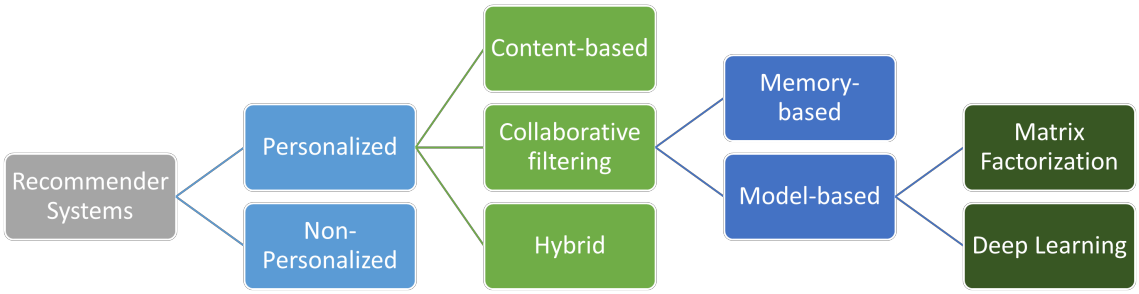


Figure 2.2: Categorization of recommender systems.

based on transformers are particularly employed in sequential recommendation, e.g., BERT4Rec [137], due to the intrinsic property of transformers of predicting the next data instance given a sequence of previous instances. Diffusion models have been proposed to solve the recommendation task by learning to reconstruct the users' past interactions when artificial noise is added to them, e.g., DiffRec and LDiffRec [148].

Nonetheless, several studies have still been focused on improving classic model architectures, such as matrix factorization and autoencoders, For instance, ENMF [26] extended the matrix factorization architecture to a neural recommendation model that can learn the users' preferences without the need of sampling techniques, which is a common practice in recommendation. Another example is represented by EASE [136], a shallow autoencoder that basically corrects the conceptual flaws of neighborhood-based approaches, and that is able to generate recommendations by learning to reproduce the users' interaction history.

Memory-based

The recommender systems that fall in the category of *memory-based* measure the similarity among users on the basis of the ratings they assign to the items. Based on the computed similarity, the items to recommend to a certain user u are derived from the users most similar to u. Some of the items such users interacted with could be recommended to u if the latter do not include such items in the history. This approach is also denoted as *neighborhood-based*, given that it leverages the most similar users to perform the recommendation task, i.e. the neighbors of a certain user. Two main variants exist, *UserKNN* and *ItemKNN*. The former learns the similarity across users by processing the interactions vectors of each user, while the latter process the interactions vectors from the viewpoint of each item by computing the similarity of a possible item to recommend with those the user under consideration interacted with.

Matrix Factorization

Matrix factorization approaches leverage latent representations of each entity in the recommendation scenario, i.e. users and items. Such latent representations, i.e.

embeddings, represent each user and each item in a N-dimensional space by N numerical features. The interest of a user towards an item is predicted by performing a matrix multiplication between the users and items embedding. This is the general operation of *MF*, or *Generalized Matrix Factorization (GMF)*, but other variants have been designed that change the way the entities are represented or learn to discriminate different embeddings. For instance, *Singular Value Decomposition (SVD)* performs a more articulated factorization than MF, while *Bayesian Personalized Rank (BPR)* leverages a pairwise loss to guide the embeddings generation towards discriminating positive from negative interactions.

Deep Learning-based

Novel recommender systems are *deep learning-based*, leveraging the deep networks available in the machine learning literature to learn high-order preferences between users and items. The main addition in deep learning-based systems is a neural network that extends other collaborative filtering algorithms or performs an end-to-end recommendation process. For instance, *Neural Collaborative Filtering (NCF)* extends the classic matrix factorization algorithm by adding a series of MLPs to learn a stack of latent representations. Other architectures re-design the task a neural network is typically used for to treat it as a recommender system. GNN-based recommender systems, for instance, are neural networks devised to perform a linking prediction task in a graph, and given that user-item interactions can be represented by a graph, such task also models a recommendation scenario.

2.3.3 Methods for Algorithmic Fairness Analysis

In the last years, several techniques have been designed to counter unfairness issues in different downstream tasks, such as speaker verification and recommendation. Researchers mainly focused on assessing, explaining and mitigating unfairness issues exhibited by the decisions of models employed in the addressed tasks.

Preliminary research [51] on speaker verification uncovered that a deep-learning model exposes different equal error rates among individuals, based on their language, gender, and age. Data balancing and pre-training strategies across groups were proposed as countermeasures in speaker [171, 108] and face [130] recognition, due to the lack of training data representing the minority demographic group. Subsequent progress included unfairness treatments based on group-adapted encoders [131] or adversarial and multi-task learning techniques [118]. Evaluation frameworks aimed to investigate performance disparities across different demographic subgroups represent another line of research in voice-only [76] and audio-visual biometrics [52]. However, none of the above studies questioned the origin of the disparities, beyond data imbalance. Understanding *why* a speaker recognition system may lead to disparate performance for different (groups of) users is still an under-explored topic, though being essential to enable such systems for everyone.

On recommender systems, preliminary studies exhibited the lacking of a consensus on how to perceive unfairness from a consumer perspective. Indeed, prior works were built upon a certain notion of fairness according to diverse aspects. Such notions can be categorized in two principles: equity of certain metric scores (e.g., the recommendation utility scores) between demographic groups (`EQ`) [47, 21, 155, 81, 9, 124, 92, 141]; independence of a certain outcome (e.g., the predicted relevance scores or recommended lists) from the sensitive attribute (`IND`) [93, 154, 157, 58]. Usually, prior studies proposed mitigation procedures coherently built on top of the fairness notion they tackled, e.g., by balancing the representation of groups in the training set [47], re-ranking items such that both recommendation utility and fairness are improved [92, 141], and decoupling the user and item latent representations from sensitive attribute information [93, 154, 157]. Nonetheless, consumer unfairness was assessed by varied evaluation protocols, comprising a diverse range of datasets, utility and fairness metrics, resulting in a convoluted landscape for fairness analysis in recommendation. Differently from the literature in speaker verification, explainability concerns have been remarkably explored in recommendation. Several methods [20, 174, 10, 169, 61, 32, 140, 97] have been especially proposed to provide predictions explanations (also denoted as *local* [127], *instance-level* [168]) about why each individual item was suggested to each user [31]. Conversely, explanations generated for aspects related to the whole system (denoted as *global* [5], *model-level* [168] or *dataset-level* [39]) are still under-explored.

We present a categorization of fairness-aware methods aimed to explain and mitigate unfairness (*unfairness mitigation methods*), with a particular focus on techniques devised for the latter task. In particular, we focus on the user-side fairness and on personalization systems, e.g., recommender systems. In such systems, fairness at the user side is typically denoted as *consumer fairness*, i.e. the users receiving the recommendations are denoted as *consumers*.

Down-Sampling Techniques

Down-sampling techniques are based on the assumption that unfairness is solely intrinsic to the data fed in input to a model. Indeed, such methods work in a *preprocessing* [107] fashion by modifying the data through balancing and pre-training strategies. Recent works investigated the application of such techniques by balancing the demographic groups representation in the dataset as an unfairness countermeasure in speaker [171, 108] and face [130] recognition. Other studies in personalization systems, such as recommender ones, adopted the same strategy to balance the consumers' representation on two datasets on the movie and music domain [47]. This type of unfairness mitigation can be helpful to counter unfairness, but its positive effects are not reliable and consistent across several experiments, datasets and considered sensitive attributes.

Regularization-based Techniques

Differently from pre-processing techniques as the down-sampling ones introduced in Section 2.3.3, the *regularization-based* methods account also for the model impact on the outcomes unfairness, denoted indeed as *in-processing* methods. The learning process of a recommender system is modified by an extended loss function that simultaneously optimize also for another task, with the goal of mitigating the unfairness. [21] devised an unfairness mitigation method tailored for neighborhood-based systems, which constraints the training process towards considering neighborhoods that are balanced across demographic groups. [81] studied a suite of objective functions with the goal of making recommender systems learn the consumers' preference patterns independently from the sensitive information of the users. [155] leveraged the optimization of multiple objectives unified in a single loss function to find the optimal configuration resulting in all the objectives being minimized or maximized. Among the multiple objectives, two of them focus on mitigating unfairness on the consumer-side and on the items side.

Post-Processing Techniques

Following the taxonomy proposed by [107] to distinguish unfairness mitigation methods, we proceed to describe *post-processing* techniques. Differently from pre- and in-processing methods, post-processing ones are applied directly on the model outcomes, e.g., predictions or recommendation lists. Recent works proposed to mitigate unfairness by leveraging integer programming to maximize utility and minimize fairness on recommendation lists [92], adding fake users as "antidote" data to the unfairness issue [124], or reducing bias disparity with regard to each demographic group by adopting a greedy algorithm [141]. Re-ordering the lists recommended to the consumers is a popular strategy in recommendation, and it is indeed leveraged to improve fairness together with other properties, such as explainability, serendipity, coverage [10].

Counterfactual Techniques

Here we delve into how *counterfactuality* is used in recent studies to examine the unfairness issue based on the discussion on counterfactual reasoning previously introduced in Section 2.2.3. Counterfactual techniques study the unfairness by analyzing the effects of altering how the information related to the individuals in the data is processed. Recent works adopted different type of alterations in terms of which information was targeted, but also varied across which stage leveraged counterfactual reasoning in the learning process. For instance, [93] adopted the proper notion of *counterfactual fairness* to devise a method that modifies the latent representation of the consumers, such that their sensitive attributes were not encoded in such representation. In this way, a model would perform unbiased decisions on the basis of the resulting bias-free embeddings, satisfying counterfactual fairness. Conversely, other

works [44, 60, 59] focused on altering the initial representation of the consumers or the relationship among them. Such works lie withing the sphere of global explanations, such as to uncover the data features that caused a certain issue, such as model instability [117] and unfairness [39, 59]. A similar line of research on explainability is employed in the GNNs literature under binary classification and recommendation tasks. Specifically, counterfactual reasoning drives methods to find the minimal perturbation to the graph (in terms of nodes or edges) fed in input to the GNN such that the prediction changes [97, 82, 32].

Sensitive Information Independence through Adversarial Techniques

A recent and effective technique adopted in machine learning to counteract unfairness is leveraging *adversarial* techniques. Such methods are used to optimize simultaneously two main objectives: the first aims to solve the task under consideration, e.g., recommendation, while the other seeks to "fool" an additional model. The additional model can be denoted as *sensitive attribute predictor* or *discriminator*, whose goal is to predict the sensitive information of a user from its latent representation. Therefore, unfairness mitigation methods based on adversarial reasoning are trained to also alter the users' latent representations, such that it is harder for a discriminator to predict the sensitive information and the resulting representations are independent of the protected attributes. This is done by learning a set of *filters* to apply on the latent representations, that lead to an altered one where the users' sensitive attributes are obfuscated to the discriminator. This techniques was successfully applied in speaker recognition to treat unfairness in a multi-task learning process [118]. In recommendation, [93] generated bias-free embeddings to improve counterfactual fairness, [154] adopted adversarial methods to generate embeddings that could lead to unbiased predictions for news recommendation, and [157] leveraged a set of filters and discriminators that resulted in embeddings being independent of multiple sensitive attributes in a GNN-based recommendation task.

Chapter 3

Fairness in Speaker Verification

3.1 Introduction

Increasingly adopted in online and on-life applications, *speaker recognition systems* aim to confirm or refute the user's identity based on the characteristics of the user's voice [77, 33]. Current successful applications include scanning passengers during border controls, checking identities for bank transactions, forensics analysis, and remote access to computers (e.g., online exams) [138]. In particular, speaker recognition is a driver for *personalization* in *voice-based interfaces* and *assistants*, such as Amazon Alexa and Google Home, to detect the active speaker based on the *voiceprints*, i.e. a unique voice signature, and provide personalized responses. Personalizing the handling of voice queries amid social voice environments is a key feature, considering that voice-based interfaces are becoming commonplace in workspaces [75]. In a common speaker recognition system, speech samples are provided by the user, and the resulting *utterances* are processed to create the enrolled speech model for that user. The vocal sample presented at authentication time is then compared with the enrolled speech model to make the decision.

Achieving the highest possible accuracy has been a primary goal along the years [42, 79, 111, 115]. However, recent literature in the machine-learning community highlighted that achieving impressive accuracy cannot be the sole goal for machine-learning models shaped for our society [65, 15, 24]. When consequential decisions are made about individuals on the basis of the outputs of speaker recognition systems, concerns about *discrimination* and *fairness* inevitably arise. Indeed, it may happen that the systems outputs result in decisions systematically *biased* against individuals in certain *demographic groups*. This might be due to differences in *dialects* (e.g., because of regional accent), *inter-group heterogeneity* (e.g., age, gender, or ethnicity), or *speech pattern variability* of each individual in the group (e.g., people with disabilities).

This behavior may result in certain groups being offered limited services from personalization systems (e.g., Alexa, Google Home), being unfairly denied access to a platform or being more vulnerable to attackers, with both *usability* and *security*

issues. Cognisant of this problem, a timely research paradigm of fair machine learning emerged, attempting to *mitigate* [51, 45, 129, 130, 37] and *explain* [39, 89] this unfairness, often referring to fairness as a concept of non-discrimination based on the membership to protected groups. However, several questions connected to how much unfairness issues affect speaker recognition systems still remain *unanswered*, remarked by the different and under-explored fairness notions recently proposed [107].

In this chapter, we extend the literature of fairness in speaker recognition systems b: (i) providing a *fairness-benchmarking protocol* for assessing how much speaker recognition systems are fair, (ii) investigating the relationship among *group fairness metrics* in speaker recognition to determine the metric yielding more disparity in recognition performance across groups, and (iii) uncovering the *influence of voice characteristics* on the disparate error rates emphasized by speaker recognition models. To this end, we designed two frameworks to inspect multiple speaker recognition systems: (i) a framework based on automated pipelines to measure the unfairness of these systems decisions, and (ii) an explanatory framework to analyze the voice covariates affecting the fairness of these models outcomes. Several deep-learning architectures were trained on the multi-language audio samples contained in *FairVoice* [51], a dataset based on the resources provided by *Mozilla Common Voice*[1]. Our experiments highlight the benefits of unfairness mitigation and explanation techniques applied to speaker recognition models.

This chapter presents in detail the following contributions:

- We propose a *multi-architecture framework* which makes it possible to train, evaluate, and inspect multiple speaker recognition systems by means of automated pipelines to measure the accuracy and fairness of identification trials of speakers from different demographic groups.

- We performed an *extensive analysis* of the adoption of an unfairness mitigation procedure on state-of-the-art speaker recognition models. We capitalized on a setting with multiple sensitive attributes and group fairness metrics to underline the generalizability of this approach under various scenarios.

- We extracted several voice characteristics reflecting speakers' unique traits. They were leveraged by an *explanatory framework* to provide key observations on the impact of these voice characteristics on the outcomes unfairness of speaker recognition architectures.

The rest of this chapter details the listed contributions as follows: Section 3.2 formulates the speaker recognition task, and defines accuracy and fairness metrics for speaker verification, Section 3.3 presents a methodology aimed to mitigate the biased outcomes of speaker recognition systems, Section 3.4 shows an approach to uncover the voice features leading such models towards unfair predictions. Finally,

[1] `https://commonvoice.mozilla.org/`.

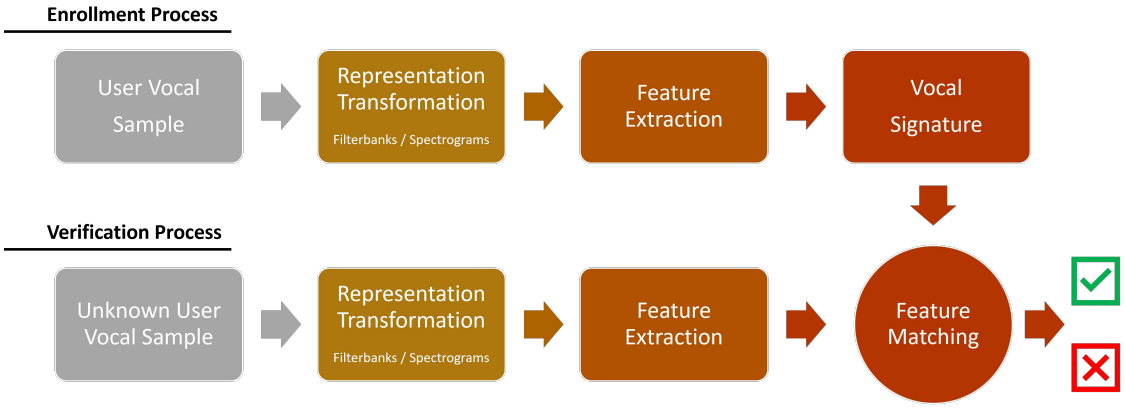


Figure 3.1: *Enrollment* (at the top) and *Verification* (at the bottom) process in a speaker recognition system.

Section 3.5 summarizes the findings resulting from the work presented in this chapter and future works to extend the research in this field.

3.2 Problem Formulation

In this section, we formalize the speaker recognition task, and present the metrics selected to measure the accuracy and to operationalize several fairness notions.

3.2.1 Speaker Recognition Task

Given a set of speakers U, we can denote as $A \subset \mathbb{R}^*$ the domain of *audio signals* with unknown length produced by the speakers in U. We can consider a traditional processing pipeline stage $\mathcal{F} : A \to S$, that generates an *intermediate acoustic representation* $S \subset \mathbb{R}^{k \times *}$, e.g., a spectrogram, where k is the feature vector dimensionality. This intermediate representation is leveraged in an explicit *feature extraction step* by an encoder $\mathcal{D}_\theta : S \to D$, that produces fixed-length representations $D \in \mathbb{R}^e$, where e is the embedding dimensionality. Given a decision threshold τ, a *verification trial* can be defined as:

$$v_\tau : D_{\theta,u'} \times D_{\theta,u} \to \{0, 1\} \tag{3.1}$$

where, under the feature extraction parameters θ, an input feature vector $d_{u'}$ from an unknown user u' is compared with a feature vector d_u from user u to *confirm* or *refute* the identity of the speaker u (1 and 0, respectively).

Given that multiple audio samples could be available for the enrolled user u, the respective feature vector d_u could be generated in several ways, such as the average of

the enrollment speaker embeddings or the creation of a single embedding by pooling utterances. In our study, we consider a *one-shot verification policy* to align with the evaluation protocol of relevant prior works in traditional speaker recognition. Such protocol assesses the performance of the model by evaluating it along a list of trial verification pairs, denoted as $\mathcal{P} = \{(d_u, d_{u'}) \mid u, u' \in U \wedge u \neq u'\}$, such as the trial test pairs in the VoxCeleb-1 set [33]. Our verification policy relies on a similarity function $\phi : D \times D \to \mathbb{R}$ and it is formally defined as:

$$v_\tau(d_{u'}, d_u) = \mathbb{1}[\phi(d_{u'}, d_u) > \tau] \tag{3.2}$$

Based on this verification policy, training a speaker recognition model becomes an optimization problem. This means finding the model parameters θ and verification threshold τ that maximize the expectation on the following objective function:

$$\underset{(\theta,\tau)}{\operatorname{argmax}} \underset{u,u' \in U}{\mathbb{E}} \begin{cases} v_\tau\left(D_{\theta,u'}, D_{\theta,u}\right) & u' = u \\ 1 - v_\tau\left(D_{\theta,u'}, D_{\theta,u}\right) & u' \neq u \end{cases} \tag{3.3}$$

In other words, we aim at maximizing the cases where $v_\tau\left(D_{\theta,u'}, D_{\theta,u}\right) = 1$ when $u' = u$ and those where $v_\tau\left(D_{\theta,u'}, D_{\theta,u}\right) = 0$ when $u' \neq u$.

3.2.2 Task Optimization Targets

Accounting for unfairness issues in machine learning establishes a multi-criteria setting, where both accuracy and fairness are relevant. We focused on *group fairness*, whose goal is to guarantee the decisions of machine learning model are fair across demographic groups. Therefore, we use U_1, U_2 in this section to denote two partitions of the user set U based on the belonging of speakers to two demographic groups, e.g., males and females. It should also be noted that the accuracy and fairness metrics adopted in a speaker verification task rely on the decision threshold selected to decide if two audio samples belong to the same speaker or not (see (3.2)). Hence, the notation M_τ will be used to denote any metric M that depends on the value of the decision threshold τ. We first describe the metrics used to measure the identification accuracy of speaker recognition systems. Then we present how the notions in Section 2.2.3 can be operationalized to estimate these systems fairness.

False Acceptance Rate (FAR)

Grounded to evaluation metrics used in classic machine learning, this score is equivalent to the *False Positive Rate (FPR)*, which measures the probability that a given condition exists when it does not. Given the identification task performed by speaker recognition systems, the literature uses the term *false acceptance* instead of false positive to denote an impostor being incorrectly identified as the legitimate user. Consequently, FAR measures the probability that the speaker recognition system incorrectly accepts an access attempt by an impostor. Formally, for a speaker u:

$$FAR_\tau(u) = \frac{|\{(d_u, d_{u'}) \in \mathcal{P} \mid v_\tau(d_{u'}, d_u) = 1 \wedge u \neq u'\}|}{|\{(d_u, d_{u'}) \in \mathcal{P} \mid u \neq u'\}|} \tag{3.4}$$

In other words, FAR is measured by dividing the number of false acceptances by the number of impostor attempts, and it is then associated to the *security* of the system. The lower the FAR, the higher the security.

False Rejection Rate (FRR)

As FAR is equivalent to FPR, the false rejection rate is equivalent to the *False Negative Rate (FNR)*, which measures the probability that a given condition does not exists when it does. In the context of identification systems, FRR measures the probability that the model incorrectly fails to authenticate a legitimate user. For a speaker u, it is formally defined as:

$$FRR_\tau(u) = \frac{|\{(d_u, d_{u'}) \in \mathcal{P} \mid v_\tau(d_{u'}, d_u) = 0 \wedge u = u'\}|}{|\{(d_u, d_{u'}) \in \mathcal{P} \mid u = u'\}|} \tag{3.5}$$

In other words, FRR is calculated as the ratio between the number of false rejects and the number of genuine attempts, and it is associated to the *usability* of the system. The lower the FRR, the higher the usability.

Equal Error Rate (EER)

This score represents the error obtained at the threshold where the FAR and the FRR are equal. Hence, EER outlines the threshold where a speaker recognition system reports the highest security and highest utility, while, at the other thresholds, one of the two aspects is less guaranteed. It is formally computed as:

$$EER_\tau(u) = \frac{FAR_\tau(u) + FRR_\tau(u)}{2} \tag{3.6}$$

In other words, EER estimates the average error done by the speaker recognition system in terms of both security (FAR) and utility (FRR).

Disparity Score (DS)

Prior studies identified as fair a system able to recognize the speakers with the same performance across demographic groups. Denoting EER as the performance of speaker recognition systems, it follows that unfairness can be estimated as the disparity of EER across demographic groups. The disparity score reflects this concept by measuring unfairness as the absolute value of the difference between two EERs, associated with two different demographic groups, e.g., male users' EER and female users' EER or over-40 users' EER and under-40 users' EER. DS can be defined as:

$$\begin{aligned} DS =& |\overline{EER}_\tau(U_1) - \overline{EER}_\tau(U_2)| \\ \text{s.t} \quad \overline{EER}_\tau(U_z) =& \frac{\sum_{i=1}^{|U_z|} EER_\tau(u_{z,i})}{|U_z|} \end{aligned} \tag{3.7}$$

Precisely, a disparity in EER across groups means that, for that model, can be easier/harder to recognize users within certain groups.

Disparity in Demographic Parity (DP)

As described in Section 2.2.3, *demographic parity* is satisfied when the likelihood of a speaker being positively recognized is the same regardless of the demographic group. Given that the classic evaluation protocol for binary classification tasks in machine learning consists in counting the *true positives (TP)*, *false positives (FP)*, *true negatives (TN)*, *false negatives (FN)*, we can denote the likelihood of a speaker being positively recognized as the *positive rate* $PR_\tau = (TP_\tau + FP_\tau)/(TP_\tau + FP_\tau + TN_\tau + FN_\tau)$. Then, the metric DP implies that PR should be the same regardless of the demographic group and can be instantiated as follows:

$$DP_\tau(U_1, U_2) = |PR_\tau(U_1) - PR_\tau(U_2)| \tag{3.8}$$

Disparity in Equal Opportunity (EOpp)

As described in Section 2.2.3 and following the classic evaluation protocol in machine learning as for demographic parity, *equal opportunity* implies that the probability of a speaker being correctly verified should be equal across demographic groups. In other words, the equal opportunity definition states that all the demographic groups should have equal true positive rates (TPR), measured as $TPR_\tau = TP_\tau/(TP_\tau + FN_\tau)$. This notion is operationalized as follows:

$$EOpp_\tau(U_1, U_2) = |TPR_\tau(U_1) - TPR_\tau(U_2)| \tag{3.9}$$

Disparity in Equalized Odds (EOdd)

As described in Section 2.2.3 and following the classic evaluation protocol in machine learning as for demographic parity and equal opportunity, *equalized odds* implies that the likelihood of a speaker being correctly verified and of being incorrectly verified should both be the same across demographic groups. In other words, the equalized odds definition states that the demographic groups should have equal rates for true positives (TPR) and false positives (FPR) ($FPR_\tau = FP_\tau/(FP_\tau + TN_\tau)$)[2]. Equalized odds is instantiated as follows:

$$\begin{aligned} EOdd_{\tau,TPR}(U_1, U_2) &= |TPR_\tau(U_1) - TPR_\tau(U_2)| \\ EOdd_{\tau,FPR}(U_1, U_2) &= |FPR_\tau(U_1) - FPR_\tau(U_2)| \end{aligned} \tag{3.10}$$

[2]As mentioned in Section 3.2.2, it is equivalent to FAR, but here we denote it as FPR for consistency with the fairness literature.

Fairness Discrepancy Rate (FDR)

Fairness Discrepancy Rate is a metric proposed in [37] that takes simultaneously into account the false positive rate (FPR) and the false negative rate (FNR) ($FNR_\tau = FN_\tau/(FN_\tau + TP_\tau)$). Even though the authors mention FDR as being an operationalization of the equalized odds notion, it instantiates the notion of *treatment equality*, which *"is achieved when the ratio of false negatives and false positives is the same for both protected group categories"*, according to [107]. FDR is formally defined as:

$$FDR_\tau(U_1, U_2) = 1 - (\alpha FPR_\tau + (1 - \alpha)FNR_\tau) \tag{3.11}$$

where α is a hyper-parameter that defines the importance of false positives, i.e. the security of the system.

3.3 Techniques for Unfairness Assessment and Mitigation

This section describes the proposed adoption of a pre-processing technique aimed to mitigate unfairness across demographic groups in speaker recognition systems. The work carried out in this study resulted in a fairness-aware framework composed of three main components:

1. A *testing balancing module* that generates testing sets for a comprehensive assessment of the fairness of speaker recognition systems. A subset of speakers that equally represent all the demographic groups is sampled to prepare the testing sets, such that the verification pairs are meticulously selected to challenge the fairness of models.

2. A *training balancing module* that samples the same number of speakers for each demographic group from the data subset left by the first component. The sampled speakers are processed to extract their utterances to generate a training set. Thus, the training split consists of audio samples coming from a set of speakers where each demographic group is *equally represented*.

3. A *fair evaluation module* consisting of an automated pipeline that gathers the results of the speaker recognition systems under the verification trail evaluation (based on the testing sets of the first module), and assess the fairness level under several metrics, as the ones introduced in Section 3.2.2.

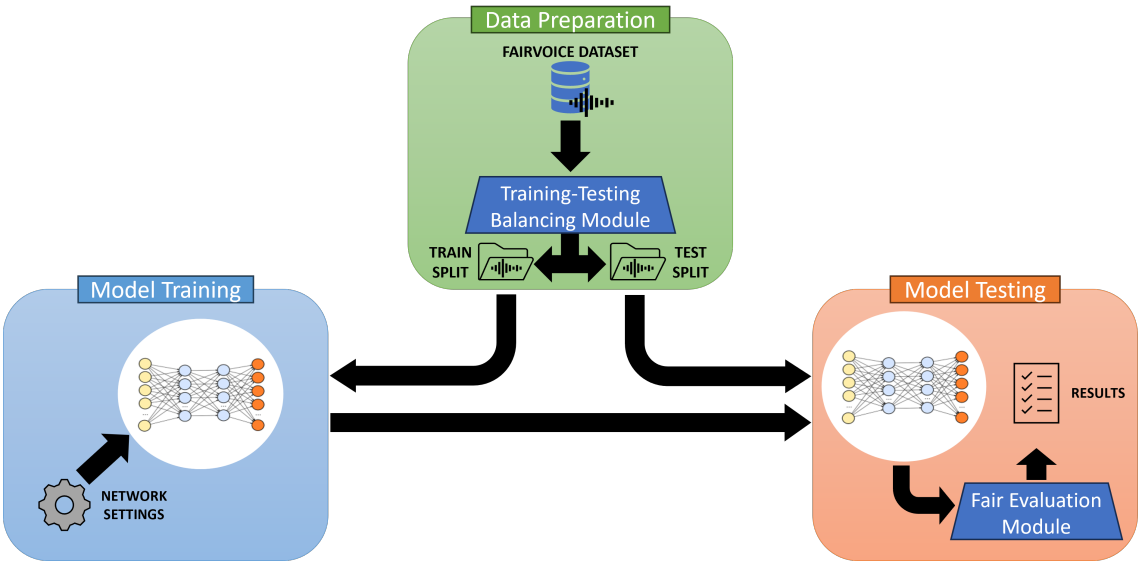


Figure 3.2: Structure of the automated pipelines in our framework: (i) training and testing sets are created for fair learning and evaluation procedures, (ii) the speaker recognition system is trained under a fairness-aware context, (iii) the evaluation results of the model are processed to measure its accuracy and fairness performance.

3.3.1 Methodology

Testing Balancing Module

This module samples the speakers from which the trail verification pairs will be created following a specific methodology, ad-hoc for fairness evaluation. Specifically, N speakers are sampled from the dataset, such that each demographic group is equally represented, and we denote them as *test speakers*. For instance, given a scenario where each speaker is characterized by the sensitive binary attributes gender and age, and $N = 100$, this module would sample 25 test speakers to represent each demographic group, i.e. 25 younger males, 25 older males, 25 younger females, 25 older females. Additionally, we also adopt a constraint on the number of utterances provided by each one of the N speakers, such that at least δ verification trial pairs could be created to guarantee a comprehensive evaluation of the speaker recognition systems. Specifically, for each test speaker u_i, δ unique trials pairs against utterances of the same speaker u_i and δ unique trial pairs against utterances from a different speaker u_j, $i \neq j$ are considered. The trial verification pairs between the different speakers u_i and u_j are created considering a demographic group shared by both users. Given that our evaluation protocol considers two sensitive attributes (gender, age), it follows that this module generates two types of testing sets, an *intra-gender* one and an *intra-age* one. Specifically, the set of intra-gender trial pairs have been constructed such that u_i and u_j belong to the same gender group, while the intra-age trial pairs have been constructed with u_i and u_j coming from the same age group.

Intra-group trial pairs have been often proved to be the most challenging ones to recognize, so our study uses the intra-age testing set when assessing unfairness on age and intra-gender testing set when assessing unfairness on gender. This ensures an adequate representation of each demographic group, making these testing sets a suitable tool for fairness evaluation in speaker recognition.

Training Balancing Module

The remaining speakers in the dataset that are not test speakers, i.e. they are not part of the testing set, are used to train the speaker verification models, and we denote them as *train speakers*. The training balancing module applies a *balancing strategy* to generate a training set with the same number of speakers for each demographic group. For instance, considering the gender, this means having an equal distribution of male and female speakers. In this way models can be trained on a more balanced dataset, a strategy that reported promising results in unfairness mitigation in other machine learning fields [130]. The audio files used for training included only those of the train speakers, i.e. those who are not part of the testing set. The balancing process is based on the less represented group, such that the number of speakers of the least represented group remains constant, while the other groups' representation follows the least represented one. This methodology results in two types of training sets:

- **NB**: we consider the full dataset of utterances without any type of balancing, i.e. fully unbalanced dataset.
- **UB**: we randomly sampled the same number of train speakers for each demographic group to create a *user-balanced* set.

In case of a multi-language dataset, we generate the two training sets types for each language, e.g., *Spanish NB*, *English NB*. Moreover, under each setup, we controlled that the same number of train speakers was included across languages, for fair comparison of the results across languages as well. In fact, our study is also interested in evaluating whether the language may be a covariate that leads to unfair performance of a model. This point can promote a better understanding of how a model fairly performs in real world.

Fair Evaluation Module

The first two modules perform operations only at the pre-processing level in our framework. Then, the speaker recognition models are trained on the training sets generated by the training balancing module, and the performance of such systems is evaluated on the testing sets generated by the testing balancing module. The *fair evaluation module* is based on an automatic pipeline that gathers the performance of a speaker recognition system on each trial verification pair of a testing set.

The performance is measured by the similarity function ϕ used in (3.2), such that the fairness of a model can be evaluated under several decision thresholds τ. The module takes the performance scores and estimates the fairness using five metrics described in Section 3.2.2: disparity score (DS), disparity in demographic parity (DP), disparity in equal opportunity (EOpp), disparity in equalized odds (EOdd), fairness discrepancy rate (FDR).

3.3.2 Experimental Setup

Dataset

Despite the existence of several datasets for speaker verification evaluation [7, 119, 33], our fairness study was conducted on FairVoice [51], a dataset that offers a large number of utterances and labeled speakers across several sensitive attributes and languages. The dataset was sampled from *Mozilla Common Voice*, one of the largest corpora including unconstrained speech from diverse acoustic environments. All the waveforms were single-channel, 16-bit recordings sampled at 16 kHz. We selected this dataset since it covers a wide range of demographic groups identified by their language, gender, and age, and the labels which describe such sensitive attributes are available for each individual speaker. However, not all the languages provided by FairVoice include *enough female speakers* in order to set balanced datasets that are sufficiently large to train state-of-the-art deep speaker recognition systems. Specifically, among the specific-language datasets of *English*, *Spanish*, *French*, *German* speakers, we only considered the first two languages because they embrace enough utterances for each demographic group.

The pre-selected dataset has been filtered by the number of utterances per speaker. Specifically, only the speakers who have provided *at least five samples* were taken into consideration in our analyses. This step is essential because we require to create trial verification pairs with both utterances coming from the same speaker in order to simulate legitimate authentication scenarios. Hence, if a speaker does not provide a minimum number of utterances, we cannot create enough trial pairs. This filtering step led to a total of 6,321 English speakers and 1,298 Spanish speakers. In our study, we analysed disparities conveyed by speaker recognition systems on four demographic groups per language, based on their *gender* (female, male) and their *age* (speakers under and over 40 years old), so only speakers with both gender and age labels were considered, resulting in 6,246 English speakers and 1,280 Spanish speakers. We selected 40 as a splitting age, since it allows us to better balance the representation of the resulting age groups, while maintaining a reasonable size of the training dataset. Based on the *testing balancing module*, we set $\delta = 64$ for the trail verification pairs and $N = 100$, so considering for each language 100 speakers evenly distributed across the demographic groups (25 speakers for each demographic group). Given that the module creates an intra-gender and an intra-age testing set for each language, the resulting testing sets are four, namely Spanish *Test-Gender*,

Test-Age, and English *Test-Gender*, *Test-Age*. These criteria led to the following distribution of the remaining speakers across languages and demographic groups:

- *English*: 400 over-40 females (6.5%), 718 under-40 females (11.7%), 1,093 over-40 males (17.8%), 3,935 under-40 males (64.0%)
- *Spanish*: 281 over-40 females (23.8%), 155 under-40 females (13.1%), 351 over-40 males (29.8%), 393 under-40 males (33.3%)

As mentioned earlier in Section 3.3.1, the balancing strategy is based on the least represented group. For instance, when balancing the Spanish dataset on the number of speakers per group, we can observe that under-40 females are less represented, with 155 speakers. Hence, we filter only 155 speakers of each group to perfectly balance data across these groups for Spanish. Additionally, this module controls that the same number of training speakers is included across languages. It follows that the balanced training sets *Spanish UB* and *English UB* include utterances of 620 speakers in total, 155 for each demographic group, while the unbalanced training sets *Spanish NB* and *English NB* are composed of utterances coming from the speakers distribution just detailed for each language.

Speaker Recognition Models

The speaker recognition models trained and tested for these experiments are some of the ones previously introduced in Section 2.3.1, namely *X-Vector*, *ResNet-34*, *ResNet-50*. For clarity, we specify the architecture structure and the data features dimensionality these models work with. X-Vector takes 24 dimensional filterbanks of size 24×300 (frequency×temporal) as input with a frame-length of 25ms, mean-normalized over a sliding window of up to 3s. Spectrograms of size 257 x 250 are generated by a 512-point Fast Fourier Transforms (FFTs) to be fed in input to ResNet-34 and ResNet-50 for 3s of speech using a hamming window of width 25ms and step 10ms. The difference between these two models is that the former is composed of 34 residual layers, while the latter by 50 residual layers. The ResNets models were adapted from computer vision to spectrogram inputs by replacing the last fully-connected (FC) layer with two layers: an FC one with support in the frequency domain and average pooling with support in the time domain. On the other hand, X-Vector includes five layers that operate on speech frames, with a time context centered at the current frame. A pooling layer aggregates frame-level outputs and computes mean and standard deviation. Two FC layers aggregate statistics across the time dimension. We used a GhostVLAD pooling [159].

From each speaker's utterance we randomly sampled segments and standardized the inputs to 2-second clips (by cropping or padding). No voice activity detection or silence removal was applied. Each acoustic vector was normalized by subtracting the mean and dividing by the standard deviation of all frequency components in a single time step. The models were trained for classification using Softmax, and

Table 3.1: Performance of the considered speaker recognition models at the EER security level.

Train set	Test set	English			Spanish		
		ResNet-34	ResNet-50	X-Vector	ResNet-34	ResNet-50	X-Vector
NB	Intra-age	*0.08*	0.09	*0.08*	*0.06*	**0.05**	*0.06*
	Intra-gender	0.11	0.13	0.11	0.08	0.08	0.08
UB	Intra-age	**0.07**	**0.07**	*0.08*	0.07	*0.06*	*0.06*
	Intra-gender	0.11	0.1	0.11	0.09	0.08	0.08

served with 32-sized batches. We used the Adam optimizer, with an initial learning rate of 0.001, decreased by a factor of 10 every 10 epochs, until convergence.

By leveraging the training files previously arranged, we were able to train several speaker recognition models under different setups. Specifically, we performed 12 model learning processes, given that three models were trained on four training sets, two for each language. For the sake of clarity, we considered the same model parameters and the same training parameters described by the authors of each deep-learning architecture. Even the parameters for acoustic extraction, i.e. spectrogram or filterbank computation, were kept consistent with respect to the original papers [133, 114]. Our framework allows to setup the parameters of a training process, e.g., the architecture type, the batch size, the learning rate, and so on.

3.3.3 Results

Speaker Verification Performance

We first evaluate the performance of the speaker recognition models in terms of *equal error rate* (EER) and report the results in Table 3.1. Even though other security levels are considered in the literature, e.g., FAR1%, we focus on the EER because the fairness performance analyzed in the next sections accounts for the EER security level. All the models in all the settings have *comparable performance*, and no one outperforms the others. Evaluation on the English intra-gender set counts more errors in the verification pairs for all the models, with EER values reported over 0.1 only in these settings. The balancing strategy applied on the training set (UB) does *not impact* the performance of the models, making this method reliable to counteract unfairness while maintaining a good EER level. In order to support this claim, the following experiments evaluate the speaker recognition systems in terms of the fairness estimated by the metrics used by the *fair evaluation module*.

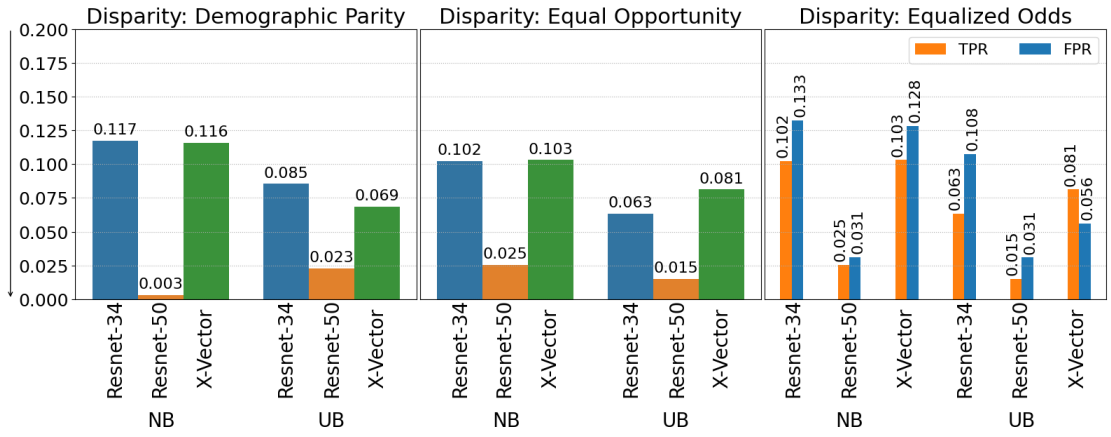
Table 3.2: Disparity score between age groups (DS Age) and gender groups (DS Gender) of the speaker recognition systems under the unbalanced (NB) and user-balanced (UB) training settings, and the intra-gender and intra-age testing settings.

	Train Set	Test Set	DS Age			DS Gender		
			ResNet-34	ResNet-50	X-Vector	ResNet-34	ResNet-50	X-Vector
English	NB	Intra-age	5.58	4.82	6.99	3.21	2.45	4.32
		Intra-gender	*3.46*	4.52	5.42	1.18	**0.12**	1.84
	UB	Intra-age	**1.95**	4.33	6.97	4.27	4.02	4.86
		Intra-gender	4.61	6.84	5.23	0.54	*0.48*	3.14
Spanish	NB	Intra-age	1.18	**0.10**	0.95	*0.29*	1.41	0.54
		Intra-gender	2.18	3.37	0.67	1.33	1.14	**0.13**
	UB	Intra-age	1.29	*0.36*	1.07	1.09	0.80	1.09
		Intra-gender	2.33	2.79	0.57	1.65	0.47	0.94

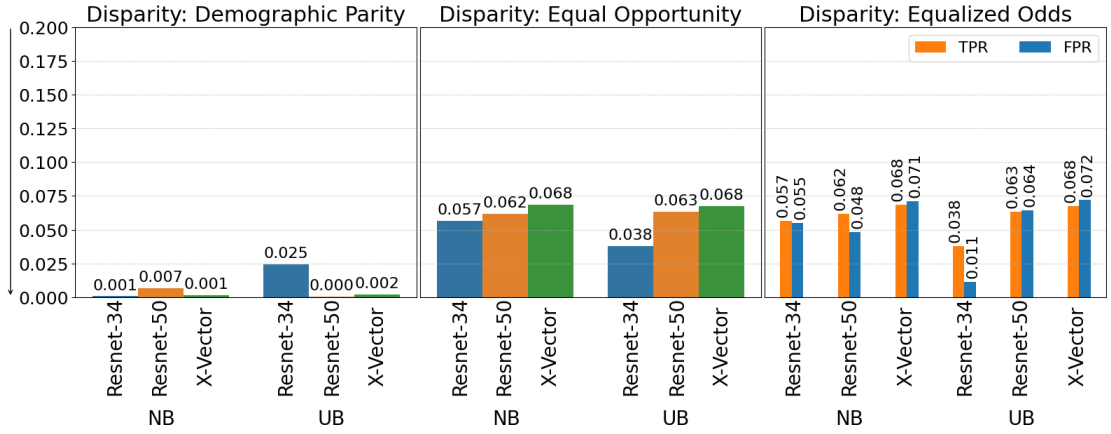
Fairness Performance in Error Rates

Before delving into the impact of our balancing strategy on various fairness notions, we focus on estimating the system fairness by the *disparity in error rates*, similarly done by [51]. To do so, we employ the metric DS used by the fair evaluation module to estimate the difference in EER experienced by the demographic groups defined by the same sensitive attribute, e.g., between males and females by the gender attribute. The results of our experiments are depicted in Table 3.2. To highlight the extent to which each testing set is challenging on the respective setting, we report DS Age and DS Gender even when intra-gender and intra-age testing sets are used respectively.

Several experiments report lower DS values, i.e. fairer, in the settings where the measured DS and the testing set focus on the same sensitive attribute. Especially on English, DS Gender is lower when the intra-gender testing set is used for evaluation compared with the setting where the intra-age testing set is adopted. The user-balanced training set (UB) had a *positive impact* on several settings, but this observation does not *hold systematically*. Focusing on the experiments where the measured DS is related to the testing set, e.g., DS Age with intra-age test set, the balancing strategy is effective on mitigating the disparity in EER between over-40 and under-40 users on English, but it is the opposite on Spanish. Results on DS Gender with the intra-gender test set are effective on some settings such as ResNet-34 on English and ResNet-50 on Spanish. It follows that the balancing strategy *can help in mitigating* the disparity in error rates, but it is *not enough reliable* to be effective in most of the experiments it could be adopted for. Further experiments with different languages, datasets, and models could better confirm the unreliability of this balancing method in mitigating unfairness across speakers' groups.



(a) EER - Gender



(b) EER Age

Figure 3.3: Fairness estimates of three deep neural architectures (X-Vector, ResNet34, ResNet50), under different training data balancing (NB: unbalanced; UB: user-based balance across demographic groups). The lower the metric is, the fairer the model is.

Extended Fairness Evaluation

The fair evaluation module is equipped with a wide set of fairness metrics, each one reflecting a different notion and a different perspective on what is perceived fair or unfair. As done for DS, we study the other four fairness metrics, namely *DP*, *EOpp*, *EOdd*, *FDR*, based on their disparity operationalization. However, given that the size of Spanish NB in terms of speakers is much lower than the English NB set, this extended fairness evaluation is carried out only for English to provide findings with statistical significance. We also do not report the results measured with the fairness discrepancy rate (FDR) [37] because the resulting patterns were similar to those obtained by DP.

Figure 3.3 reports the fairness metric scores on each sensitive attribute, under

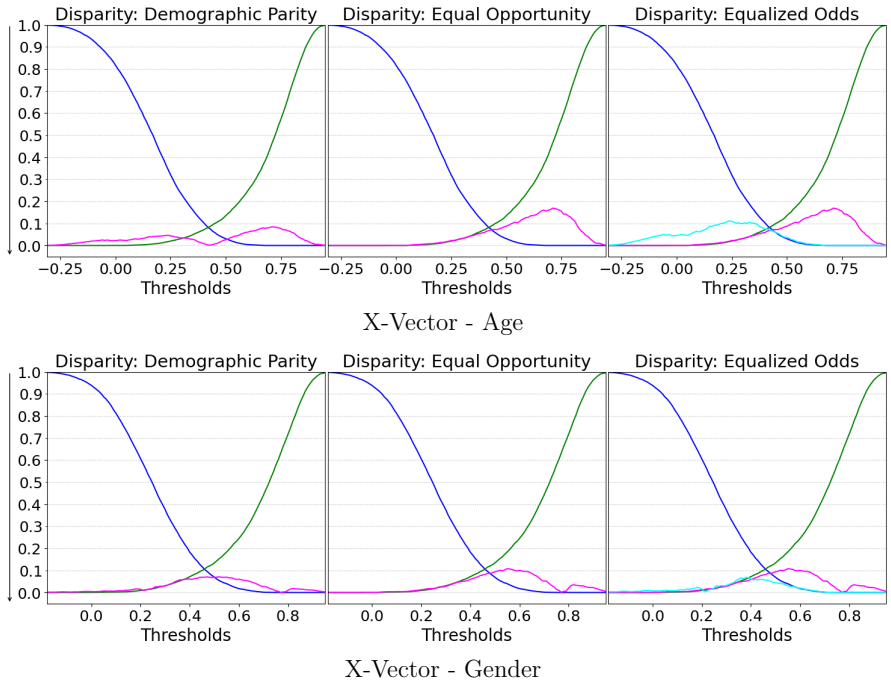
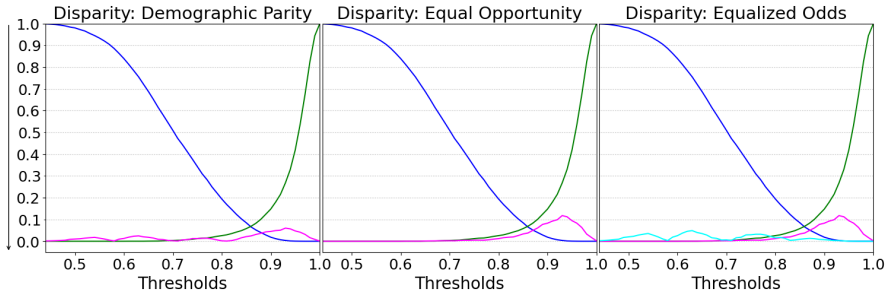


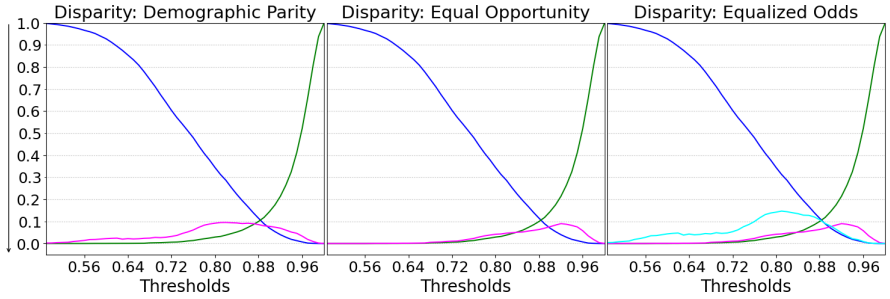
Figure 3.4: The impact of the decision threshold on the trade-off between fairness, security, and usability for X-Vector, three fairness metrics, for the user-based balanced training set (UB). ■ represents the FAR, ■ represents the FRR, ■ represents the respective fairness metric value. For Equalized Odds ■ represents $EOdd_{\tau,TPR}$, while ■ represents $EOdd_{\tau,FPR}$.

different training balancing setups. Balancing users across demographic groups often helps mitigating unfairness. Specifically, the disparities between males and females are *mitigated for all models* under all fairness metrics, except DP for ResNet-50 at EER. The fairness scores on the age-based groups highlight a *good level of mitigation* of the disparity between under- and over-40 users as well, but not for all models, e.g., DP for ResNet-34 at EER. Indeed, ResNet-34 is the one being influenced the most by the balancing of the dataset, followed by X-Vector. Surprisingly, the *ResNet-50* architecture tends to be *fairer* on the *gender-based groups*, while the other two architectures are often *fairer* than *ResNet-50* for *age-based groups*.

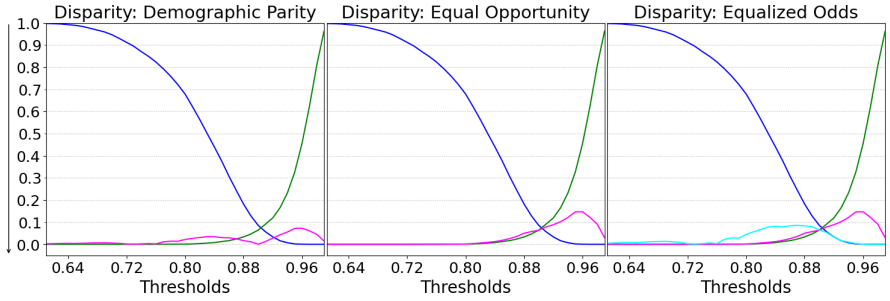
Additionally, we assess the impact of the recognition threshold on the trade-off between (security, usability) and fairness. Given that models trained on UB often led to the *fairest* results, we focused only on these models. For each model, for each threshold between 0 and 1, we computed FAR, FRR, and the fairness estimate, to understand the relation between *security* (FAR), *usability* (FRR), and *fairness* under different fairness notions. Figures 3.4-3.5 report the fairness score, FAR, and FRR as a function of the *recognition threshold*, for each model. For almost all settings, the disparity scores show their *peaks nearby the EER and FAR*



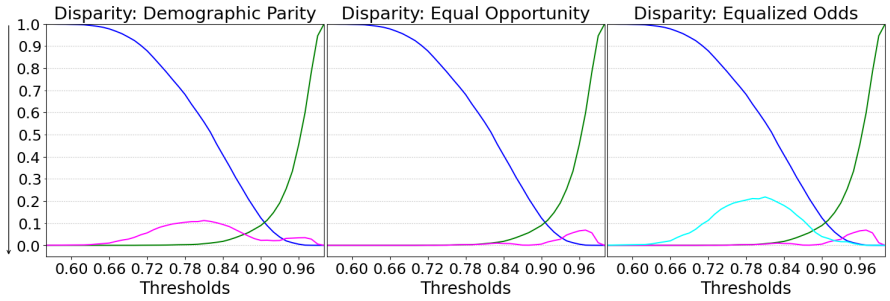
ResNet-34 - Age



ResNet-34 - Gender



ResNet-50 - Age



ResNet-50 - Gender

Figure 3.5: The impact of the decision threshold on the trade-off between fairness, security, and usability for ResNets, three fairness metrics, for the user-based balanced training set (UB). ■ represents the FAR, ■ represents the FRR, ■ represents the respective fairness metric value. For Equalized Odds ■ represents $EOdd_{\tau,TPR}$, while ■ represents $EOdd_{\tau,FPR}$.

1% security levels. Our analyses on the age-based groups shed light on unfairness near the FAR 1% threshold, while experiments with the gender-based groups often show unfairness at thresholds close to the EER one. On gender-based groups, the thresholds at which a model achieves *lower disparities vary across models*. On age-based groups, the thresholds close to EER lead to a degree of unfairness, but not as much as thresholds slightly higher than EER, where the disparities achieve the highest peak. These results highlight the *friction between fairness and accuracy* (FAR and FRR), confirming the trade-off usually experienced in this task.

3.4 Counterfactual Reasoning for Unfairness Explanation

In this section, we describe the foundations of our explanatory framework that aims to study the impact of voice characteristics on the performance of different speaker encoders. The central question in the explanatory modeling research is the choice of explanatory variables. Two main approaches exist for choosing them, based on confirmatory or exploratory research. They can be regarded as two complementary components of the same goal, i.e. finding relevant variables in the most efficient, reliable, and replicable way. The difference is that, in confirmatory research, the potential impact of different variables is hypothesized a-priori, based on existing theories. The confirmatory research approach is useful when researchers have a theory (or theories) supported by facts. The second approach is exploration-driven, which is used when there exists a lack of sufficient theory foundations. Exploratory research could likewise produce new hypotheses that could formally be evaluated later. Our study belongs to the second category, as we design a general framework. With it, we studied the impact of voice characteristics on speaker verification performance, in terms of false acceptance rates, to highlight how the security could be affected by such characteristics. Our explanatory framework will be outlined in three parts:

1. First, we describe the numerous *explanatory variables* selected for our framework, where several of them reflect *unique characteristics* of each individual's voice.

2. Then, we introduce the *dependent variable* representing the performance of the systems. As previously mentioned, the performance was estimated as the rate of false acceptances to examine the *influence* of the explanatory variables on the *security* of the speaker recognition models, but we first processed the performance to be adapted to our explanatory model.

3. We present the formulation of the *explanatory model* that processes the distribution of the vocal characteristics of the speakers to estimate their impact on the systems performance.

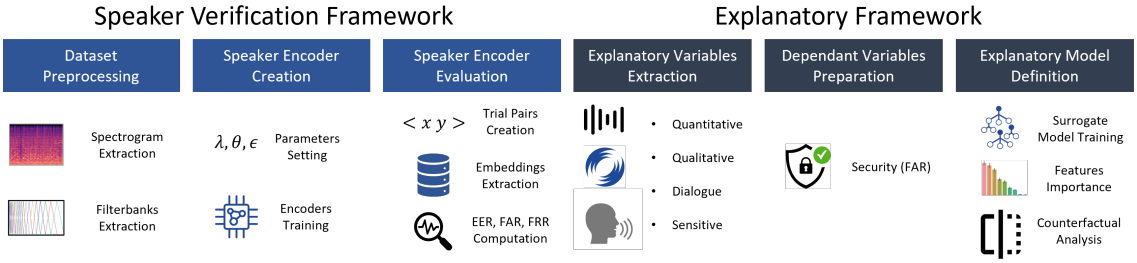


Figure 3.6: Once the dataset is pre-processed, we created two speaker encoders and ran an explanatory analysis to assess the impact of voice characteristics.

The experiments devised to analyze the impact of voice characteristics on the system security leverage *counterfactual reasoning*. As introduced in 2.3.3, counterfactual reasoning studies the hypothesis of a counterfactual world, where the change of one or more features could modify a certain event occurred in the real world. Grounded to the concept of *counterfactual fairness*, we select the speakers' sensitive attributes as the changing features to observe whether the models predictions would differ if in a counterfactual world a speaker belonged to a different demographic group, e.g., if a male was considered a female. An overview of our method is depicted in Figure 3.6.

3.4.1 Methodology

Explanatory Variables Extraction

The explanatory variables considered in our study include voice characteristics pertaining to a wide range of perspectives. Each speaker u was represented in terms of two main categories of characteristics: *protected* and *non-protected*. Non-protected characteristics, e.g., jitter, shimmer, were extracted from each speech waveform w belonging to speaker u and averaged across speech waveforms of that speaker u to obtain a vector x of size P, where P is the number of non-protected characteristics. As we will describe later, non-protected characteristics can be further divided into three sub-categories: *quantitative*, *qualitative*, and *dialogue*. Conversely, protected characteristics, e.g., gender and age sensitive attributes, were defined at the speaker level and represented with the vector z of size T, where T is the number of protected attributes. Formally, a speaker u was represented as a vector $c_u = [z; x] \in \mathbb{R}^{P+T}$, where $[;]$ is the concatenation operator.

Indeed, speech can be influenced by protected attributes, such as age and body conformation. Even though the availability of protected explanatory variables in corpora adopted for speaker recognition is limited, our study considered the following three protected attributes included into `FairVoice`:

- **Gender** of the speaker, self-reported by users, represented as a binary label (male, female).

- **Age Range** of the speaker, with the label "younger" assigned to those with age ≤ 40, "older" otherwise.

- **Language** spoken by the speaker (English, Spanish).

Non-protected quantitative variables measure properties common to any audio signal and do not have *any* direct *relation* with *personal speaker traits*. Specifically, we considered:

- **Root mean square (RMS)** is the loudness of the audio signal, measured as the power of the wave averaged across its length; a low-volume audio sample could negatively affect recognition performance.

- **Decibels relative to full scale (dBFS)** represents the loudness of the audio signal in decibel (dB) units, under a logarithmic scale, relative to the maximum possible loudness.

- **Maximum Amplitude** that is reached by the sound wave.

- **Intensity** (Mean, Std. Dev., Skewness, Kurtosis) is the power of the audio signal per unit area perpendicularly to that area, measured in dB SPL (Sound Pressure Level).

- **Signal-to-Noise Ratio (SNR)** measures the noise of the audio signal in dB, where a lower value reveals a high noise in the audio signal.

Non-protected qualitative variables include all those characteristics of a audio signal that differ depending on the source that generated it. The vocal folds that produce the human voice are an organic structure. Hence, the oscillations of the *voice* could contain *significant fluctuations*. Characteristics like fundamental frequency or jitter are affected by the context of the dialogue. Specifically, we considered:

- **Harmonics-to-Noise Ratio/Harmonicity (HNR)** (Mean, Std. Dev., Skewness, Kurtosis) represents the degree of acoustic periodicity, with high values for signals where most of the energy is in the periodic part. This measure is influenced by personal traits and medical conditions [14].

- **Fundamental Frequency F0** (Mean, Std. Dev., Skew, Kurtosis) of a speech signal refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. The sound wave is divided into several windows, and F0 is extracted for each one as the average number of oscillations per second and expressed in Hertz. This property depends on gender, age, overall body size, and cultural aspects [22].

- **Formants F1, F2, F3, F4 Frequencies** (Mean, Std. Dev., Skewness, Kurtosis) are the first four lowest resonant frequencies of the vocal tract [22]. There is a significant positive correlation between vocal tract length and body size (either height or weight), but also clear differences in male and female vocal tract morphology [56]. After data cleaning, F1 skewness, F3 kurtosis and F4 kurtosis were maintained.

- **Formant Position** is the average standardized formant value for the first n (we use $n = 4$) formants [50].

- **Jitter** is the variation in signal frequency caused by irregular vocal fold vibration, included in all natural speech. This measure is influenced by several factors, such as loudness, language, gender, and personal habits, e.g., smoking or alcohol consumption [22]. In our study, jitter is measured with the *local* variation of [14] implementation, defining it as the average absolute difference between consecutive periods divided by the average period.

- **Shimmer** is similar to jitter, but accounts for the variation in amplitude. This measure depends on personal traits. We adopted the *local dB* variation from [14], defined as the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

When considering sound waves containing human voices, aspects related to what a person is saying and how the speech is made can influence the performance of a speaker recognition system. Non-protected dialogue variables include properties related to *the way speech is generated* by a speaker, such as:

- **Number of Syllables** could impact the recognition task.

- **Number of Pauses** in a speech could be relevant for the speaker recognition system.

- **Rate of Speech** is the number of syllables pronounced along the entire audio sample, related to the propensity of the user to speak in a certain amount of time.

- **Articulation Rate**, differently from the rate of speech, is the number of syllables pronounced only over the speaking duration; so it describes how fast the user speaks.

- **Speaking Duration Without Pauses (SDWP)** counts the total duration in seconds of the portions of the audio example where the user is speaking.

Dependent Variables Preparation

The dependent variable is the performance of the speaker recognition model at the user level. Performance is estimated using the False Acceptance Rate (FAR) experienced by the user, to align our evaluation protocol with other works analyzing the impact of voice data manipulation [104, 105] focused more on security against imposture. As previously described and formally defined in Section 3.2.2, FAR is the measure of the likelihood that the biometric security system will incorrectly accept an attempt by an unauthorized user. For simplicity, with FAR as dependent variable, in our experiments we consider secure (label 1) a value of $FAR = 0$ and insecure (label 0) a value of $FAR > 0$.

Explanatory Model Creation

Explanatory modeling refers to the application of statistical models to data for testing causal hypotheses about theoretical constructs. In this study, on the basis of the components described in the previous sections, the causal hypothesis is: *can explanatory variables, capturing voice characteristics, explain the variation of the dependent variable related to speaker verification performance*? To this end, we introduce a surrogate model as the explanatory model, which, for a certain speaker recognition system, is optimized to explain the dependent variable (FAR) by means of the explanatory variables extracted from the speakers' utterances. To analyze the dependency of the performance on the explanatory variables, we considered random forests and linear models as a surrogate model, leaving the usage of other families of explanatory models for future works. Formally, an exploratory statistical model $\mathcal{G}$ for a speaker verification system $\mathcal{V}$ can then be defined as:

$$\mathcal{G}_\theta : f(c_u) = \hat{y}^{\mathcal{V}} \quad \text{s.t.} \qquad h(\mathcal{G}) = \Psi^{\mathcal{V}} \tag{3.12}$$

where $\hat{y}^{\mathcal{V}}$ is the prediction of $\mathcal{G}$, h is a function that from $\mathcal{G}$ returns the importance weights $\Psi^{\mathcal{V}} \in \mathbb{R}^{P+T}$, which are the hypothesized impactful parameters that vary in terms of the information captured from each characteristic in c_u. Training a surrogate model becomes then an optimization problem:

$$\underset{\theta}{\operatorname{argmin}} \, |f(c_u) - y^{\mathcal{V}}| \tag{3.13}$$

where $\theta \in \mathbb{R}^*$ is a set of parameters, i.e. rules used internally by $\mathcal{G}$ to be optimized. Surrogate models were optimized via a grid search on all the audio samples included in the testing set. Specifically, the vector c of explanatory variables characterizing each user was fed as input of the surrogate model. The dependent variable was considered as the ground truth value to predict. Since we are interested in the explanation power of the surrogate models, no further split of this set is performed.

Table 3.3: Performance of the considered speaker encoders under negative pairs created with another user from the same age range or the same gender (more challenging scenario).

	ResNet-34		**X-Vector**	
Negative pair type	*EER*	$FRR_{FAR1\%}$	*EER*	$FRR_{FAR1\%}$
Same age range	**0.08**	0.27	**0.08**	**0.2**
Same gender	0.11	0.43	0.11	0.3

3.4.2 Experimental Setup

Dataset

The dataset used to carry out the experiments on our explanatory model is FairVoice, the corpora previously described in Section 3.3.2. We embraced the same configuration idea adopted for the training and testing files of our previous study, but we applied some modifications to increase the significance of the experiments with our explanatory framework. We considered a *balanced multi-language training set* setting, where the speakers' representation was balanced by gender and age for both languages in the same training set. Specifically, we merged together the training sets UB of English and Spanish in one single training set, i.e. the training set includes 155 speakers for each of the eight demographic groups obtained by combining gender, age, and language, for a total of 1,240 speakers. We also performed a merge operation on the testing set, so including *100 English speakers* and *100 Spanish speakers* evenly distributed across gender and age groups. Differently from the previous experiments, for each speaker in this merged testing set we generated 55 trial verification pairs: 5 *positives* (other utterances from the same speaker) and 50 *negatives* (other utterances from another speaker) divided in 25 intra-age and 25 intra-gender verification pairs of the same language group (no pair includes an English and a Spanish speaker, nor vice versa).

Speaker Recognition Models

We focused on a subset of the models adopted to study the balancing strategy previously described. Specifically, among the three models *X-Vector*, *ResNet-34*, *ResNet-50*, our explanatory framework was applied to study the impact of the voice characteristics on the performance of the first two architectures. Given the objective of examining the extent to which voice features affect the models prediction process and their security, we selected speaker recognition systems with architectures substantially different. It follows that ResNet-50 was not considered due to the structure similarity with ResNet-34 and no relevant aspects differ between these two models, e.g., EER performance. We maintained the same hyper-parameters for this study, but the models were re-trained given that we applied a multi-language

setup, which was not contemplated in our previous study. We report in Table 3.3 the models performance in FRR at the FAR1% security level and in EER, which were measured according to the training set and the testing set introduced earlier, but still highlighting the nature of the trail verification pairs.

3.4.3 Results

Relationship between Explanatory Variables

In a first analysis, we investigated whether protected explanatory variables have any relationship with other speech covariates we considered as explanatory variables (quantitative, qualitative, and dialogue). To this end, Figure 3.7 shows the Pearson correlation among the considered explanatory variables. For conciseness, we present only the results for the Random Forest (RF) as a surrogate model, since it achieved a value close to 1 for both F1 score and AUC and can well explain the relationship of the dependent variable with the explanatory variables. We also played with linear models, but achieved both F1 Score and AUC lower than 0.65. For clarity, we also removed non-protected variables highly correlated with each other.

It can be observed that there was a high absolute correlation between gender and other voice characteristics, such as *F0 mean* and statistical moments measured on the distribution of the 4 formants (F1, F2, F3, F4). Similarly, *jitter local* and *shimmer local dB* had a positive correlation with gender as well, confirming that these characteristics are able to encode personal traits of each individual. Conversely, age range and language did not report any significant correlation with other speech covariates.

Influence of Speech Covariates on Performance

In a second analysis, we analyzed which speech covariates influence speaker recognition performance the most. In particular, we examined the dependency of the FAR (security) from the considered explanatory variables by means of the surrogate models included in our explanatory framework. Before training the surrogate models, a variance inflation factor [39] was applied to remove multi-collinearity among explanatory variables (threshold equal to 5.0), which resulted in removing a range of less influential explanatory variables. The remaining ones were used to train the surrogate models. In order to uncover the influence of explanatory variables on recognition performance, we leverage techniques of permutation feature importance on the surrogate models, applied over 10 repetitions to ensure statistical significance.

Figure 3.8 collects the explanatory variable importance scores on ResNet-34 and X-Vector, respectively. It can be observed that the RF surrogate model considered the formants F1, F3, and F4 as well as the fundamental frequency F0 to be the most important variables for both speaker encoders. None of the protected explanatory variables were considered as important by the surrogate model, except for language

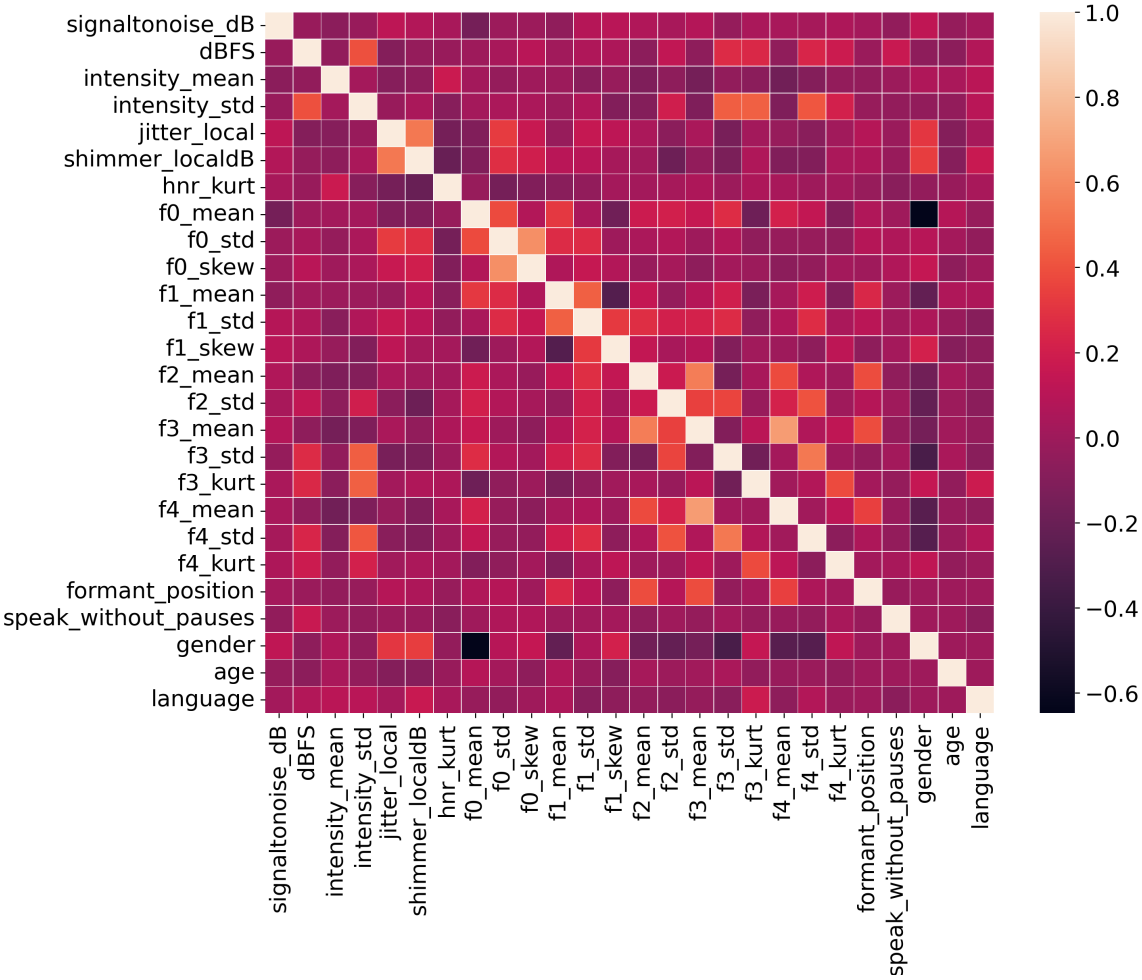
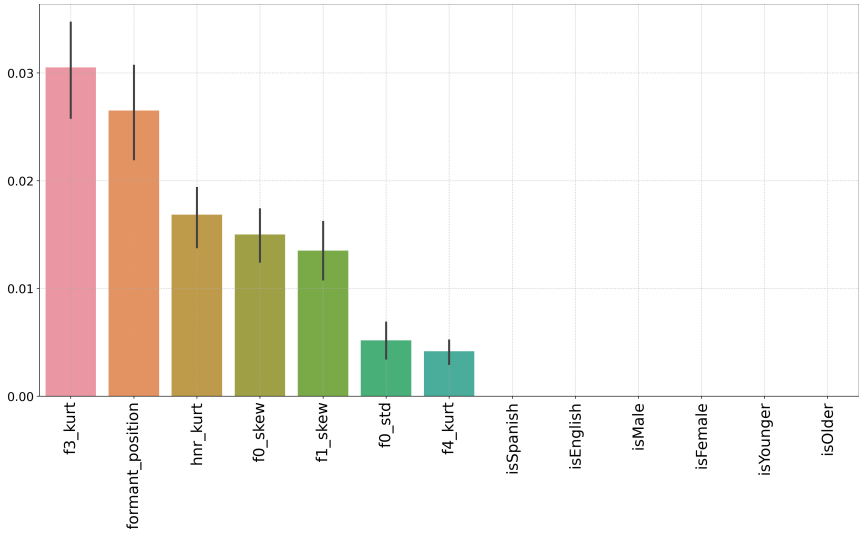


Figure 3.7: Pearson correlation between explanatory variables.

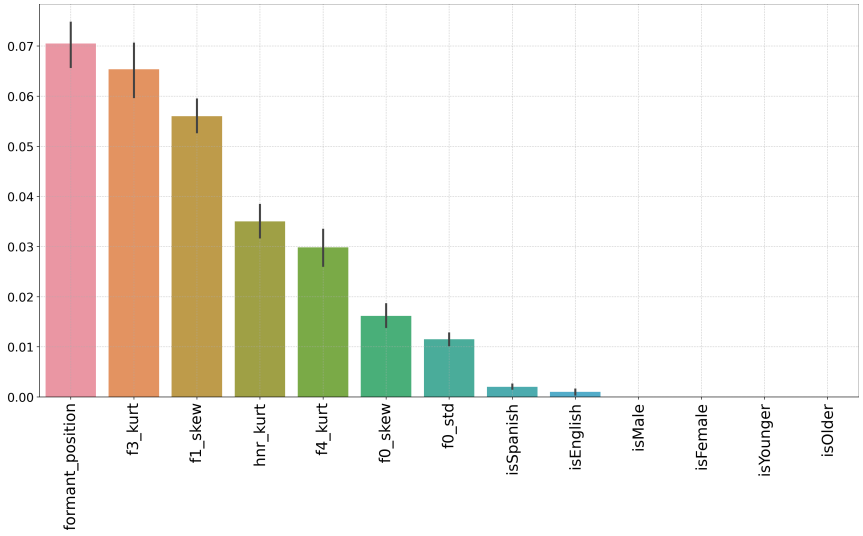
in X-Vector. Overall, our results reveal that protected attributes do not directly affect the performance of the speaker recognition system. Other speech covariates, despite still being correlated with protected attributes, can be used to interpret and then counteract unfairness in terms of security on both speaker encoders.

Impact of Protected Class Changes

Previous experiments revealed that there exists a relationship between protected attributes (especially gender) and other voice characteristics and that some key voice characteristics can explain the error rates experienced by a speaker recognition system (especially FAR) to a good extent. In our third and last analysis, we therefore leveraged the surrogate model to investigate what happens to the dependent variable when we flip the protected class of a user in his/her vector c, e.g., by modifying the gender (age; language) of a user from female (younger; English) to male (older; Spanish). Our goal is to compare the predictions of the surrogate model when the



(a) ResNet-34

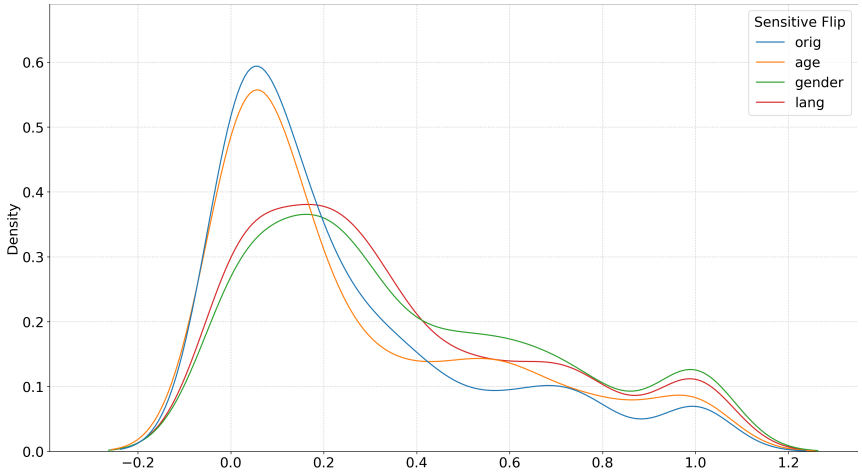


(b) X-Vector

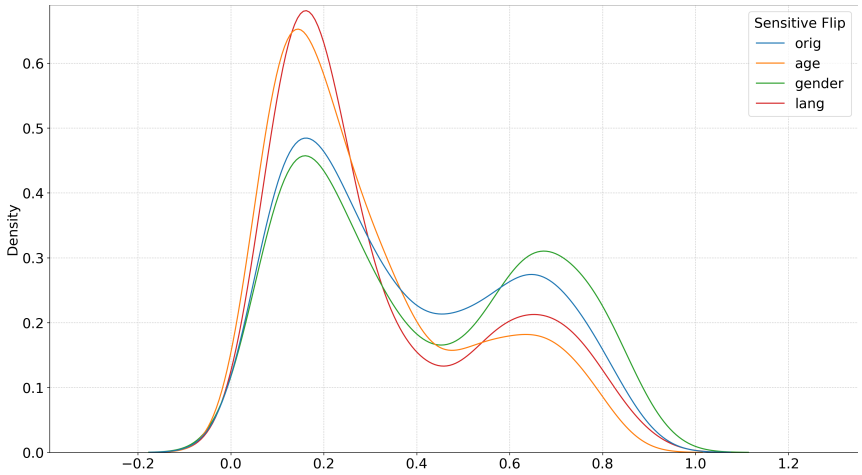
Figure 3.8: Permutation feature importance of voice characteristics on the predicted FAR over 10 repetitions.

original vector and the vector with the flipped protected attribute are fed, respectively. Through our surrogate model, we provided "what if" feedback of the form "if an input data point was c_u' instead of c_u, then a speaker encoder outcome would be $\hat{y}$', and not $\hat{y}$".

Figure 3.9 reports the predicted FAR for original vectors and vectors with a protected class flipped on ResNet-34 and X-Vector respectively. The *orig* curve depicts the density distribution of the predicted FAR when the original vector of each speaker was used. The other curves represent the predicted FAR when the



(a) ResNet-34



(b) X-Vector

Figure 3.9: Counterfactual analysis of the effect of flipping the protected class on predicted FAR.

corresponding protected attribute is "flipped" for each user, i.e. the curve labeled *gender* was generated by flipping the gender class of each speaker and similarly for *language* and *age* range. It can be observed that flipping the gender and language classes resulted in a significant increase of predicted FAR level on ResNet-34. Conversely, flipping the language and age classes positively affected FAR predictions on X-Vector. Hence, the sensitive attributes, especially the language, are able to modify the predictions of RF. This observation is also in line with what we observed in the first two analyses.

3.5 Findings and Discussion

This chapter outlined the contributions conceived by our research work in the speaker recognition field. Our studies raised awareness about fairness in machine learning applied for speaker recognition tasks through the analysis of methods to mitigate and explain unfairness experienced in such systems outcomes. We summarize the findings of our work in speaker recognition as follows.

- Speakers balancing techniques applied to the training data *do not affect the recognition performance* of the systems.
- The balancing strategy *can help in mitigating the disparity in error rates*, but it is not systematically effective.
- Under an *extended fairness evaluation*, *under-sampling* the speakers in the training set systematically *mitigates disparities across gender groups*, while the *mitigation across age groups is model-dependent*.
- Studying the impact of the balancing strategy along decision thresholds, we observed that *unfairness across age groups* is more prominent *near the FAR 1% threshold*, while experiments with the *gender-based groups* often show *unfairness* at thresholds *close to the EER* one.
- There are *significant relationships between gender and other speech covariates*. Age and language do not relate significantly with any covariates.
- Speech covariates pertaining to *vocal frequency aspects explain the most the disparate security estimates* across individuals.
- The spoken *language has the strongest impact on the security* of the two considered speaker recognition systems.

The proposed research opens to several future works. For instance, we can investigate on approaches able to better mitigate unfairness and on multi-class sensitive attributes beyond gender and age. Additionally, adversarial methods could be leveraged to obfuscate the sensitive attribute latent representation in the speaker encoders embeddings to guarantee unbiased predictions. The studies on our explanatory framework proved that the causes of disparate performances go beyond mere memberships to certain demographic groups, but they result from fine-grained voice characteristics (some of them related to the group membership). This opens a new perspective for analysis and mitigation of unfairness in speaker recognition where it might be no longer required to know the (hard to retrieve, especially due to privacy constraints) protected attribute labels. Other voice covariates emerged from our analysis can be used as a real proxy of such labels and as drivers for specific mitigation strategies, e.g., clustering users based on those characteristics and

provide treatments to the disadvantaged clusters. Another line of research can also focus on making input waveform statistically indistinguishable from the perspective of the relevant voice characteristics, for instance through the use of autoencoders, in order to make speaker encoders robust to these characteristics.

Chapter 4

Fairness in Recommendation

4.1 Introduction

With the large adoption of decision-support systems, humans' intervention has been increasingly supported by automated intelligence in various real world, high-stakes environments. A notable example of decision-support system is represented by recommender systems, where people are provided with personalized suggestions generated by a certain model (e.g., [8, 106]). Recommender systems filter the tremendous amount of products and services available in streaming platforms, e-commerce, social media, and so on, to help us make decisions, from selecting books to meeting new friends [128]. Their wide adoption has spurred investigations into possibly unfair practices in the systems' mechanisms [27, 46, 39, 101, 17]. Prior studies have shown that recommender systems often lead to discriminatory outcomes [41, 91, 116]. These phenomena can occur in the form of unfavorable outcomes, affecting the entity being ranked (*item* or *provider* unfairness) or the users the recommendations are targeted to (*user* or *consumer* unfairness) [21, 27, 4].

An abundance of consumer fairness notions have been consequently proposed, along with procedures for unfairness mitigation [21, 58, 92, 47, 81]. Despite the growing interest in providing fair recommendations to consumers, the landscape is convoluted, with often diverging definitions. The fragmented conception of consumer fairness has led to unfairness mitigation procedures built on top of heterogeneous evaluation protocols. *Why*, *when*, and *how* to apply a certain mitigation procedure, another, or all of them is still unclear. The current state of progress on consumer unfairness mitigation therefore calls for a discussion on what consumer fairness is and which properties a mitigation procedure against consumer unfairness should be evaluated on to let scientists select more consciously the procedure to apply according to the circumstances and conditions they experience.

Meanwhile, as recommender systems become more and more effective and sophisticated, the complexity of their functioning increases dramatically. The recommendations of novel systems improve the satisfaction of the users, but their *lack of interpretability* lays the groundwork for worrying questions [60]. The issue of in-

terpretability comes in addition with the prominent importance of preserving properties that go beyond recommendation effectiveness, such as *trustworthiness* [147], *fairness* [150], and *explainability* [170]. However, all these issues (from model interpretability to results that go beyond accuracy) are usually treated by the modern literature as independent perspectives, mostly tackled one at a time. Taking as an example algorithmic fairness (which is also the main case study in this thesis), while it is of uttermost importance to provide the end users and the content providers with equitable recommendations, it is also important for service providers (e.g., an online platform) to understand *why* the model behind their platform is unfair. Hence, tackling algorithmic fairness in an explainable way is a central yet under-explored area, contemplated only in a few works [59, 39].

This perspective does also apply to unfairness mitigation methods, since the existing ones have often relied on mathematical formulations of fairness principles, but have been rarely informed from explanatory analyses on such unfairness [21, 47, 92, 58]. Indeed, the few existing approaches that explain unfairness in recommender systems did not lead to a mitigation procedure that leverages the identified explanations to mitigate the measured unfairness [59, 39]. Moreover, their unfairness explanation methods exploit user and item features, which might be challenging to obtain, given that most recommendation models work with user-item interaction data, and to be used to mitigate the examined issue.

Shifting the focus to other areas within explainable artificial intelligence, various techniques have been employed to determine the relevant data entities that can serve as explanations for diverse tasks. Counterfactual methods have recently emerged as an effective way to explain the predictions produced by models based on Graph neural networks (GNNs) [67, 84, 143, 172], which have proven to be effective in modeling graph data in several domains, such as information retrieval [34], recommender systems [71, 165], natural language processing [164] and user profiling [29, 30, 162]. Approaches driven by counterfactual reasoning have also been used to guarantee algorithmic fairness in GNN-based models, for various downstream tasks, by manipulating the topological structure [6, 99, 146]. However, to the best of our knowledge, no approach was ever proposed to explain unfairness in GNN-based recommender systems or leverage explainability techniques to mitigate unfairness in such systems. Filling this research gap goes beyond a simple application of counterfactual explanations methods for GNNs, so as to uncover and subsequently mitigate unfairness in recommender systems. Indeed, the original methods that generate explanations to *explain the predictions* [97, 82, 32, 167] or to *mitigate the unfairness* [43] in GNN-based models are applied for classic tasks, e.g., classification, and classic graphs, while recommender systems are characterized by a bipartite nature, since they bridge the interactions between two types of entities (nodes), i.e. users and items.

In this chapter, we account for the issues previously highlighted and extend the literature of fairness in recommender systems by (i) providing a comprehensive overview of existing mitigation procedures through a systematic reproducibility

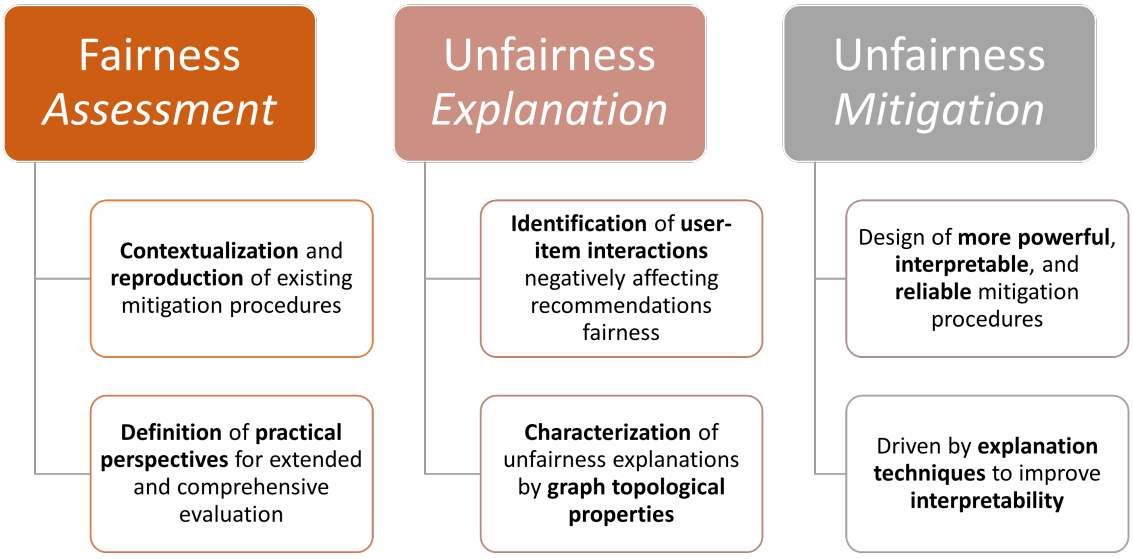


Figure 4.1: Overview of the research conducted to assess, explain, and mitigate unfairness issues in recommender systems.

study, (ii) identifying which properties a mitigation approach for consumer unfairness should meet to be effective and reliable, (iii) devising a framework to generate global explanations of unfairness across consumer groups in GNN-based recommender systems to explain, but also mitigate such issue, through user-item interactions and no additional features.

The contributions of this chapter is five-fold:

- we conducted a *systematic reproducibility study* on algorithmic procedures for mitigating consumer unfairness in rating prediction or top-k recommendation tasks, identifying *15 relevant papers* and reproducing *8 of them*.
- we defined a *common evaluation protocol*, including two public datasets, two sensitive attributes and two fairness notions to assess the fairness level of the recommendation models reported in the reproduced papers, with and without the proposed mitigation procedure.
- we identified a set of *eight technical properties* a given mitigation procedure against consumer unfairness should meet for being effective in practice, and assessed the extent to which existing mitigation procedures against consumer unfairness meet the defined properties, qualitatively and quantitatively (when possible).

- we formulated the problem of *explaining unfairness in GNN-based recommender systems* and proposed a framework to generate global counterfactual explanations of the unfairness they propagate.

- we leveraged these global counterfactual explanations to *augment the bipartite graph* used during the GNN-based model inference step, such that the altered recommended lists could be fairer across demographic groups.

The rest of this chapter details the listed contributions as follows: Section 4.2 formulates the recommendation task adopted by recommendation systems in general and by systems based on GNNs, Section 4.3 contextualizes existing unfairness mitigation procedures under a common evaluation protocol and assesses the level they meet eight practical properties, which we designed ad-hoc for consumer unfairness evaluation, Section 4.4 shows an approach that identifies a set of interactions between users and items as an explanation of the unfairness across consumer groups in GNN-based recommender systems, Section 4.5 describes a data augmentation method that finds new user-item interactions to generate a new graph with the goal of mitigating unfairness in recommendation utility across demographic groups at the GNN model inference-level. Finally, Section 4.6 summarizes the findings resulting from the outcomes observations of our multifaceted experimental analysis and illustrates future works to extend the research in consumer unfairness in recommendation.

4.2 Problem Formulation

4.2.1 Model-based Recommendation Task

In recommendation, the goal of the preference model is typically predicting whether or to what extent an (unseen) item would potentially be of interest for a user. In a common scenario, the model uses past interactions between two main entities, namely users U and items I, to learn preference patterns. Each user $u \in U$ is assumed to have interacted with a certain item $i \in I$ in case they rated, liked, or clicked on such item, depending on the applicative scenario. The set of items I_u a user interacted with is referred to as the u's history.

A main categorization in recommendation is the type of feedback given by a user to an item, which can be *implicit* or *explicit*. Specifically, an *implicit* feedback implies a user u interacted with an item i without information on the interest level for such item, e.g., watching a movie, listening to a song. On the other hand, an *explicit* feedback given by u to i denotes the extent to which i is of interest to u, e.g., rating a movie, putting "like" on a song. The feedback of all the users and items can be gathered to generate a feedback matrix $R \in \mathbb{R}^{|U| \times |I|}$, where $R_{u,i} \neq 0$ if a user u interacted with an item i, otherwise $R_{u,i} = 0$. The goal of a recommender

system is to predict the relevance of the missing entries in R, i.e. the entries that are equal to 0. A general recommender system can then be defined as:

$$f(R; W) \rightarrow \hat{R} \in \mathbb{R}^{|U| \times |I|} \tag{4.1}$$

Thus, f is parameterized by the weight matrix W and predicts a relevance matrix $\hat{R}$ to fill the missing entries in R.

Additionally, we can distinguish among three main sub-tasks of recommendation: *top-k recommendation*, *next-item recommendation*, *rating prediction*. One does not negate the other, since the first two could be performed on the basis of a rating prediction task, but they are typically treated separately.

- *Top-k recommendation* typically exploits an implicit feedback matrix, where $R_{u,i} = 1$ denotes an interaction between a user u and an item i. The goal in top-k recommendation is to suggest a list of k items based on the predicted relevance. Given a user u, items in I are sorted based on their decreasing relevance in $\hat{R}_u$, and the top-k items are recommended to user u.

- *Next-item recommendation* typically exploits an implicit feedback matrix as well as the top-k recommendation one. The goal is to learn the list-wise and sequence-aware preference of each user, so as to predict the next-item a user would prefer to interact with, e.g., a product to add to an online cart. While it can be perceived as a top-1 recommendation based on the item with the respective highest relevance, next-item recommendation is performed by models that account for the sequence order of the users' interactions to learn which item would better fit the next entry in the sequence.

- *Rating prediction* exploits an explicit feedback matrix, where $R_{u,i} \neq 0$ is equal to a numeric value on a rating scale, e.g., between 1 and 5, where 1 denotes the user u did not minimally like item i, whereas 5 denotes i fully satisfied u. The goal is to predict the extent to which each item is of interest to each user to fill the missing entries in R. In this case, the predicted relevance matrix $\hat{R}$ would be better denoted as a predicted rating matrix, because it includes the interest levels that a model predicts they would have been given by the users if they actually interacted with those items. As mentioned earlier, top-k and next-item recommendation could be performed on the basis of the predicted ratings.

4.2.2 Graph-based Recommendation Task

Graphs are structures that represent a set of entities (nodes) and their relations (edges). GNNs operate on graphs to produce representations that can be used in downstream tasks. In our case, user-item interactions can be represented by means of an undirected bipartite graph $\mathcal{G} = (U, I, E)$, where E is the set of edges representing

the interactions and $U \cup I$, with $n = |U| + |I|$, is the set of nodes. No edge exists between nodes of the same type, i.e. $E = \{(u, i) \mid u \in U, i \in I\}$ highlighting that E only contains edges between users and items, and not among users or among items. GNNs can then solve the recommendation problem by treating it as a linking prediction task, in order to predict potentially interesting links between users U and items I in the bipartite graph $\mathcal{G}$.

The graph $\mathcal{G}$ is a different way to represent the feedback matrix R. Given that a graph is typically represented by a $n \times n$ adjacency matrix A, we use A to denote the feedback matrix R. A GNN-based recommender system can then similarly be defined as:

$$f(A; W) \rightarrow \hat{R} \in \mathbb{R}^{|U| \times |I|} \tag{4.2}$$

In this case, the relevance $\hat{R}_{u,i}$ is better denoted as the linking probability between user u and item i. In a common scenario, the adjacency matrix A is normalized on the basis of the degree matrix. Therefore, f predicts the user-item relevance matrix $\hat{R}$ by combining the normalized adjacency matrix $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where $D_{j,j} = \sum_k A_{j,k}$ are entries in the degree matrix D, with the learned weight W according to the GNN implementation. Analogously to the general recommender system, given the matrix $\hat{R}$ and a user u, items in I are sorted based on their decreasing linking probability in $\hat{R}_u$, and the top-k items are recommended to user u. Consequently, the list of items recommended to user u is referred to as Q_u as and the set of all recommended lists as Q.

4.2.3 Task Optimization Targets

Metrics used to evaluate recommender systems are particularly different from the ones used for other tasks, e.g., classification. Indeed, top-k recommendation task metrics typically account for the length of the list Q_u recommended to each user, and they could also depend on the order the items are presented in Q_u. Additionally, in top-k recommendation it is more common to deal with the *utility* of a recommendation list instead of its accuracy, given that we evaluate the extent to which a preference model recommends items that are useful and of interest for a user. Focusing on *group fairness*, we describe several metrics devised to assess the unfairness of a recommender system, usually based on the disparity in utility across demographic groups or on other perspectives, e.g., independence from the sensitive attribute. Given $T = \{r_{u,i} \mid r_{u,i} \in R \wedge r_{u,i} \neq 0\}$ the number of nonzero entries in the feedback matrix R, i.e. the number of user-item interactions in a dataset, we present some of the most used metrics in recommendation to measure the performance in different tasks and to measure the fairness.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is generally used in statistics to measure the errors between paired observations. Referring to recommendation, in particular the rating

prediction task, MAE is used to measure the error in predicting the real rating given by a user u to an item i. Formally:

$$MAE(R, \hat{R}) = \frac{1}{|T|} \sum_{j=1}^{|T|} |R_{T_j} - \hat{R}_{T_j}| \tag{4.3}$$

In other words, MAE measures the absolute difference between R_{T_k}, which represents the real rating given by u to i as a data instance of the dataset, and $\hat{R}_{T_k}$, which is the rating predicted by a model.

Root-Mean Squared Error (RMSE)

Root-Mean Squared Error (RMSE) is used similarly as MAE, but they have a main difference. In the RMSE formula the errors are squared before they are averaged, so the RMSE gives a relatively high weight to large errors, while MAE is a linear score, which means that all the individual differences are weighted equally in the average. Formally:

$$RMSE(R, \hat{R}) = \sqrt{\frac{\sum_{k=1}^{|T|} (R_{T_j} - \hat{R}_{T_j})^2}{|T|}} \tag{4.4}$$

F1 Score (F1)

F1 Score (F1) is a metric used to measure the accuracy of a test, given by the harmonic mean of *precision* and *recall*. In top-k recommendation, precision and recall account for the size of the list recommended to the users, hence, we deal with precision@k and recall@k, where the former is the fraction of relevant retrieved instances among the retrieved instances (i.e. k instances), and the latter is the fraction of relevant retrieved instances among all relevant instances. Hence, we define a list-aware F1 Score, denoted as F1@k, as follows:

$$F1(R, \hat{R})@k = 2\frac{precision@k \cdot recall@k}{precision@k + recall@k} \tag{4.5}$$

It should be noted that F1@k does not account for the position of the items in the list, i.e. the value of F1@k does not change if a relevant item is recommended at the top or at the bottom of the list.

Normalized Discounted Cumulative Gain (NDCG)

Normalized Discounted Cumulative Gain (NDCG) is a metric designed to capture the notion that relevant items should be ranked higher in the recommendation list and that the importance of an item decreases as its position goes down in the list. It considers the graded relevance of items, e.g., user ratings or relevance scores, and

calculates a discounted cumulative gain, denoted as DCG, based on the relevance and the position of each recommended item. DCG is then normalized by an ideal ranking that represents the perfect order of relevant items. For each user u and the corresponding recommendation list $Q_u = q$, NDCG can be formally defined as:

$$NDCG(R_u, \hat{R}_u)@k = \frac{DCG(R_u, \hat{R}_u)@k}{DCG(R_u, R_u)@k} \; ; \; DCG(R_u, \hat{R}_u)@k = \sum_{j=1}^{k} \frac{2^{r_{u,q_j}} - 1}{log_2(j+1)} \tag{4.6}$$

Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank (MRR) measures the effectiveness of a ranking algorithm by considering the rank of the first relevant item in the list of recommended items. Specifically, for each user, the reciprocal rank is calculated as the inverse of the rank of the first relevant item in the list. If no relevant item is present in the recommendation list, the reciprocal rank is set to 0. For each user u and the corresponding recommendation list $Q_u = q$, MRR can be formally defined as:

$$MRR(R_u, \hat{R}_u)@k = max(\{1/j \mid 1 \leq j \leq k, \; r_{u,q_j} \neq 0\}) \tag{4.7}$$

Disparity in Demographic Parity (DP)

Disparity in Demographic Parity (DP) estimates fairness based on the notion of demographic parity, which in prior works in top-k recommendation [155] it was operationalized as the disparity in recommendation utility across demographic groups. Let S be a metric that measures the recommendation utility (e.g., NDCG, MRR), G the set of demographic groups, and denoting the feedback sub-matrix and the predicted relevance sub-matrix with respect to a demographic group g as R^g and $\hat{R}^g$, respectively, we can formally define DP as:

$$DP(R, \hat{R})@k = \frac{1}{\binom{|G|}{2}} \sum_{1 \leq i < j \leq |G|} \left\| S(R^{g_i}, \hat{R}^{g_i})@k - S(R^{g_j}, \hat{R}^{g_j})@k \right\|_2^2 \tag{4.8}$$

Category Equity Score (CES)

Category Equity Score (CES) is the metrics proposed in [21] to measure whether the distribution of a desired item category is similar across demographic groups. The definition is not perfectly clear, since it is a consumer-side metric, but the paper mentions item categories as "protected". We studied this metric considering each item category, e.g., movie or song genre, as protected for each iteration and measuring CES for each one. We report the original definition with our notation:

$$CES(R, \hat{R})@k = \frac{\sum_{u \in U^{g_i}} \sum_{q \in Q_u@k} \gamma(q)/|U^{g_i}}{\sum_{u \in U^{g_j}} \sum_{q \in Q_u@k} \gamma(q)/|U^{g_j}} \tag{4.9}$$

where g_i and g_j are two different demographic groups of the set G, U^g denotes the users belonging to the demographic group g, $\gamma : q \to 0, 1$ is a function that maps to 1 if the recommended item is in a "protected" category.

ϵ-Fairness (EPS)

ϵ-Fairness (EPS) is a metric proposed in [58] that accounts for the preference distribution for any two items across demographic groups. In other words, a recommender system is said to be ϵ-fair if, for any two items, the proportion of users with the same preference is approximately identical in all the subpopulations of users defined by the same sensitive attribute. In their definition, the term "preference" is used exactly with respect to the considered two items to study if the users prefer an item over another with an approximately identical distribution across demographic groups. Formally, for any two items i and i':

$$EPS(R, \hat{R}) = \left| \frac{|\{u \mid u \in g \wedge \hat{R}_{u,i} > \hat{R}_{u,i'}\}|}{|\{u \mid u \in g\}|} - \frac{|\{u \mid u \in g_\neg \wedge \hat{R}_{u,i} > \hat{R}_{u,i'}\}|}{|\{u \mid u \in g_\neg\}|} \right| \leq \epsilon \tag{4.10}$$

where g and $g_\neg$ are two different demographic groups of the set G.

Bias Disparity (BD)

Bias Disparity (BD) is a metric proposed in [141] that studies the propagation of preference bias from the interaction data to the recommendation. In other words, BD estimates how different is the distribution of a certain item category in the lists recommended to the users in a demographic group with respect to the distribution of these users' interactions with the considered item category. Let $c \in C$ be an item category, I_c be the subset of items of category c and $g \in G$ a demographic group, BD can be formally defined as:

$$PR(R) = \frac{\sum_{u \in g} \sum_{i \in I_c} R_{u,i}}{\sum_{u \in g} \sum_{i \in I} R_{u,i}} \; ; \; B(R) = \frac{PR(R)}{P(c)} \; ; \; BD(R, \hat{R}) = \frac{B(\hat{R}) - B(R)}{B(R)} \tag{4.11}$$

where PR is the *preference ratio* of a demographic group g for the item category c, $P(c) = |I_c|/|I|$ is the probability of selecting uniformly at random an item of category c, B is the *preference bias* of a group g for the item category c with respect to its representation in the dataset, and $B(\hat{R})$ represents the preference bias measured on the recommended lists by first measuring the preference ratio on the latter.

Kolmogorov-Smirnov Test (KS)

Kolmogorov-Smirnov Test (KS) in its two-sample form can be used to test whether two one-dimensional probability distributions differ. It was used in [80] to test whether the distributions of the predicted ratings for two demographic groups differed, such that a smaller value of KS $\in [0 + \inf)$ indicates that the predicted ratings and sensitive attribute are more independent. In can be easily interpreted as the area between two empirical cumulative distributions of predicted ratings for two demographic groups and it can be formally defined as:

$$KS(\hat{R}) = \sup_x |F_{\hat{R},g}(x) - F_{\hat{R},g^\neg}(x)| \tag{4.12}$$

where sup is the supremum function, $F_{\hat{R},g}(x)$ and $F_{\hat{R},g^\neg}(x)$ are the empirical distribution functions of the predicted ratings for the demographic group $g \in G$ and $g^\neg \in G$, respectively.

Group Loss Variance (GLV)

Group Loss Variance (GLV) is a metric used in [124] to estimate the variance across demographic groups of the mean squared estimation error reported by a recommender system in a rating prediction task. Formally:

$$L_j = \frac{\left\| \hat{R}_T^{g_j} - R_T^{g_j} \right\|_2^2}{|T^{g_j}|} \; ; \; GLV(R, \hat{R}) = \frac{1}{|G|^2} \sum_{1 \leq j < k \leq |G|} (L_j - L_k)^2 \tag{4.13}$$

where $T^{g_j} = \{r_{u,i} \mid r_{u,i} \in R \wedge u \in g_j\}$ is the set of the real ratings given by the users in the j-th demographic group $g_j \in G$.

Generalized Entropy Index (GEI)

Generalized Entropy Index (GEI) is an inequality index used in [9] to measure the inequality of the *benefit* distribution over all the users or demographic groups. The concept of *benefit* is better described in the work [134] from which the metric is taken from, whose goal is to estimate how unequally the outcomes of an algorithm benefit different individuals or groups in a population. The definition of the benefit function b_j needs to be determined for a specific task. For a constant $\alpha \neq \{0, 1\}$, GEI can be formally defined as:

$$\bar{b} = \frac{1}{|T|} \sum_{j=1}^{|T|} b_j \; ; \; GEI = \frac{1}{|T|\alpha(\alpha - 1)} \sum_{j=1}^{|T|} \left[\left(\frac{b_j}{\bar{b}} \right)^\alpha - 1 \right] \tag{4.14}$$

Area Under the Curve (AUC)

Area Under the Curve (AUC) is a metric typically used in classification problems. AUC is derived from the *receiving operating characteristic curve* (ROC curve), which is the plot of true positive rates and false positive rates at various decision thresholds. Specifically, AUC is the area under the ROC curve, which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. We define this metric as a fairness tool leveraged in [93, 154] to estimate the accuracy of a predictor model with the goal of predicting the sensitive attribute of the users from their latent representation. For an unbiased predictor f, AUC can be defined as follows:

$$AUC(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbb{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|} \tag{4.15}$$

where $\mathcal{D}^0$ and $\mathcal{D}^1$ denote the set of negative and positive examples, respectively.

4.3 Techniques for Unfairness Assessment and Mitigation

This section describes our study aimed to shape recommender systems that account for consumer fairness. A common understanding and practical benchmarks on how and when each procedure can be used in comparison to the others is crucial. As a response, we conducted a systematic reproducibility study of algorithmic procedures for mitigating consumer unfairness in rating prediction or top-k recommendation tasks on the basis of a four-step pipeline:

1. A *paper collection process* was performed to gather recent studies proposing mitigation methods to counter consumer fairness, by scanning Information Retrieval scientific journals, as well as conferences and workshops proceedings with high impact.

2. We defined a *common evaluation protocol* to assess the unfairness level of the models adopted in the collected papers, with and without applying the corresponding mitigation procedures.

3. *Eight technical properties* were devised to increase the perspectives of consumer fairness evaluation in top-k recommendation, such that a given mitigation procedure against consumer unfairness would be reliable and effective in practice if it meets the proposed properties.

4. A *comprehensive assessment* of the mitigation level of the reproduced methods for consumer unfairness was carried out using our common evaluation protocol under both a rating prediction and top-k recommendation task. Finally, the

mitigation procedures devised for consumer unfairness in top-k recommendation were tested in terms of the tailor-made technical properties under our and original evaluation protocol.

Figure 4.2 provides a visual representation of our study based on the aforementioned four-step pipeline.

4.3.1 Methodology

Paper Collection Process

To collect existing mitigation procedures against consumer fairness, we systematically scanned the recent proceedings of top-tier Information Retrieval conferences and workshops, namely CIKM, ECIR, ECML-PKDD, FAccT, KDD, RecSys, SIGIR, WSDM, WWW, and journals edited by top-tier publishers, namely ACM, Elsevier, IEEE, and Springer. The keywords for our manual research were composed of a technical term, "*Recommender System*" or "*Recommendation*", and a non-technical term, "*Consumer Fairness*" or "*User Fairness*". We marked a paper to be relevant if (a) it focused on a personalized recommendation task, (b) it proposed a mitigation procedure, and (c) that procedure targeted the end users receiving the recommendations. Papers on other domains (e.g., non-personalized rankings), other stakeholders (e.g., providers only), and on pure conceptualization only (e.g., proposing a fairness notion without any mitigation) were excluded. Papers addressing both consumer and provider fairness were included, since they also target the end users. Finally, *15 relevant papers* were considered in our study.

We then attempted to reproduce the mitigation procedure proposed in each relevant paper, relying as much as possible on the source code provided by the authors themselves. We hence tried to obtain the source code for each relevant

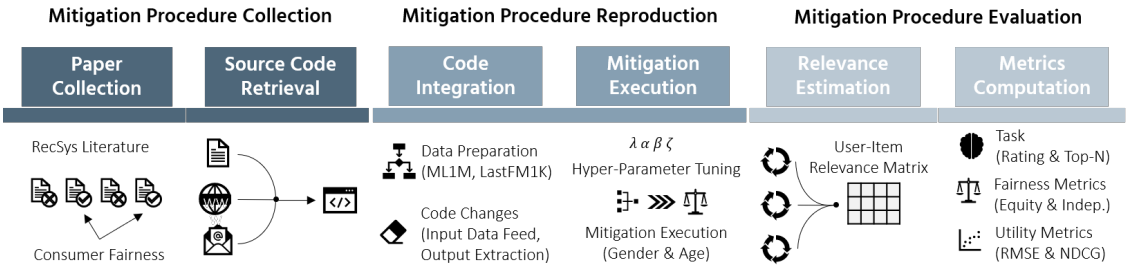


Figure 4.2: **Method**. We systematically collected papers and retrieved their source code. We processed the datasets used in our evaluation protocol, formatted them as per each mitigation requirements, and made the format of the mitigation results uniform. We trained the recommendation models included in the original papers, with/out mitigation, and computed fairness and utility metrics for the target recommendation task.

paper, by searching for the link into the paper, browsing for the official repository on the Web, and sending an e-mail to the authors as a last resort. We considered a mitigation procedure to be reproducible if a working version of the source code was obtained, and required minimal changes to accept another dataset and extract the final recommendations. Otherwise, we considered a paper to be non-reproducible given our reproduction approach. We also considered works to be non-reproducible when the source code was obtained but included only a skeleton version of the procedure with many parts and details missing. At the end, *8 out of 15 relevant papers* could be reproduced with a reasonable effort.

In Table 4.1, for each reproducible paper, we identified the recommendation task (RP : Rating Prediction; TR : Top-k Recommendation), the notion of consumer fairness (`EQ` : equity of the error/utility score across demographic groups; `IND` : independence of the predicted relevance scores or recommendations from the demographic group), the consumers' grouping (G : Gender, A : Age, O : Occupation, B : Behavioral), the mitigation type (PRE-, IN- or POST-Processing), the evaluation datasets (ML : MovieLens 1M or 10M, LFM : LastFM 1K or 360K, AM: Amazon, SS: Sushi, SY: Synthetic), the utility/accuracy metrics (NDCG; F1; AUC; MRR; RMSE; MAE), and fairness metrics (EPS; CHI; KS; GEI; TI; DP; CES; GLV). The reproducibility ratio was of 53% (8/15) in total: 50% (4/8) for top-k recommendation[1] and 57% (4/7) for rating prediction. We identified [154, 155, 93, 141] and [13, 74, 49] as non-reproducible procedures according to our criteria for top-k recommendation and rating prediction, respectively. Given that the most recent works addressing consumer fairness in recommendation focus on the top-k recommendation task, we also report in Table 4.1 the non-reproducible mitigation procedures devised for top-k recommendation task, denoted by the symbol (*). In light of this, specialized evaluation tools, particularly the technical properties mentioned at the bullet 3 in the Section 4.3 introduction, will be examined in the next sections only on papers focused on top-k recommendation, even for non-reproduced algorithms when possible.

Common Evaluation Protocol

To ensure evaluation consistency and uniformity across mitigation procedures, given the heterogeneity of the original experimental evaluations, we mixed replication and reproduction [1, 36]. For readability, we use the term "*reproducibility*". So, we used the source code provided by the original authors to run their models and mitigation procedures, and our own artifacts (data and source code) to (a) pre-process the input datasets as per their requirements and (b) compute evaluation metrics based on the relevance scores or recommendations they returned.

[1]During our studies, some authors published the source code of their papers for consumer fairness mitigation in top-k recommendation, increasing the reproducibility ratio from 50% (4/8) to 75% (6/8), but we could not reproduce them due to our work coming to an end.

Table 4.1: The considered reproducible mitigation procedures for consumer fairness. The works with the symbol (*) were not available or not reproducible at the time of this study.

Task	Paper	Year	Mitigation			Evaluation		
			Notion	Groups	Type	datasets	Utility Metrics	Fairness Metrics
TR	Burke et al. [21]	2018	EQ	G	IN	ML	NDCG	CES
	Tsintzou et al. [141]*	2019	EQ	G	POST	ML 1M-SY	-	BD
	Frisch et al. [58]	2021	IND	G-A	IN	ML	NDCG	EPS-CHI
	Li et al. (A) [92]	2021	EQ	B	POST	AM	NDCG-F1	DP
	Li et al. (B) [93]*	2021	IND	G-A-O-MS	IN	ML 1M-INS	NDCG-HIT	AUC
	Wu et al. (A) [154]*	2021	IND	G	IN	NEWS	AUC-NDCG-MRR	AUC-F1
	Wu et al. (B) [155]*	2021	EQ	G-A	IN	ML 1M-LFM 1K	RECALL-NDCG	DP
TR + RP	Ekstrand et al. [47]	2018	EQ	G	PRE	ML-LFM	NDCG-MRR	DP
RP	Kamishima et al. [81]	2018	IND	G-A	IN	ML-SS	MAE	KS
	Rastegarpanah et al. [124]	2019	EQ	B	POST	ML	RMSE	GLV
	Ashokan & Haas [9]	2021	EQ	G	POST	ML-SY	RMSE-MAE	GEI
	Wu et al. (C) [157]	2021	IND	G-A-O	IN	ML-LFM	RMSE	AUC-F1

[1] **Notion**: Equity (EQ), Independence (IND).
[2] **Groups**: Gender (G), Age (A), Occupation (O), Country (C), Marital Status (MS).
[3] **Type**: Pre-Processing (PRE), In-Processing (IN), Post-Processing (POST).
[4] **datasets**: MovieLens 1M (ML 1M), LastFM 1K (LFM 1K), Amazon (AM), Synthetics (SY), [153] (NEWS).

Datasets. The assessment of consumer fairness is challenging due to the lack of public datasets with ratings and *sensitive attributes* of the consumers. In our analysis, we considered all the public datasets that (a) were used in at least one reproduced paper, (b) reported at least one sensitive attribute, and (c) included enough ratings to reasonably train a recommender system ($\geq$ 200,000 ratings). We hence evaluated the reproduced mitigation procedures on two public datasets on the movies, MovieLens-1M (ML 1M), and music, Last.FM 1K (LFM 1K), domains (Table 4.2). Each dataset was downloaded from the original website and pre-processed according to our common evaluation protocol, in response also to some limitations of the reproduced mitigations. For instance, given that the existing mitigation procedures are often tailored to binary groups only, we grouped users in two groups in case of datasets with multi-class sensitive attributes (*while attributes like gender and age are by no means a binary construct, what we are considering is a binary feature*).

Gender labels were already binary in ML 1M. We binarized age labels, organized in seven age ranges, such that the two groups included consecutive age ranges and had the most similar representation possible. For LFM 1K, we considered only users reporting both their gender and age and filtered those with wrong ages (≤ 0 or ≥ 125). Interactions of a user for the same artist were aggregated, using the number of plays of a user for an artist as a proxy of the rating, similarly done by [47]. We filtered users interacting with less than 20 artists (as in ML 1M), and ratings were log-normalized and scaled in $[1, 5]$. Gender labels were already binary. We binarized age labels (integer) with the same criteria used in ML 1M.

Protocol. Each reproduced paper applied the corresponding mitigation procedure to a set of state-of-the-art recommendation models, which was quite het-

Table 4.2: The datasets with consumer's sensitive attributes included in this study.

dataset	#Users	#Items	#Ratings	Sensitive Attributes
ML 1M [69]	6,040	3,952	1,000,209	Gender (M : 71.7%; F : 28.3%) Age ($< 35 : 56.6\%; \geq 35 : 43.4\%$)
LFM 1K [25]	268	51,609	200,586	Gender (M : 57.8%; F : 42.2%) Age ($< 25 : 57.8\%; \geq 25 : 42.2\%$)

erogeneous across papers due to authors' arbitrary choices or the focus on a specific type of model. These models covered several families, including memory (`ItemKNN` [47, 9], `UserKNN` [47]), matrix factorization (`BiasedMF` [92, 9], `PMF` [92, 81, 157], `FunkSVD` [47]), learning-to-rank (`NCF` [92], `LBM` [58], `SLIM-U` [21], `ALS` [124], `LMaFit` [124]), graph (`GCN` [157]), and session-based (`STAMP` [92]). In line with our reproduction approach, we applied a given mitigation on the same models considered by the original authors[2].

Specifically, given a dataset, a sensitive attribute, and a reproducible paper, we considered the following evaluation protocol. We first performed a train-test split per user, with 20% of the interactions (the most recent if a timestamp was available, randomly selected otherwise) being in the test set and the remaining interactions being in the train set. In case a validation set was needed for best model selection, 10% of interactions (selected in the same way) of each user from the train set were considered as a validation set and the other ones included in the final train set. To fit with the original source code, the format of the considered sets and the sensitive attribute's labels per user were adapted. No changes on the source code specific for the mitigation procedure were applied.

Using the prepared sets and an appropriate hyper-parameters grid, we ran a grid search for each recommendation model, with and without mitigation. For each paper, our source code includes the scripts to format a dataset as per the original source code requirements and to compute evaluation metrics as well as the details of models hyper-parameter tuning. For each setup, we obtained the predicted relevance scores and the recommendations, and computed utility and fairness metrics. Utility metrics included NDCG for top-k recommendation (using binary relevances) and RMSE for rating prediction, selected due to their popularity (see Table 4.1). Consumer fairness metrics monitored *equity* through Disparity in Demographic Parity (DP), computed as the difference on utility for the corresponding task between groups, and *independence* through Kolmogorov-Smirnov (KS), computed on predicted relevance scores, covering two well-known perspectives and steps of the pipeline. We left analyses on other fairness notions and implementations of the same fairness notions as a future work.

[2]Though some procedures might be applied across models, their transfer often requires arbitrary design choices and core changes that mine our rigorous reproduction.

Practical Perspectives of Technical Properties

In this section, we propose eight key properties to consider while evaluating a mitigation procedure offline, before moving it into practice (e.g., user studies or online experiments). Being recommender systems often powered by machine learning, these properties would be indeed applicable also to mitigation procedures for any machine-learning model, from a conceptual perspective. This generality would indeed allow to contextualize mitigation procedures for recommender systems with respect to those for other machine-learning models in the future, giving a high-level overview on unfairness mitigation across machine-learning applications. It is worth noticing that the specificity of these properties would be represented by the way they were operationalized and monitored practically in the recommendation scenario (experimental protocols and results reporting) in our study. For instance, the property concerning data robustness was operationalized by considering data imbalances pertaining to the popularity of items according to the user's interactions with them, which is peculiar to a recommendation scenario. The design process for these properties was based on the adopted practices in the current academic literature about mitigation procedures for recommender systems (when possible) or those for general machine-learning models for completeness.

Applicability. A notable property a mitigation procedure would need to be evaluated on to support scientists' in their selection process is represented by the extent to which it can be technically applied to many recommendation models. As in the more general machine-learning field, mitigation procedures can be applied in pre-processing, by transforming the input data, in-processing, by constraining the training process of state-of-the-art models, and post-processing, by re-arranging the original predictions. Post-processing mitigation procedures can be applied to a wider range of models, being model-agnostic per definition. Although an in-processing approach could be extremely effective, it could be limitedly applicable to another model, especially from a different family. For instance, a mitigation procedure involving knowledge about users' neighbors might be seamlessly applied to different neighborhood-based recommendation models, but not to deep-learning models not relying on such neighborhood concept. On the other hand, an upsampling strategy can be used to manipulate the training data to be fed to any model, and a re-ranking procedure would be able to act on the original predictions provided by any model. One specificity of this property with respect to the recommendation task is that even pre-processing procedures might not always be model-agnostic, since for instance an in-processing tailored for a pair-wise model would need to act on pairs of items (not single items). Similar considerations can be made for post-processing approaches, which might rely on assumptions about the original rankings (e.g., in knowledge-aware recommendation) a model was able to produce.

Property. *Applicability indicates the extent to which a mitigation procedure can be technically run on a wide range of different recommendation models without requiring any substantial change to the fundamental steps it is based on.*

Coherence. Recent literature, in both the machine-learning field and the recommendation field more specifically, has proven that models can result in unfair outcomes for the users belonging to a certain demographic group. In some cases, mitigating unfairness for that demographic group might result in reversing the disparity towards the other group(s) (instead of merely reducing it for the originally disadvantaged groups). For instance, [17] showed that applying upsampling techniques to the training data and regularizing the training process, along with their combination, often leads to introducing disparities for the group the recommender system was originally performing the best on. A mitigation procedure would need to (at least) reduce the unfairness towards the originally disadvantaged group, without introducing any unfairness towards another group.

Property. *Coherence indicates the extent to which a mitigation procedure tends to reduce the biased outcomes for the originally disadvantaged group, without reversing the disparate outcome towards the other group(s).*

Consistency. One of the primary properties a mitigation procedure should effectively satisfy is being able to reduce the unfairness estimate measured on the original recommendation model, according to the targeted fairness notion. Notably, this property is the most popular one among those implicitly considered so far in the literature. This observation is expected, since it represents the property that motivates the design and development of a mitigation procedure in general. Being able to reduce unfairness only under a specific dataset and for a specific set of demographic attributes would not be enough, however, for a mitigation procedure to be considered consistent. Unfairness mitigation should ideally be possible regardless of the dataset and the demographic attribute.

Property. *Consistency indicates the ability of a mitigation procedure to substantially reduce the model's unfairness according to the pursued fairness notion, given any dataset and any consumer grouping method.*

Data Robustness. User-item interaction datasets collected from online platforms are usually characterized by imbalances with respect to the collected feedback (e.g., due to the popularity bias recommender systems often emphasize) and the stakeholders influenced by the provided recommendations (e.g., under-represented provider or consumer groups). Furthermore, depending on the domain, there might be features having a causal relationship with consumer unfairness. For being practically successful, mitigation procedures should be able to deal with data characterized by

(even extremely) imbalances and evident causal relationships. Inspecting robustness with respect to the data would be especially important when mitigation procedures do not report good results. Indeed, given a dataset and a mitigation procedure which shows a limited consistency, it could be better to analyze the underlying data, explain the causes behind unfairness from a data perspective, and inform the mitigation procedure about them.

Property. *Data robustness indicates the ability of a mitigation procedure to reduce unfairness also in challenging cases related to data distribution (e.g., imbalances) and relationships between unfairness and other features.*

Reproducibility. The source code of an approach, by producing results that are then analyzed and interpreted, allows to elaborate scientific conclusions. This imposes specific constraints on the code that are often overlooked in practice. Being able to make a mitigation procedure reproducible, by sharing for instance the original source code, would allow the community to test it even under different conditions (e.g., datasets or data splits). This sheds light on *reproducibility*, a key property to meet for progressing on unfairness mitigation. Other than enabling other researchers to assess the work's quality in more detail, sharing the source code would be important to increase the visibility of a study and further prove its potential advancement from the community.

Property. *Reproducibility indicates the ability of taking the original source code that implements a mitigation procedure and being able to execute it under the same or a different evaluation protocol, with respect to the one used in the original paper.*

Scalability. The increasing number of users and items in online platforms that leverage recommender systems is demanding a high amount of computational resources in terms of memory and execution time. Requirements pertaining to these resources should not be ignored, especially when a mitigation procedure against consumer unfairness should be run within or on top of the original recommendation model. Existing mitigation procedures were originally evaluated on barely small datasets, questioning the extent to which such procedures could be applied in real-world online platforms that involve thousands or even millions of users, items, interactions, and so on.

Property. *Scalability indicates the ability of a mitigation procedure to scale well when the number of interactions, users, items, and sensitive attributes, and other relevant features increases consistently.*

Trade-off Management. In the fair machine learning literature, it has been often highlighted a trade-off between accuracy/utility and fairness. This phenomenon particularly characterized the recommendation domain as well, e.g., [17]. A notable observation coming from these prior studies is that such trade-off often appeared for the best performing recommendation models, while was less evident for the ones reporting lower utility. This would be expected since less accurate models tend to provide random predictions more frequently. Inspecting the trade-off between utility and fairness in the context of a mitigation procedure for consumer unfairness would therefore be more relevant in case the original model already achieved a good performance. In the same way, while applying a mitigation procedure on a high-performing model, in an ideal case, we expect that the mitigation effectively reduces unfairness while preserving as much as possible the original model's performance.

Property. *Trade-off management indicates the ability of a mitigation procedure to preserve the performance estimate achieved by the target recommendation model originally (before the mitigation was applied).*

Transferability. In recent years, a large number of studies addressed consumer unfairness in recommender systems. However, these studies often examined the impact of the respective mitigation procedure on a restricted set of recommendation models or, in extreme cases, only on a single one. It therefore remains unclear the extent to which a mitigation procedure, originally evaluated and proven to be effective on a range of recommendation models, transfers well and it is equally effective (and so reduces unfairness) on another recommendation model. From a conceptual perspective, *transferability* is related to *consistency* (being effective across datasets and users' groups) and *applicability* (being applicable to a recommendation model). The main difference is that, with *transferability*, we aim to highlight the fact that, even if a mitigation procedure has been proven to be effective on a model, there are no guarantees that it will be equally effective on other models it could be applied to. Being *transferable* implies, for a mitigation procedure, that the consistency is preserved regardless of the model.

Property. *Transferability indicates the ability of a mitigation procedure to be effective (and not only applicable) on a wide range of recommendations models, even those it was not originally designed for or tested on.*

4.3.2 Results

Equity and Independence Assessment in Rating Prediction and Top-k Recommendation

Our first experiments focus on both the rating prediction and top-k recommendation task, by analyzing the extent to which the mitigation procedures impact on recommendation utility and unfairness. To this end, we report recommendation utility

and fairness scores obtained under the above evaluation protocol, for Top-k Recommendation (TR) across gender groups in Table 4.3 and across age groups in Table 4.4, and for Rating Prediction (RP) across gender groups in Table 4.5 and across age groups in Table 4.6. DP was tested for statistical significance via a Mann-Whitney test. For KS, we used its own score. The symbols (*) and ($\wedge$) meant significance at p-values 0.05 and 0.01, respectively.

Impact on Recommendation Utility. In a first analysis, we assess the impact of mitigation procedures on recommendation utility, focusing on the NDCG/RMSE columns provided in the aforementioned tables.

In a TR task, we observed that the NDCG achieved by the untreated models (Base) in ML 1M was in the range $[0.110, 0.140]$, except for SLIM-U, FunkSVD, LBM, and STAMP, whose NDCG was lower (≤ 0.084). Mitigating unfairness (Mit) in ML 1M did not generally result in a substantial change in utility (± 0.006 gender; ± 0.003 age). Higher changes were observed in two cases: SLIM-U treated with Burke et al.'s mitigation (stable for gender; -0.036 age) and LBM treated with Frisch et al.'s (-0.023 gender; stable for age). In LFM 1K, the untreated models (Base) got an NDCG in $[0.204, 0.406]$, overall higher than ML 1M. The models ranking based on NDCG differs for several models from ML 1M. Though their utility was relatively high, PMF, FunkSVD, LBM, and STAMP were still under-performing in LFM 1K. The treated models (Mit) showed changes in NDCG (± 0.009 gender; ± 0.018 age) larger in magnitude than ML 1M. SLIM-U with Burke et al.'s mitigation (-0.019 gender; -0.113 age) and LBM with Frisch et al.'s mitigation ($+0.068$ gender; $+0.069$ age) led to higher changes in NDCG.

Considering an RP task, the untreated models (Base) achieved an RMSE in

Table 4.3: [Top-k recommendation (TR) - Consistency - *Gender* Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Base*) and after mitigating (*Mit*) for *gender* groups.

		ML 1M						LFM 1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Burke et al. [21]	SLIM-U	0.084	0.084	^-0.022	^-0.028	^-0.032	^-0.115	0.320	0.301	*-0.060	^-0.072	^-0.006	^-0.142
Frisch et al. [58]	LBM	0.044	0.021	^-0.006	^-0.004	^-0.013	^-0.025	0.144	0.212	*-0.035	*-0.058	^-0.120	^-0.126
Li et al. (A) [92]	BiasedMF	0.112	0.051	^-0.017	**^-0.001**	^-0.035	**^-0.006**	0.287	0.114	^-0.095	^-0.060	^-0.012	**^-0.001**
	NCF	0.117	0.057	^-0.016	**^-0.001**	^-0.022	**^-0.006**	0.250	0.138	*-0.073	-0.026	^-0.033	**^-0.001**
	PMF	0.119	0.056	*0.013	^-0.002	^-0.023	**^-0.006**	0.200	0.071	*-0.062	-0.027	^-0.010	**^-0.001**
	STAMP	0.022	0.020	***0.003**	^-0.003	**^-0.006**	**^-0.006**	0.160	0.113	-0.021	**0.002**	**^-0.001**	**^-0.001**
Ekstrand et al. [47]	FunkSVD	0.018	0.015	^-0.004	0.002	^-0.027	^-0.018	0.010	0.013	**-0.006**	-0.003	^-0.107	^-0.119
	ItemKNN	**0.140**	**0.134**	^-0.038	^-0.030	^-0.030	^-0.031	0.287	0.286	^-0.127	*-0.116	^-0.019	^-0.022
	UserKNN	0.137	0.131	^-0.031	^-0.024	^-0.074	^-0.052	**0.406**	**0.411**	^-0.110	^-0.106	^-0.067	^-0.067

Configurations that resulted in a statistically significant difference in NDCG (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "^" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

Table 4.4: [Top-k recommendation (TR) - Consistency - *Age* Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Base*) and after mitigating (*Mit*) for *age* groups.

		ML 1M						LFM 1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Burke et al. [21]	SLIM-U	0.084	0.048	ˆ-0.022	ˆ-0.014	ˆ-0.009	ˆ-0.095	0.320	0.207	-0.026	ˆ-0.145	ˆ-0.017	ˆ-0.082
Frisch et al. [58]	LBM	0.044	0.042	ˆ-0.005	ˆ-0.006	ˆ-0.021	ˆ-0.027	0.144	0.213	-0.011	-0.021	ˆ-0.125	ˆ-0.152
Li et al. (A) [92]	BiasedMF	0.112	0.051	ˆ-0.015	**ˆ-0.000**	ˆ-0.042	**ˆ-0.006**	0.287	0.111	*-0.079	0.010	ˆ-0.019	**ˆ-0.005**
	NCF	0.117	0.057	ˆ-0.018	ˆ-0.002	ˆ-0.029	**ˆ-0.006**	0.250	0.137	*-0.067	-0.015	ˆ-0.055	**ˆ-0.005**
	PMF	0.119	0.056	ˆ-0.020	ˆ-0.004	ˆ-0.027	**ˆ-0.006**	0.200	0.071	*-0.046	**0.004**	ˆ-0.008	**ˆ-0.005**
	STAMP	0.022	0.020	**0.000**	**ˆ-0.000**	**ˆ-0.006**	**ˆ-0.006**	0.160	0.113	-0.031	-0.008	**ˆ-0.005**	**ˆ-0.005**
Ekstrand et al. [47]	FunkSVD	0.018	0.016	ˆ-0.008	ˆ-0.006	ˆ-0.029	ˆ-0.021	0.010	0.016	**0.002**	**-0.004**	ˆ-0.054	ˆ-0.047
	ItemKNN	**0.140**	**0.138**	ˆ-0.027	ˆ-0.024	ˆ-0.029	ˆ-0.033	0.287	0.269	0.010	0.020	ˆ-0.133	ˆ-0.118
	UserKNN	0.137	0.137	ˆ-0.028	ˆ-0.023	ˆ-0.060	ˆ-0.051	**0.406**	**0.397**	-0.023	-0.031	ˆ-0.036	ˆ-0.031

Configurations that resulted in a statistically significant difference in NDCG (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "ˆ" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

Table 4.5: [Rating Prediction (RP) - *Gender* Groups] Rating prediction error (*RMSE*, the lower it is, the more accurate the rating prediction), equity (RMSE Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Base*) and after mitigating (*Mit*) for *gender* groups.

		ML 1M						LFM 1K					
		RMSE ↓		DP $\downarrow_0$		KS ↓		RMSE ↓		DP $\downarrow_0$		KS ↓	
Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Ekstrand et al. [47]	FunkSVD	0.881	0.894	ˆ-0.032	**-0.023**	ˆ-0.052	ˆ-0.051	1.255	1.268	*0.039	0.039	ˆ-0.040	ˆ-0.052
	ItemKNN	0.865	0.882	ˆ-0.034	*-0.026	ˆ-0.055	ˆ-0.056	1.218	1.230	*0.037	*0.035	ˆ-0.064	ˆ-0.072
	UserKNN	0.896	0.911	ˆ-0.035	-0.025	ˆ-0.056	ˆ-0.058	1.226	1.239	ˆ-0.047	*0.054	**ˆ-0.036**	ˆ-0.045
Kamishima et al. [81]	PMF BDist	**0.863**	0.870	**ˆ-0.029**	ˆ-0.046	ˆ-0.056	**ˆ-0.032**	1.172	1.179	**0.014**	*0.029	ˆ-0.067	ˆ-0.029
	PMF Mean	**0.863**	0.870	**ˆ-0.029**	ˆ-0.048	ˆ-0.056	ˆ-0.056	1.172	1.179	**0.014**	*0.025	ˆ-0.067	ˆ-0.054
	PMF Mi	**0.863**	0.870	**ˆ-0.029**	ˆ-0.046	ˆ-0.056	**ˆ-0.032**	1.172	1.179	**0.014**	*0.029	ˆ-0.067	ˆ-0.029
Rastegarpanah et al. [124]	ALS	0.894	0.890	ˆ-0.034	ˆ-0.034	**ˆ-0.035**	ˆ-0.033	1.490	1.189	ˆ-0.145	0.029	**ˆ-0.036**	ˆ-0.114
Ashokan & Haas [9]	ALS Par	0.867	0.868	ˆ-0.030	ˆ-0.029	ˆ-0.056	ˆ-0.034	**1.145**	**1.146**	0.016	**0.018**	ˆ-0.047	***0.017**
	ALS Val	0.867	0.867	ˆ-0.030	ˆ-0.030	ˆ-0.056	ˆ-0.057	**1.145**	1.150	0.016	**0.018**	ˆ-0.047	ˆ-0.050
	ItemKNN Par	0.865	0.866	ˆ-0.034	ˆ-0.033	ˆ-0.055	ˆ-0.036	1.176	1.183	*0.033	*0.045	ˆ-0.061	ˆ-0.058
	ItemKNN Val	0.865	**0.865**	ˆ-0.034	ˆ-0.034	ˆ-0.055	ˆ-0.052	1.176	1.173	*0.033	*0.036	ˆ-0.061	ˆ-0.046
Wu et al. (C) [157]	FairGo GCN	0.895	0.892	ˆ-0.038	ˆ-0.034	ˆ-0.048	ˆ-0.045	1.609	1.283	ˆ-0.151	0.038	ˆ-0.113	ˆ-0.113

Configurations that resulted in a statistically significant difference in RMSE (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "ˆ" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

the range $[0.863, 0.905]$ in ML 1M. By mitigating (Mit) in ML 1M, no substantial changes were observed (± 0.017 gender; ± 0.013 age). In LFM 1K, the untreated models (Base) achieved a higher RMSE, in the range $[1.145, 1.255]$. ALS and GCN are the lowest performers (1.490 and 1.609, respectively). The treated models (Mit) showed minimal (± 0.0135 gender; ± 0.012 age) which are similar to the changes in ML 1M. ALS under Rastegarpanah et al.'s mitigation lowered RMSE (-0.301 gender; -0.305 age), as well as GCN under Wu et al. (C)'s mitigation (-0.326 gender; -0.332 age).

Table 4.6: [Rating Prediction (RP) - *Age* Groups] Rating prediction error (*RMSE*, the lower it is, the more accurate the rating prediction), equity (RMSE Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Base*) and after mitigating (*Mit*) for *age* groups.

		ML 1M						LFM 1K					
		RMSE ↓		DP $\downarrow_0$		KS ↓		RMSE ↓		DP $\downarrow_0$		KS ↓	
Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Ekstrand et al. [47]	FunkSVD	0.881	0.886	˙0.042	˙0.045	˙0.073	˙0.081	1.255	1.264	0.032	0.035	˙0.083	˙0.086
	ItemKNN	0.865	0.875	˙0.039	˙0.042	˙0.074	˙0.079	1.218	1.226	**0.019**	**0.028**	˙0.088	˙0.092
	UserKNN	0.896	0.902	˙0.047	˙0.050	˙0.092	˙0.103	1.226	1.233	0.034	0.031	˙0.087	˙0.095
Kamishima et al. [81]	PMF BDist	**0.863**	0.872	˙0.039	˙0.031	˙0.084	**˙0.018**	1.172	1.183	0.045	˙0.065	˙0.124	˙0.047
	PMF Mean	**0.863**	0.872	˙0.039	**˙0.027**	˙0.084	˙0.045	1.172	1.184	0.045	˙0.069	˙0.124	˙0.042
	PMF Mi	**0.863**	0.872	˙0.039	˙0.031	˙0.084	**˙0.018**	1.172	1.183	0.045	˙0.064	˙0.124	˙0.047
Rastegarpanah et al. [124]	ALS	0.894	0.892	**˙0.034**	˙0.040	**˙0.034**	˙0.037	1.490	1.185	0.033	*0.052	**˙0.017**	˙0.064
Ashokan & Haas [9]	ALS Par	0.867	0.871	˙0.041	˙0.048	˙0.074	˙0.026	**1.145**	**1.146**	0.043	*0.046	˙0.082	***0.015**
	ALS Val	0.867	0.866	˙0.041	˙0.042	˙0.074	˙0.079	**1.145**	1.149	0.043	*0.046	˙0.082	˙0.077
	ItemKNN Par	0.865	0.870	˙0.040	˙0.048	˙0.074	˙0.031	1.176	1.177	0.029	0.031	˙0.085	˙0.029
	ItemKNN Val	0.865	**0.864**	˙0.040	˙0.042	˙0.074	˙0.071	1.176	1.172	0.029	0.032	˙0.085	˙0.083
Wu et al. (C) [157]	FairGo GCN	0.895	0.908	˙0.040	˙0.044	˙0.070	˙0.074	1.609	1.277	0.043	*0.056	˙0.079	˙0.120

Configurations that resulted in a statistically significant difference in RMSE (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "˙" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

Impact on Group Unfairnes. In a second analysis, we investigated the impact of mitigation procedures on unfairness. For each table and dataset, we consider the DP and KS columns.

We start from a TR task, focusing our presentation on the subset of models that achieved a reasonable NDCG (≥ 0.110 for ML 1M; ≥ 204 for LFM 1K). In ML 1M, the DP and KS achieved by the untreated models (Base) laid in the ranges ($[0.013, 0.038]$ gender; $[0.015, 0038]$ age) and ($[0.022, 0.074]$ gender; $[0.027, 0.060]$ age), respectively. Without any mitigation, in terms of DP, BiasedMF, NCF, and PMF (≤ 0.017 gender; ≤ 0.020 age) were fairer than UserKNN, and ItemKNN (≤ 0.031 gender; ≥ 0.027 age). To some surprise, when KS was considered, we observed a different pattern. The fairest models in order were NCF and PMF (0.022 and 0.023 gender; 0.029 and 0.027 age), ItemKNN and BiasedMF (0.030 and 0.035 gender; 0.029 and 0.042 age), and UserKNN (0.074 gender; 0.060 age). By mitigating (Mit), DP went down to the range ($[-0.002, 0.030]$ gender; $[0.000, 0.034]$ age), while KS laid in the range ($[0.006, 0.052]$ gender; $[0.006, 0.051]$ age). In LFM 1K, models were less fair than in ML 1M. The untreated models (Base) achieved a DP in the ranges ($[-0.060, -0.127]$ gender; $[0.010, -0.079]$ age) and a KS in the ranges ($[0.001, 0.067]$ gender; $[0.017, 0.133]$ age). The models ranking in terms of DP and KS was similar between LFM 1K and ML 1M. Once mitigated (Mit), interestingly, we observed that re-sampling by Ekstrand et al. resulted in a decrease of fairness for ItemKNN and UserKNN in terms of DP on age groups (≥ 0.06). These findings are replicated for ItemKNN in terms of KS on gender groups (0.03), while, for age groups KS was substantially lowered (0.015). Other cases did not lead to substantial changes.

In a RP task, in ML 1M, untreated models (Base) achieved a DP in $[-0.038,$

$-0.025]$ (gender) and $[0.034, 0.051]$ (age), and a KS in $[0.035, 0.056]$ (gender) and $[0.034, 0.092]$ (age). With no mitigation, there were minimal differences in terms of DP between models for the attribute gender (avg. 0.033, std. dev. 0.003). For the attribute age, the untreated models had similar DP (avg. 0.041, std. dev. 0.005). Considering KS, comparable estimates across models were observed (avg. 0.053, std. dev. 0.003 gender; avg. 0.076, std. dev. 0.007 age). ALS (0.035 gender; 0.034 age) resulted in fairer outcomes in terms of KS. Treated models (Mit) showed stable fairness (± 0.010 gender; ± 0.008 age) in all cases, except for Kamishima et al. (± 0.019 gender; ± 0.012 age) when DP was considered. In terms of KS, models treated with Kamishima et al.'s mitigation (for gender only PMF BDist and PMF Mi) and Ashokan et al.'s mitigation (parity setting) were substantially fairer (≥ 0.019 gender; ≥ 0.039 age), while other treated models did not benefit from the mitigation (± 0.003 gender; ± 0.011 age). In LFM 1K, untreated models (Base) achieved a DP in $[0.014, 0.151]$ (gender) and $[0.019, 0.045]$ (age), and a KS in $[0.036, 0.113]$ (gender) and $[0.017, 0.124]$ (age). Without mitigating, findings in ML 1M held in LFM 1K, except for the high DP (0.151) and KS (0.113) of GCN for gender. Treated models (Mit) instead showed stable fairness (≤ 0.015 gender; ≤ 0.009 age) except for Kamishima et al. (≥ 0.019 age), ALS (0.116 gender; 0.019 age), GCN (0.113 gender; 0.013 age), in terms of DP (opposite to ML 1M). In terms of KS, except the mitigations of Kamishima et al. and Ashokan et al. (parity), treated models did not benefit from mitigation (≤ 0.015 gender; ≤ 0.005 age).

Top-k Recommendation Evaluation under Technical Properties

Although other recommendation tasks, e.g., rating prediction, exist in many commercial systems, the users are presented with a personalized ranking of items, but the predicted rating values are not. With a practical focus on personalization, we therefore evaluated our specialized technical properties only on consumer unfairness mitigation procedures adopted on recommendation models that aim to find a few specific items supposed to be most appealing for the user, based on their interests.

Applicability. Fairness is an objective that should be tackled regardless of the implementation of a recommendation model. Models used in the literature often belong to similar families, e.g., model- or memory-based, making it easier the development of mitigation procedures *applicable* to different models. The models' family is particularly important in the context of in-processing approaches, where aspects related to the implementation are relevant with respect to the range of models the mitigation is *applicable* to. Furthermore, the possible flexibility of an in-processing procedure could allow a broader range of models to benefit from it. This perspective can be envisioned for [21]'s work, where the authors added a regularization term to the loss function of SLIM, a hybrid model, to balance the neighborhoods. Such term can be also used on a simpler memory-based model, e.g., UserKNN, by regularizing the weights of the neighborhoods. In addition, in-processing proce-

dures like [58, 93, 155, 154] can be extended to different architectures by modifying some aspects that pertain to several families, e.g., multi-objective optimization and feature-independent user embeddings.

Other mitigation procedures are instead designed without considering in advance the recommendation model they will be applied to. Pre-processing approaches, such as data transformation or user representation balancing [47], are examples that potentially have a very high *applicability* for top-k recommendation. Conversely, the *applicability* of post-processing approaches [92, 141] could depend on other aspects related to the adopted fairness notion and how this notion needs to be satisfied. [141] proposed an approach to reduce the amplification of input data biases over specific categories of items by a recommender system; without datasets including category labels, this mitigation could not be applicable to a model.

Coherence. To investigate this phenomenon, we study the Tables 4.3-4.4, where the demographic parity score and the Kolmogorov-Smirnov test score are reported for the recommendation models originally considered in each paper, before and after applying the respective mitigation procedure. We were particularly interested in observing whether the sign of the fairness metric score changed between the two (reversed unfairness, so low coherence) or was the same (the same group was disadvantaged, so high coherence). The results show that some configurations reported a DP opposite in sign when a mitigation procedure was used, compared to the original recommendation model. A value of DP less than zero exhibits a bias towards the minority group, while the majority group is advantaged when DP is higher than 0.

A notable example that showed low coherence was SLIM-U when the mitigation of [21] was applied on gender groups of LFM 1K. When the original model was treated with the mitigation procedure, the outcomes advantaged male users instead of female users. This suggests that biases are not always caused by the characteristics of the dataset, and the way the model learns from the data is also relevant. In particular, due to the neighborhood balancing, the model learnt patterns on the basis of a different viewpoint on the data, which resulted in a higher recommendation utility for male users. Several experiments with [92]'s procedure were characterized by this phenomenon as well. On ML 1M, the mitigation procedure advantaged the minority group to a small extent when applied to NCF and PMF, while on LFM 1K the bias slightly shifted towards male users for STAMP and towards younger users for BiasedMF and PMF. A low coherence was also reported for FunkSVD while considering age groups. However, the bias towards the demographic groups was not substantial, neither for the original model (bias towards younger user) nor for the mitigated one (bias towards older users).

Coherence was not satisfied in ML 1M and in LFM 1K by two and five models out of ten respectively, which raises the question whether the bias in the outcomes can be reduced by just re-arranging the training data. The recommendation model and mitigation procedure should be thoroughly examined to understand the causes that

led to unfairness towards one or more demographic groups. We therefore encourage researchers to take into account this property, so as to devise mitigation procedures with a high coherence in future works addressing consumer unfairness.

Consistency. In order to carry out a comprehensive analysis of this property, we both consider our unified evaluation protocol as well as the original evaluation process in the corresponding paper. It would indeed be inconclusive to analyze the mitigation procedures consistency based solely on out protocol, given that the authors could have devised their algorithms to counter unfairness under a specific fairness notion and the evaluation should reflect the same viewpoint.

Unified Evaluation. Tables 4.3-4.4 report the performance of the recommender systems, before (`Base`) and after (`Mit`) unfairness was mitigated, in terms of recommendation utility (NDCG@10) and fairness (Demographic Parity and Kolmogorov-Smirnov test). We assumed that a mitigation procedure was consistent on a recommendation model when there was a significant decrease of unfairness compared to that of the original recommendation model.

Our results on ML 1M show that the mitigation procedures were able to improve DP for all the models, except for SLIM-U on gender groups. On the other hand, KS was decreased only by [92][3] and [47]'s mitigation procedures (except for ItemKNN, where we observed a minimal increment). In LFM 1K, unfairness was mitigated between gender groups in terms of DP by all the mitigation procedures, with the exception of LBM and SLIM-U. Only [92]'s method improved DP for age groups. KS was reduced by both [92] and [47]'s mitigation procedures, the former for both sensitive attributes, while the latter only for age groups. Overall, [92] was the only consistent mitigation procedure across datasets and sensitive attributes.

Original Evaluation. Even though Tables 4.3-4.4 present a complete overview of the *consistency* achieved by the considered mitigation procedures, NDCG DP and KS were often not monitored in the original papers. The two metrics we adopted monitor unfairness from two viewpoints not necessarily corresponding to the fairness notions addressed by the authors of the original papers. We were therefore interested in investigating the extent to which the consistency patterns observed under the fairness metrics adopted in our study were confirmed (or reversed) in case we monitored fairness through the metrics proposed in the original papers. To this end, we experimented with the three mitigation procedures whose original paper differs substantially from this perspective, with our framework.

Specifically, [92]'s paper originally monitored demographic parity with respect to the F1 Score (F1@10) as a fairness evaluation metric. We therefore computed this metric under our evaluation protocol and reported the results in Table 4.7. It can be observed that no recommendation model, after mitigating through the

[3] [92] mitigation extracts only the top 10 items that maximize utility and fairness, setting the other prediction scores to 0. KS is lower for this approach because most of the prediction scores are equal to 0.

Table 4.7: [Consistency] Recommendation utility (*F1* score, the higher it is, the more useful the recommendations) and equity (F1 Demographic Parity - *DP*, the closer to zero it is, the fairer the model) assessment for the recommendation models before (*Base*) and after (*Mit*) [92]'s mitigation. These two metrics were those used in the original paper.

	Gender								Age							
	ML 1M				LFM 1K				ML 1M				LFM 1K			
	F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$	
Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
BiasedMF	0.057	0.024	*0.003	-0.003	**0.043**	**0.018**	˙-0.014	˙-0.006	0.057	0.024	***0.001**	-0.002	**0.043**	0.017	**0.002**	˙0.009
NCF	0.056	**0.026**	*0.004	-0.003	0.035	**0.018**	˙-0.008	-0.002	0.056	**0.026**	˙0.003	**-0.001**	0.035	**0.018**	0.003	**0.004**
PMF	**0.058**	**0.026**	**0.000**	*-0.004	0.029	0.011	˙-0.007	-0.002	**0.058**	**0.026**	*0.002	**-0.001**	0.029	0.010	0.003	0.006
STAMP	0.008	0.008	*0.001	***0.001**	0.021	0.014	**-0.002**	**0.000**	0.008	0.008	**-0.001**	**-0.001**	0.021	0.014	**0.002**	**0.004**

Configurations that resulted in a statistically significant difference in F1 score (for *DP*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "˙" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

Table 4.8: [Consistency] Fairness assessment of recommendation models before (*Base*) and after (*Mit*) [58]'s mitigation, on the metric used in the original paper, i.e. ϵ-fairness (the closer to zero it is, the fairer the model).

	Gender				Age			
	ML 1M		LFM 1K		ML 1M		LFM 1K	
Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit
LBM	0.022	**0.012**	**0.009**	**0.009**	0.094	**0.092**	**0.010**	**0.010**

above mitigation procedure, reported a lower unfairness estimate for all the settings. Therefore, in contrast to the patterns observed in the unified protocol above, the method was not *consistent* on any of the models used in the original work in terms of F1 DP. Conversely, [58]'s work original assessed fairness with the ϵ-fairness score. This score is closer to 0 if, for any two items[4], the proportion of users with the same preference is approximately the same in all the demographic groups. The results in Table 4.8 show a decrease in unfairness only on ML 1M. No improvement was reported on LFM 1K, making [58]'s mitigation procedure not *consistent* on LBM.

Finally, [21]'s work monitored consumer unfairness using the category equity score (CES) across item categories as a proxy. For any item category, when the items distribution for that category in the recommendation lists is equal across demographic groups, CES is equal to 1 (the recommender system is considered fair). ML 1M originally included the movie categories. On the other hand, LFM 1K did not contain song categories, but just the MusicBrainz[5] ID of the artist. This ID was used to retrieve the genres from the artist profile in the platform (if available), assigned by sorting the artist's genres by vote and taking the most voted genres (more than one in case they had the same number of votes). Then, for any two genres A, B, A $\neq$ B, genre B was replaced by genre A if the name of genre A was a

[4]A maximum of 5000 unique combinations of items have been used to reduce execution time.

[5]`https://musicbrainz.org/`

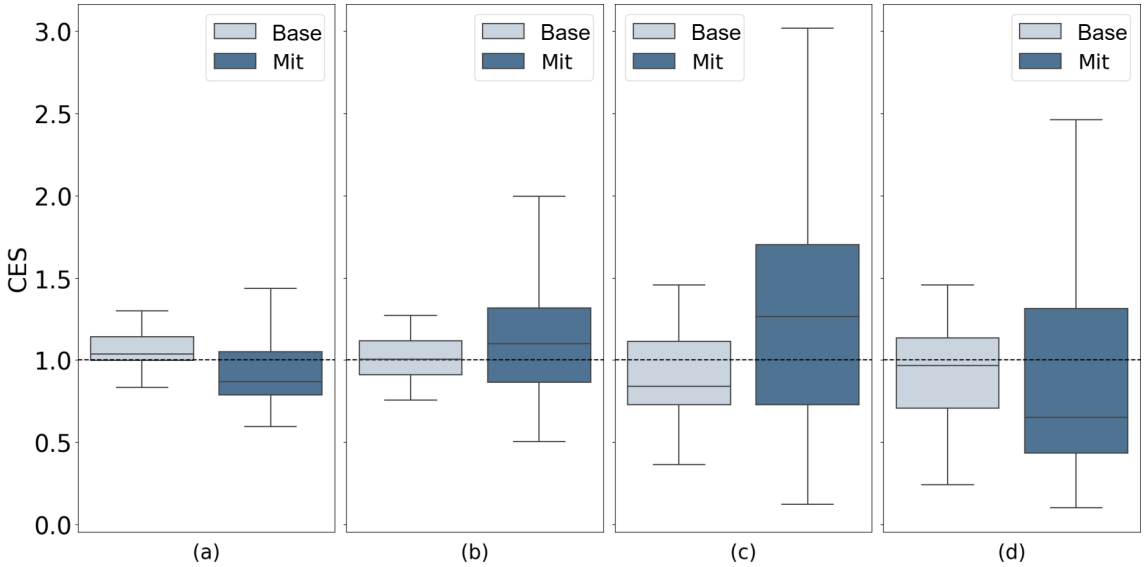


Figure 4.3: [Consistency] Category equity score distribution across item categories (*CES*, the closer to 1 it is, the more similar the category representation between the interactions and the recommendations) before (*Base*) and after (*Mit*) [21]'s mitigation under for the following combinations of datasets and demographic groups: (a) ML 1M - Gender, (b) ML 1M - Age, (c) LFM 1K - Gender, (d) LFM 1K - Age.

sub-string of the name of genre B (e.g., *death metal* was replaced by *metal*). Each box plot in Figure 4.3 represents the averaged CES distribution on a dataset for a certain attribute, before and after applying [21]'s mitigation procedure. It can be observed that the CES obtained after mitigating had a higher variance and the average score was less close to 1 (and so the recommender system was less fair), compared to the original model's CES. Overall, this mitigation procedure would not be considered *consistent* on SLIM-U according to our definition.

Data Robustness. To investigate *data robustness*, we analyzed the correlation between the satisfaction of the users and the distribution of popularity over items. For every row, Figure 4.4 shows three sub-plots, with each tick of the x-axis representing a group of items with similar popularity. The groups were formed by taking 1,000 consecutive items from an item list sorted by decreasing popularity. Considering only the interactions in the training set, each heatmap reports the difference in percentage concerning a certain metric (NDCG, NDCG DP, and KS), computed between the majority and the minority group. Each cell in sub-plots (a)-(d) represents the difference in percentage of the users who interacted with the respective group of items, (b)-(e) show the difference in percentage of users who received the items in the respective group as the top-10 recommendations, finally (c)-(f) collect the difference in percentage of users satisfied by the items in the respective group in the top-10 recommendations. We consider a user as satisfied from receiving an

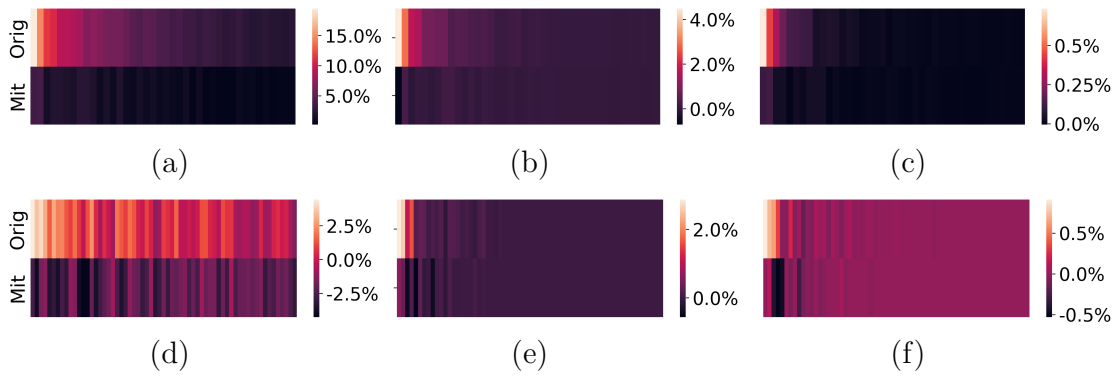


Figure 4.4: [Data Robustness] User interaction, recommendation, and relevant recommendations drift across item groups formed based on their popularity. The results were obtained by mitigating unfairness through [47]'s procedure on the original UserKNN recommendation model, on ML 1M for gender groups (a, b, c) and on LFM 1K for age groups (d, e, f). Each tick of the x-axis represents a group of $1,000$ items with similar popularity, with groups formed by sorting items based on decreasing popularity and taking disjoint sets of 1,000 consecutive items for each group. The sub-plots (a, d) represent the percentage drift of the users who interacted with that group of items, (b, e) represent the percentage drift of users who received the items in that group within the top-10 recommendations, and (c, f) represent the percentage drift of users who have been satisfied by receiving items of that group in the top-10 recommendations.

item as a recommendation if that item was included in the recommended list, and the testing set of that user included that item.

Sub-plots (a)-(b)-(c) show an analysis on gender groups by considering the predictions of UserKNN on ML 1M and the respective mitigation procedure of [47]. In particular, this method balances the consumer group representation in the training set, which leads to a more balanced interactions distribution for each demographic group over items. This aspect was highlighted by the lower differences in sub-plot (a). Sub-plot (b) reveals that also the items distribution in the recommendations was more balanced: the cells are darker for the mitigation with respect to the original model, which reported a bias towards male consumers. [47]'s mitigation on UserKNN successfully reduced unfairness. Sub-plot (c) reports similar satisfaction levels between male and female consumers, depicted as a lower intensity within the respective cells, compared to the ones associated with the original model. These results show that [47]'s procedure satisfies *data robustness* for ML 1M on gender groups. The method generates fairer recommendations by taking into consideration also the popularity bias as a possible causal reason of the unfairness between gender groups.

Sub-plots (d)-(e)-(f) analyze the UserKNN predictions on LFM 1K, when [47]'s mitigation was applied for age groups. UserKNN did not seem to be significantly

affected by the mitigation method, when item popularity was considered. Instead, the balancing leads to an higher percentage of interactions for the minority group (sub-plot (d)), and so a more equal items distribution in the recommendations across demographic groups (sub-plot (e)). However, the satisfaction level switched from the majority to the minority group for the most popular items (sub-plot (f)). Interestingly, some item groups recommended by the original model satisfied more the minority group. This suggests non-sensitive variables other than item popularity could affect model's unfairness. Hence, [47]'s mitigation did not satisfy *data robustness* on LFM 1K for age groups.

Overall, mitigation procedures that can leverage data characteristics causally-related to unfairness to reduce it would provide a more informed solution to the problem. Such procedures would need to satisfy *data robustness* and support the development of mitigation procedures according to the exploratory findings.

Reproducibility. To inspect the *reproducibility* of the considered mitigation procedures, we leveraged the findings derived from our paper collection process and source code retrieval, again limited to the top-n recommendation task. The analysis reported that only 50% (4/8) of the papers could be reproduced. The other works were marked as non-reproducible if the source code was not publicly shared or additional requirements were not met.

During our analysis on the unfairness level of recent mitigation procedures proposed in the recommendation literature, some authors of the papers we collected published the source code of their algorithm, as previously mentioned in Footnote 1. In light of this, without considering additional requirements, in this paragraph we consider papers as reproducible if we found evidence of source code publicly available, even though we did not actually deal with source codes made public in a second step. From our investigation, for the 75% (6/8) of the papers, a source code repository was linked in the paper or the code was found on the Web. Such analysis shows that the *reproducibility* level is higher than the 50% reported by our paper collection process in Section 4.3.1, but it remarks in any case the need to sharing the source code. In particular, no link to a public repository was reported in [141] and [155], and no trace was found querying on the web through a search engine. Indeed, the source code used by the authors of these two publications is not available anymore or it has not been made public yet.

Scalability. Table 4.9 shows the estimated requirements of time and memory for the recommendation models treated with their respective mitigation procedures. SLIM-U was the model with the highest time requirement for ML 1M, even though the amount of iterations was overall the lowest among the considered approaches. This high time requirement was caused by the in-processing mitigation procedure itself, since the original recommendation model could complete the training in less than a third of that time In LFM 1K, the model was faster probably due to the lower

Table 4.9: [Scalability] Number of iterations (*Iterations*), execution time (*Time*) and computational resources (*Memory*) required by the considered unfairness mitigations on the recommendation models considered in the original papers.

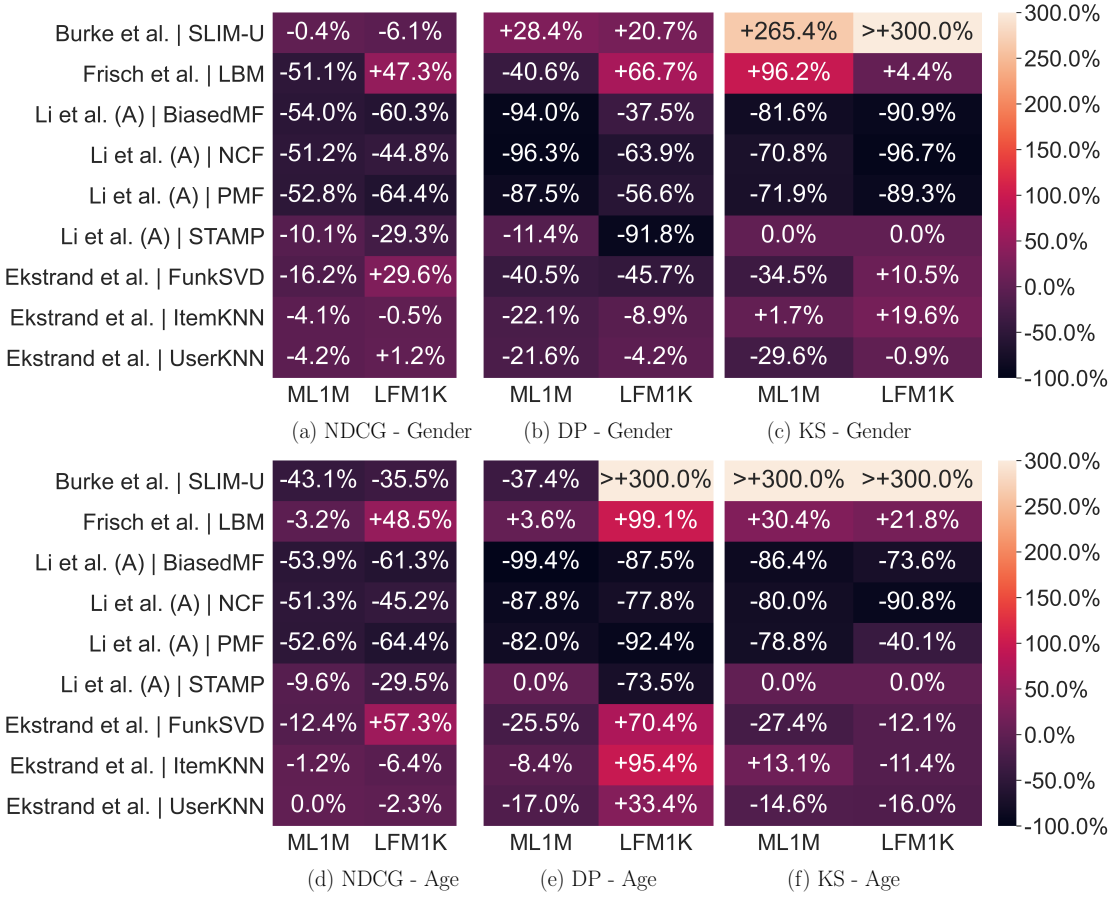
Model	**Iterations**	**ML 1M**		**LFM 1K**	
		Time	Memory	Time	Memory
SLIM-U [21]	10	˜10h	˜10 GB	˜2m	˜2 GB
LBM [58]	300	˜2h	˜500 MB	˜3h	˜400 MB
PMF [92]	100	˜5-10m	˜20 GB	˜5-10m	˜20 GB
BiasedMF [92]	100	˜5-10m	˜20 GB	˜5-10m	˜20 GB
NCF [92]	100	˜5-10m	˜20 GB	˜5-10m	˜20 GB
STAMP [92]	100	˜5-10m	˜20 GB	˜5-10m	˜20 GB
FunkSVD [47]	150	0.01s	-	0.01s	-
UserKNN [47]	-	0.01s	-	0.01s	-
ItemKNN [47]	-	0.01s	-	0.01s	-

number of users compared to ML 1M (although the number of items in LFM 1K is higher than ML 1M). The highest memory demand among mitigation procedures was required by [92], due to the mathematical optimization solver, which stored a variable for any interaction in the predictions file[6]. [47]'s mitigation was the most scalable according to our framework. This pre-processing approach did not need additional computational memory to operate and the group balancing was the fastest among mitigation procedures.

Scalability was evaluated also for the studies whose source code was not available in a first phase or it could not be reproduced. The assessment was performed on the basis of a theoretical analysis from a time complexity viewpoint. In particular, we analyzed the structure of the mitigation procedures and their scalability with regard to the increment of the number of sensitive attributes, users, items, and interactions. Specifically, [93, 154, 155] include components that can be largely affected by the number of the sensitive attributes in their in-processing approaches, e.g., discriminators to generate bias-free embeddings and pairwise fairness losses. [141]'s procedure swaps items within the recommendation lists on the basis of a fairness notion, requiring knowledge of demographic groups and items categories. Nevertheless, it is scalable because it is not affected by any of the considered factors and depends on the manually specified number of swaps.

In general, applying a mitigation procedure is an additional step in a recommendation pipeline aimed to reduce unfairness according to the pursued fairness notion. Approaches that excessively affect recommendation models in terms of time and memory would not be well suited for this purpose. On datasets with a higher number of users or interactions, the [21]'s in-processing mitigation would make SLIM-U

[6]The original paper took into account only the top-k items, which did not guarantee the fairest possible outcome.



(a) NDCG - Gender (b) DP - Gender (c) KS - Gender

(d) NDCG - Age (e) DP - Age (f) KS - Age

Figure 4.5: [Trade-off] Gain/loss in recommendation utility (*NDCG*), equity (NDCG Demographic Parity - *DP*) and independence (Kolmogorov-Smirnov - *KS*) resulted from applying the mitigations for reducing unfairness on the recommendation models used in the original papers, considering gender (a, b, c) and age (d, e, f) groups. Concerning NDCG (DP and KS), positive (negative) percentages indicate a gain in recommendation utility (fairness) after mitigating.

much slower, while [92]'s mitigation procedure would lead to unmanageable memory requirements. Since the primary task of a recommender system is suggesting items of potential interest to users, researchers should focus on designing mitigation procedures that account as much as possible for *scalability*.

Trade-off Management. To inspect trade-off management, Figure 4.5 reports, for each sensitive attribute and target metric, the variation of the latter after applying the respective mitigation. The value in the heatmap cell represents the ratio in percentage between the metric score achieved after mitigating and the one achieved by the original model. For instance, for NDCG, this means that all positive values are gains over the original model, while for fairness metrics (DP, KS) a gain in

fairness is represented by a loss in the considered metrics (the lower they are, the fairer).

For gender groups (top row of the figure grid), it can be observed that [92]'s mitigation was the best at reducing unfairness in terms of both DP and KS. However, the NDCG loss was very high, with a negative peak of -64.4% for PMF on LFM 1K. A mitigation procedure that better met the *trade-off* management property was the one proposed by [47]. Indeed, on ML 1M, both DP and KS were improved for most of the models, while the NDCG loss was negligible. On LFM 1K, only UserKNN was positively affected by the mitigation, while a higher KS was reported by the other two models under [47]'s mitigation.

Concerning age groups, no mitigation procedure was able to perform well on both datasets in terms of trade-off. [92]'s mitigation reported a very high fairness estimate, with a reduction of -99.4% of DP for BiasedMF, but again the NDCG decrease was substantially high. [47] reported the best *trade-off* on ML 1M. Their mitigation procedure led to gains on both DP and KS (except ItemKNN for KS) and a minimal loss in NDCG, e.g., UserKNN. On the other hand, only the mitigation method proposed by [92] was successful in reducing DP for LFM 1K, whereas the best result reported by the other mitigation procedures was an increase of 33.4% on UserKNN. Conversely, in terms of KS, we observed gains in fairness by most of the models and mitigation procedures. Overall, [47]'s method was the one that reported the best *trade-off* across all the datasets and sensitive attributes. It accomplished the goal of reducing unfairness, while minimally affecting the recommendation utility. Researchers are encouraged to consider this property, so as to develop mitigation procedures that could be applied in practice in an effective way.

Transferability. To inspect *transferability*, we applied the mitigation procedures proposed by [47] and [92] on the recommendation models used by the other papers. We selected those two mitigation procedures, since they had the highest applicability among the considered ones. Table 4.10 reports the recommendation utility and fairness estimated for the recommendation models treated with the [47]'s procedure for gender and age groups. The column `Paper` identifies only the paper that used the respective model (under column `Model`) in their study, not the mitigation applied.

In ML 1M, the performance measured for the considered mitigation procedure highly varied, decreasing DP for some models, e.g., SLIM-U, and increasing it for others, e.g., BiasedMF and PMF, for both age and gender groups. A similar behavior was reported for KS, but not always for the same models. The results were better for LFM 1K, but the performance still varied, reducing DP in a substantial way on SLIM-U and STAMP for both gender and age groups. KS was improved for most of the models for both sensitive attributes, except when the mitigation strategy was applied on BiasedMF, which reported a higher KS for both gender and age groups. In summary, [47] does not hold a good transferability level.

For gender and age respectively, Table 4.11 reports the results obtained after

Table 4.10: [Transferability - [47]] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Base*) and after (*Mit*) [47]'s mitigation for gender and age groups.

			ML 1M						LFM 1K					
			NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
	Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Gender	Burke et al.	SLIM-U	0.084	0.079	ˆ-0.022	*0.021	ˆ-0.032	ˆ-0.032	**0.320**	**0.356**	*-0.060	*-0.082	ˆ-0.006	ˆ-0.006
	Frisch et al.	LBM	0.044	0.020	**ˆ-0.006**	**ˆ-0.001**	ˆ-0.013	ˆ-0.012	0.144	0.034	*-0.035	**-0.008**	ˆ-0.120	ˆ-0.061
	Li et al. (A)	BiasedMF	0.112	0.101	ˆ-0.017	ˆ-0.027	ˆ-0.035	ˆ-0.034	0.287	0.261	ˆ-0.095	-0.068	ˆ-0.012	ˆ-0.015
		NCF	0.117	**0.113**	ˆ-0.016	ˆ-0.018	ˆ-0.022	ˆ-0.017	0.250	0.185	*-0.073	*-0.059	ˆ-0.033	ˆ-0.020
		PMF	**0.119**	0.101	*0.013	ˆ-0.027	ˆ-0.023	ˆ-0.030	0.200	0.194	*-0.062	-0.049	ˆ-0.010	ˆ-0.009
		STAMP	0.022	0.027	*0.003	**0.001**	**ˆ-0.006**	**ˆ-0.006**	0.160	0.150	**-0.021**	-0.027	**ˆ-0.001**	**ˆ-0.002**
Age	Burke et al.	SLIM-U	0.084	0.081	ˆ-0.022	ˆ-0.017	ˆ-0.009	ˆ-0.013	**0.320**	**0.345**	-0.026	-0.084	ˆ-0.017	ˆ-0.013
	Frisch et al.	LBM	0.044	0.026	ˆ-0.005	0.007	ˆ-0.021	ˆ-0.013	0.144	0.016	**-0.011**	**-0.007**	ˆ-0.125	ˆ-0.066
	Li et al. (A)	BiasedMF	0.112	0.107	ˆ-0.015	ˆ-0.039	ˆ-0.042	ˆ-0.032	0.287	0.271	*-0.079	-0.064	ˆ-0.019	ˆ-0.023
		NCF	0.117	**0.117**	ˆ-0.018	*0.018	ˆ-0.029	ˆ-0.019	0.250	0.186	*-0.067	-0.030	ˆ-0.055	ˆ-0.018
		PMF	**0.119**	0.107	ˆ-0.020	ˆ-0.039	ˆ-0.027	ˆ-0.034	0.200	0.175	*-0.046	-0.026	**ˆ-0.008**	ˆ-0.011
		STAMP	0.022	0.026	**0.000**	**0.003**	**ˆ-0.006**	**ˆ-0.007**	0.160	0.149	-0.031	*-0.051	**-0.008**	**ˆ-0.006**

Configurations that resulted in a statistically significant difference in NDCG (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "ˆ" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

Table 4.11: [Transferability - [92]] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Base*) and after (*Mit*) [92]'s mitigation for gender and age groups.

			ML 1M						LFM 1K					
			NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
	Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Gender	Burke et al.	SLIM-U	0.084	0.046	ˆ-0.022	ˆ-0.008	ˆ-0.032	ˆ-0.007	0.320	0.172	*-0.060	ˆ-0.033	**ˆ-0.006**	**ˆ-0.001**
	Ekstrand et al.	FunkSVD	0.018	0.016	**ˆ-0.004**	**ˆ-0.003**	ˆ-0.027	ˆ-0.007	0.010	0.010	**-0.006**	***-0.005**	ˆ-0.107	**ˆ-0.001**
		ItemKNN	**0.140**	**0.079**	ˆ-0.038	ˆ-0.008	ˆ-0.030	ˆ-0.007	0.287	0.016	ˆ-0.127	-0.006	ˆ-0.019	**ˆ-0.001**
		UserKNN	0.137	0.043	ˆ-0.031	ˆ-0.007	ˆ-0.074	ˆ-0.007	**0.406**	**0.189**	ˆ-0.110	*-0.036	ˆ-0.067	**ˆ-0.001**
	Frisch et al.	LBM	0.044	0.037	ˆ-0.006	ˆ-0.005	**ˆ-0.013**	**ˆ-0.006**	0.144	0.118	*-0.035	*-0.032	ˆ-0.120	**ˆ-0.001**
Age	Burke et al.	SLIM-U	0.084	0.046	ˆ-0.022	ˆ-0.006	**ˆ-0.009**	**ˆ-0.006**	0.320	0.175	-0.026	0.024	**ˆ-0.017**	ˆ-0.006
	Ekstrand et al.	FunkSVD	0.018	0.016	ˆ-0.008	ˆ-0.007	ˆ-0.029	**ˆ-0.006**	0.010	0.010	**0.002**	**0.002**	ˆ-0.054	ˆ-0.006
		ItemKNN	**0.140**	**0.079**	ˆ-0.027	**ˆ-0.004**	ˆ-0.029	**ˆ-0.006**	0.287	0.018	0.010	*0.012	ˆ-0.133	ˆ-0.006
		UserKNN	0.137	0.043	ˆ-0.028	ˆ-0.007	ˆ-0.060	**ˆ-0.006**	**0.406**	**0.190**	-0.023	0.032	ˆ-0.036	ˆ-0.006
	Frisch et al.	LBM	0.044	0.037	**ˆ-0.005**	ˆ-0.006	ˆ-0.021	**ˆ-0.006**	0.144	0.120	-0.011	**-0.002**	ˆ-0.125	**ˆ-0.005**

Configurations that resulted in a statistically significant difference in NDCG (for *DP*) or predicted relevance (for *KS*) distributions between the two groups under a *Mann-Whitney U* test are indicated with the symbol "ˆ" ($p < 0.01$) and the symbol "*" ($p < 0.05$) respectively.

applying [92]'s mitigation to the models included in the other papers. In ML 1M, DP was reduced in all the settings, except for LBM on age groups. In LFM 1K, the mitigation procedure improved DP by a considerable amount for gender groups, e.g., ItemKNN, whereas DP for age groups increased for two models out of five. KS cannot be analyzed for [92] as pointed out in Footnote 3. Due to the poor

mitigation performance on age groups in LFM 1K, transferability is not satisfied by [92]. We encourage readers to consider *transferability* in their future works, enforcing the solidity of the logic of the proposed technique, as well as giving a clear demonstration on how their mitigation procedure can be applied to other models.

4.4 Unfairness Explanation via Graph Perturbation

This section describes our study aimed to generate explanations of unfairness experienced by GNN-based recommender systems. The previous section highlighted the significant effort that has been recently put by the recommendation community to provide the end users with equitable recommendations. However, it is also important for service providers (e.g., an online platform) to understand *why* the model behind their platform is unfair, which is still an unexplored area with just a handful of studies. Starting from a popular technique in GNNs based on the manipulation of the graph topological structure, we devised a method that generates a global explanation of unfairness measured in a GNN-based recommender system, by analyzing the bipartite graph representing the user-item interactions and finding a subset of them affecting the system fairness. Specifically, we propose `GNNUERS` (**GNN**-based **U**nfairness **E**xplainer in **R**ecommender **S**ystems) and present its three main components:

1. An extended and improved *graph perturbation* algorithm, specifically devised for bipartite graphs, which recommendation systems based on GNNs typically use. We specialize the manipulation of the graph topology on a subset of edges, i.e. user-item interactions, to only account for the graph regions relevant for recommendation tasks.

2. A *perturbed graph generator* that leverages the previous component and queries the model to estimate the change in fairness resulting from the applied perturbation, and iteratively adjusting the set of edges to manipulate.

3. A *two-term loss function* specifically designed to guide the optimization process towards selecting user-item interactions that significantly affect the fairness of the system, such that these interactions could be provided as an explanation of the system's underlying unfairness.

A visual representation of `GNNUERS`'s operation is depicted in Figure 4.6.

4.4.1 Methodology

The next sections will not only describe the `GNNUERS`'s three components previously mentioned, but we will also dive into a constraint adopted to the loss function to

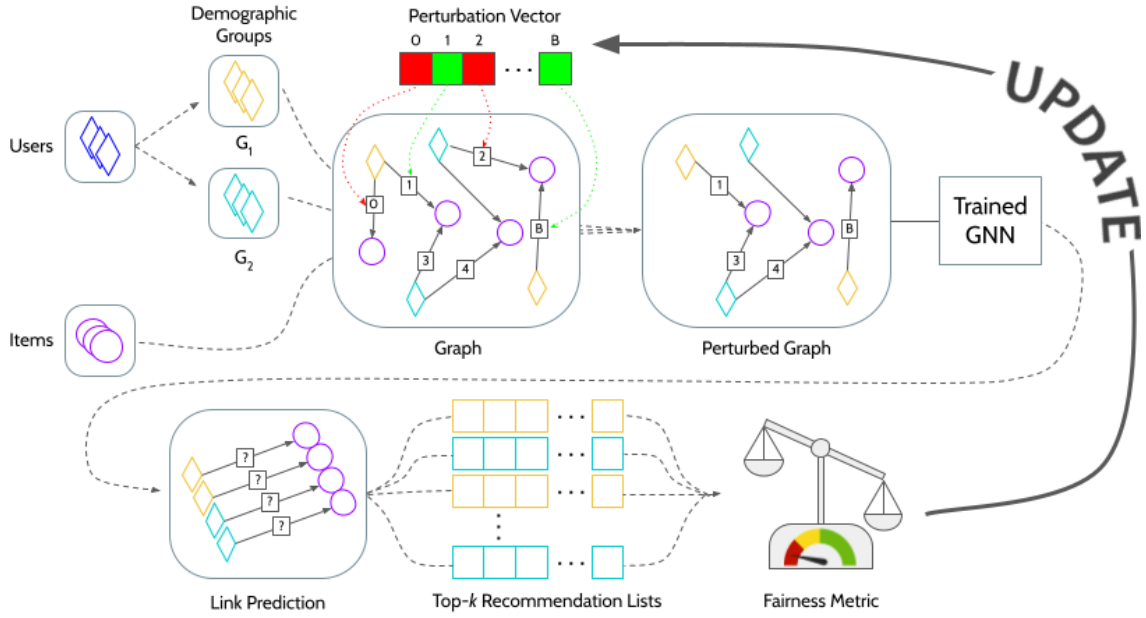


Figure 4.6: GNNUERS updates the perturbation vector such that the removed user-item interactions from the graph lead the trained GNN to generate fairer recommendations. The perturbation vector represents the counterfactual explanation of the prior unfairness across demographic groups.

improve the targeted task in Section 4.4.1, and analyze our method in terms of resources usage in Section 4.4.1.

Bipartite Graph Perturbation

Our graph perturbation approach is inspired by previous work for GNNs explanation for binary classification on plain graphs [97]. However, since GNNUERS aims to perturb a bipartite graph generated for recommender systems, it presents several differences. In [97] a perturbation matrix P is populated to then generate the perturbed matrix $\tilde{A} = P \odot A$[7]. Optimizing for P can eventually include indices for zero entries in A. While for plain graphs this method results to be efficient, for bipartite graphs it can be memory inefficient, mainly because it requires to store a perturbation value also for the user-user and item-item links, not present in bipartite graph per definitionem. To overcome these limitations, the perturbation in GNNUERS is optimized through a vector $p \in \mathbb{N}^B$, where B is the number of existing edges in the original graph. Our method is memory efficient, especially under sparse graphs, since it needs to store perturbation values only for non-zero entries of A.

Given our unfairness explanation task, we aim to find a set of interactions in A that led a GNN to generate unfair recommendations. To do so, we derive a perturbed

[7] $\odot$ denotes the Hadamard product.

matrix $\tilde{A}$, resulting in fairer recommendations when a trained GNN uses $\tilde{A}$ instead of A during the inference phase. The non-zero entries $A_{u,i} \neq 0$ are perturbed by p through a function $h : \mathbb{N}^2 \rightarrow \mathbb{N}$ that maps the 2D indices (u, i) of A to a 1D index for p. Thus, given $j = h(u, i), j < B$, an entry $p_j = 0$ denotes the edge $A(u, i)$ is deleted in the perturbed adjacency matrix, i.e. $\tilde{A}_{u,i} = 0$. In other words, the perturbed matrix $\tilde{A}$ is populated as follows:

$$\tilde{A}_{u,i} = \begin{cases} p_{h(u,i)} & if\ h(u,i) < B \\ A_{u,i} & otherwise \end{cases} \tag{4.16}$$

The perturbation mechanism is therefore driven by the way h and B are defined.

Following [97, 135], we generate p through an optimization process that leverages a real valued vector $\hat{p}$. Once optimized, we apply a sigmoid transformation, and then a binarization of the entries such that values ≥ 0.5 become 1, while values < 0.5 become 0, obtaining eventually p. The initialization of $\hat{p}$ should guarantee $\tilde{A} = A$, i.e., a real-valued α is selected to initialize $\hat{p}$, such that $p_j = 1, \forall j \in [0, B)$. In all the experiments in Section 4.4.3 we set $\alpha = 0$, leaving the analysis of other initialization values as a future work.

Perturbed Graph Generation

Based on the protocol described above, `GNNUERS` modifies the adjacency matrix edges by means of the perturbation vector p. The decision process of which edges will be deleted is performed by the counterfactual model $\tilde{f}$. $\tilde{f}(A, W; \hat{p}) \rightarrow \tilde{R}$ extends the GNN-based recommender system f using $\hat{p}$ as parameter and the frozen weights W learnt by f as additional input. In detail, $\tilde{f}$, similarly to f, predicts the altered relevance matrix $\tilde{R}$ by combining the normalized version of the perturbed adjacency matrix $\tilde{L} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, where $\tilde{D}_{u,u} = \sum_i \tilde{A}_{u,i}$, with W according to the implementation of the original model f. Therefore, $\tilde{f}$ learns only $\hat{p}$, while the weights W, already optimized by f to maximize the recommendation utility, remain constant.

As explained in Section 4.4.1, p is generated from $\hat{p}$, whose values get updated during the learning process. At different steps of the latter, the values of $\hat{p}$ could oscillate close to the threshold that determines if p will be 0 or 1 at the respective indices. Considering aspects such as floating-point errors or dropout layers, the oscillation could negatively affect the update of $\hat{p}$, due to previously perturbed edges being restored, or vice versa. To counter this phenomenon, the perturbation algorithm is constrained by the usage of a policy that prevents a deleted edge from being restored, such that the number of perturbed edges follows a monotonic trend.

Loss Function Optimization

The previous section introduced $\tilde{f}$, the counterfactual model responsible of the generation of the perturbed adjacency matrix $\tilde{A}$. In our context, we assume to model

counterfactual explanations according to the users' history. More precisely, a set of user-item interactions, perturbed with respect to the original ones, represents a counterfactual explanation in case a trained model produces at least one different recommendation to the users, when these perturbed interactions are used in the inference phase. In our graph-based approach, it means that we aim to generate a perturbed version of the adjacency matrix A, i.e. $\tilde{A}$, that leads to the recommended lists $\tilde{Q}$, with $\tilde{Q} - Q \neq \emptyset$, when a GNN uses $\tilde{A}$ (instead of A) for prediction. Following [59], if the predictions, i.e. recommendations, generated by a model are fairer when a set of modifications, i.e. the perturbed interactions in $\tilde{A}$, is applied to the original data, i.e. A, such modifications represent the counterfactual explanation of the unfairness estimated in the original predictions. Under our user unfairness explanation task, we specifically aim to produce a perturbed adjacency matrix $\tilde{A}$ (counterfactual explanation) that leads to the highest fairness across users by means of the lowest number of perturbations on the original adjacency matrix A.

Motivated by its increasingly recognized importance in prior work in top-k recommendation [155], we decided to model fairness according to the notion of demographic parity, even though our formulation and method is flexible to accommodate other notions of fairness. Under this demographic parity notion of fairness, our goal is to generate a perturbed adjacency matrix $\tilde{A}$ that modifies the predictions of a trained GNN, resulting in recommended lists $\tilde{Q}$ with closer utility estimates across demographic groups than the original recommendations Q, constrained to the number of perturbed edges with respect to the original adjacency matrix A. Formally, we seek to minimize the following objective function:

$$\mathcal{L}(A, \tilde{A}) = \mathcal{L}_{fair}(A, f(\tilde{A}, W)) + \mathcal{L}_{dist}(A, \tilde{A}) \tag{4.17}$$

where $\mathcal{L}_{fair}$ is the term monitoring fairness, operationalized according to the notion of demographic parity, $\mathcal{L}_{dist}$ is the term controlling the distance between the perturbed adjacency matrix $\tilde{A}$ and the original one A. Therefore, $\mathcal{L}$ drives the optimization of the perturbation vector $\hat{p}$, such that the perturbed graph generator updates the entries of $\hat{p}$ on the basis of the instructions provided by $\mathcal{L}$.

We follow recent works [155] to operationalize the demographic parity (DP) notion, i.e. the core of $\mathcal{L}_{fair}$, as the mean of the absolute pair-wise utility difference across all demographic groups. Formally:

$$\mathcal{L}_{fair}(A, \tilde{R}) = \frac{1}{\binom{|G|}{2}} \sum_{1 \leq i < j \leq |G|} \left\| S(\tilde{R}^{g_i}, A^{g_i}) - S(\tilde{R}^{g_j}, A^{g_j}) \right\|_2^2 \tag{4.18}$$

where S is a function that measures the recommendations utility level, G is the set of considered demographic groups, $\tilde{R}^{g_i}$ ($\tilde{R}^{g_j}$) and A^{g_i} (A^{g_j}) denote, respectively, the altered relevance sub-matrix and the adjacency sub-matrix with respect to the users belonging to g_i (g_j).

Following works that proposed methods to mitigate or explain unfairness in recommendation [92, 9, 81, 59], we focus on a binary setting, with sensi-

tive attributes comprised of two demographic groups. For instance, given $G = \{males(M), females(F)\}$, $\mathcal{L}_{fair}$ aims at minimizing the utility disparity between males and females, with the optimal result being:

$$\left\| S(\tilde{R}^M, A^M) - S(\tilde{R}^F, A^F) \right\|_2^2 = 0$$

We denote the group with higher (lower) utility on the evaluation set as *unprotected* (*protected*). This enables the reader to better contextualize the approach with respect to fairness.

NDCG was selected as the utility metric S. However, due to the non-differentiability of the sorting operation performed to compute NDCG, we adopt an approximated version [155, 120], which we refer as $\widehat{NDCG}$[8]:

$$\begin{aligned} \widehat{NDCG}(r, a) = & -\frac{1}{DCG(a, a)} \sum_i \frac{2^{a_i} - 1}{log_2(1 + z_i)} \\ \text{s.t.} \quad & z_i = 1 + \sum_{j \neq i} \sigma\left(\frac{r_j - r_i}{\gamma}\right) \end{aligned} \tag{4.19}$$

where r is the item relevance score produced by the recommender system, σ is a sigmoid function, and γ is a scaling constant. We fix $\gamma = 0.1$ for the experiments in Section 4.4.3, being the default value in the TensorFlow Ranking implementation. $\widehat{NDCG}$ is adopted only when constructing the fairness objective in Equation (4.18), while the original NDCG is used in the evaluation phase. In the same way, we denote as $\Delta\widehat{NDCG}$ the metric used to measure $\mathcal{L}_{fair}$ in our binary setting and ΔNDCG the one used to evaluate unfairness in the experiments. Given our unfairness explanation task, the ground truth labels used to measure $\widehat{NDCG}$ during the `GNNUERS` learning process are taken from the evaluation set. Such approach is justified by the explanation task of the recommendation unfairness measured on the evaluation set itself, while for other tasks, e.g., mitigation ones, having access to the ground truth labels is a less realistic assumption [121].

Any differentiable distance function can be adopted as the distance loss $\mathcal{L}_{dist}$ [97]. In `GNNUERS`, it is based on the absolute element-wise difference between $\tilde{A}$ and A, defined as follows:

$$\mathcal{L}_{dist}(A, \tilde{A}) = \beta \frac{1}{2} \sigma \left(\sum_{i,j} \left\| \tilde{A}_{i,j} - A_{i,j} \right\|_2^2 \right) \tag{4.20}$$

A sigmoid function is used to bound the distance loss to the same range of $\mathcal{L}_{fair}$, i.e. [0,1]. In particular, we used $\sigma(x) = |x|/(1 + |x|)$ which needs a higher number of perturbed edges to reach 1 compared with the popular logistic function, hence covering a wider range of values. β is a parameter that balances the two losses, due

[8] We use the TensorFlow implementation, called `ApproxNDCGLoss`.

to the trend of $\mathcal{L}_{fair}$ to report values $\ll 0.5$, while $\mathcal{L}_{dist}$ gets rapidly close to 1.0 as more edges are perturbed. We tested several values of β in the range $[0.001, 2.0]$ and the best value for each model was selected for the experiments in Section 4.4.3.

Gradient Deactivation

The optimization of (4.18) takes into account the approximate NDCG measured on the predicted recommendations for the protected and unprotected group. The update of the real-valued perturbation vector $\hat{p}$ is then affected from the viewpoint of both demographic groups. In particular, `GNNUERS` selects edges that could simultaneously optimize two objectives: increasing utility for the protected group and decreasing it for the unprotected one. However, the edges that are going to be perturbed for one of the objectives could negatively affect the other one, and vice versa. To this end, we perform a gradient *deactivation* on the recommendations generated for the protected group, i.e. the back-propagation updates the perturbation vector only from the unprotected group viewpoint. This procedure is applied only on the protected group, such that the `GNNUERS` objective is to delete the edges generating the gap in recommendation utility between unprotected and protected users.

Deactivating the gradient does not limit the group of edges that can be perturbed because the optimization does not involve only the user nodes, but also the item ones. Hence, `GNNUERS` could delete all the edges connected to an item node, both coming from user nodes of the unprotected and protected group. For conciseness, we will use the terms *deactivated* and *activated* to characterize a group associated with inactive and active gradient respectively.

Resources Usage

In this section, the two steps of the `GNNUERS` pipeline are examined in terms of memory footprint and execution time complexity. The first step regards the generation of the perturbed matrix $\tilde{A}$ at each step of the learning process by means of (4.16), which requires to store only the real-valued perturbation vector $\hat{p}$. Leveraging a sparse representation of A and $\tilde{A}$, the perturbation time complexity is dependent only on the number of perturbed edges B, i.e., $\mathcal{O}(B)$. The second step, that is the optimization process in Sections 4.4.1-4.4.1 directed towards learning p, has no memory footprint and is executed for C iterations. Hence, given Θ the execution time for the inference step of the extended GNN-based recommender system $\tilde{f}$ and Ψ the execution time of (4.17), $\mathcal{O}(\Theta\Psi CB)$ is the time complexity of the perturbed graph generation.

4.4.2 Experimental Setup

The data manipulation, training and assessment of the GNN-based recommender systems were built upon the framework Recbole [173]. The experiments were ran

on a A100 GPU machine with 80GB VRAM and 90GB RAM.

Graph Topological Properties

GNNUERS identifies explanations in the form of user-item interactions that made a GNN-based recommender system generate unfair outcomes. Each edge deleted from the graph unlinks a user and an item node, modifying the network topological structure and affecting the properties characterizing all the nodes, e.g., degree. GNNUERS edges selection process can then be described by the properties of the nodes of the removed edges.

To this end, we selected three properties that reflect different networks topological aspects and their relation to features examined in recommender systems tasks, e.g., popularity bias. Let $z \in Z$ be a generic node of $\mathcal{G}$, i.e., $Z \subset U$ if z is a user or $Z \subset V$ if z is an item, the nodes properties are defined as follows:

- **Degree (DEG)**: the number of edges connected to each node. For a user node u it represents the history length, i.e. $|I_u|$, for item nodes it represents their popularity.

- **Density (DY)**: it represents the tendency of a node to be connected to high-degree nodes. For user nodes it represents the tendency to interact with popular items, for item nodes it describes the interest of their peers, where users' interest is higher as their histories length is longer. Formally, given $\mathcal{N}_z$ the neighbors set of a node z:

$$DY_z = \frac{\sum_{i=1}^{|\mathcal{N}_z|} \frac{|\{z' \mid (\bar{z}_i, z') \in E \wedge \bar{z}_i \in \mathcal{N}_z\}|}{|Z|}}{|\mathcal{N}_z|} \tag{4.21}$$

- **Intra-Group Distance (IGD)**: it represents how a node z is close to the other nodes $z' \in Z/\{z\}$. Given the bipartite nature of recommender systems networks, we consider two users (items) being distant n if the shortest path that connects them include n items (users). IGD is the average of the number of nodes of the same type normalized by their distance to the considered node. Formally, given N the graph diameter:

$$IGD_z = \frac{\sum_{n=1}^{N} \frac{|\{z' \mid \Gamma(z, z') = n\}|}{n}}{|Z|} \tag{4.22}$$

  where Γ measures the shortest path length between two nodes of the same type.

The selected properties can describe the context on which GNNUERS operates, i.e. the GNN-based recommender systems, and insights on the unfairness can be uncovered by the variance of such properties across demographic groups, due to the

Table 4.12: Statistics of the four datasets used in our experimental protocol. *Repr.* stands for *Representation*, *Min.* for *Minimum*. *G* stands for Gender, *A* for Age, *Gini.* for the Gini coefficient applied over the values of the graph properties for each group. The line over the graph properties denotes their average.

		ML-1M [69]	FENG[9]	LFM-1K [25]	INS[10]
# Users		6,040	25,741	268	346
# Items		3,706	23,643	51,609	20
# Interactions		1,000,209	708,919	200,586	1,879
Min. User DEG		20	5	21	5
Domain		Movie	Grocery	Music	Insurance
Repr.	A	O : 43.4%; Y : 56.6%	O : 45.5% ; Y : 54.5%	O : 42.2%; Y : 57.8%	O : 49.4%; Y : 50.6%
	G	F : 28.3%; M : 71.7%	NA	F : 42.2%; M : 57.8%	F : 23.4%; M : 76.6%
$\overline{\text{User DEG}}$	A	O : 106.1; Y : 124.9	O : 20.7; Y : 19.6	O : 657.5; Y : 428.0	O : 4.3; Y : 4.5
	G	F : 101.8; M : 122.7	NA	F : 496.7; M : 545.3	F : 4.2; M : 4.5
Gini User DEG	A	O : 0.53; Y : 0.52	O : 0.45; Y : 0.44	O : 0.43; Y : 0.43	O : 0.06; Y : 0.08
	G	F : 0.53; M : 0.52	NA	F : 0.42; M : 0.45	F : 0.05; M : 0.08
$\overline{\text{User DY}}$	A	O : 0.07; Y : 0.07	O : 0.00; Y : 0.00	O : 0.03; Y : 0.03	O : 0.40; Y : 0.38
	G	F : 0.07; M : 0.07	NA	F : 0.04; M : 0.04	F : 0.40; M : 0.39
Gini User DY	A	O : 0.12; Y : 0.11	O : 0.45; Y : 0.42	O : 0.22; Y : 0.18	O : 0.12; Y : 0.14
	G	F : 0.12; M : 0.12	NA	F : 0.15; M : 0.23	F : 0.14; M : 0.13
$\overline{\text{User IGD}}$	A	O : 0.95; Y : 0.96	O : 0.57; Y : 0.56	O : 0.99; Y : 0.99	O : 0.97; Y : 0.97
	G	F : 0.95; M : 0.96	NA	F : 0.99; M : 0.99	F : 0.97; M : 0.97
Gini User IGD	A	O : 0.03; Y : 0.02	O : 0.05; Y : 0.05	O : 0.01; Y : 0.01	O : 0.01; Y : 0.01
	G	F : 0.03; M : 0.02	NA	F : 0.00; M : 0.01	F : 0.01; M : 0.01

intrinsic relationship between the given context and the unfairness. In particular, the degree (DEG) of a user node represents the interest towards the available items and the amount of information that the GNN can leverage in the aggregation step, the density (DY) of a user node captures the inclination to engage with popular items, the intra-group distance (IGD) of a user node indicates the propensity to interact with items valued by fellow users within the same demographic group. Moreover, the properties DEG and DY reflect the GNN ability to propagate information across the user nodes using the message passing mechanism, given that DEG and DY regard the nodes amount at the 1- and 2-hop distance respectively.

Data Preparation

Extensive research in user fairness in recommender systems is challenging due to the limited datasets including users' sensitive information. We relied on the artifacts of our prior study presented in Section 4.3, where we performed a fairness assessment on two corpora: MovieLens 1M (ML-1M), on the movie domain, and Last.FM 1K (LFM-1K), on the music domain. The time information of the users' interaction in LFM-1K was missing, so, given that the interactions of each user are grouped by artist and considered as a single interaction, we set the timestamp of the last interaction with an artist to be the timestamp of the relative user-artist pair in LFM-1K. We extended the set of datasets by including Insurance (INS), on the

insurance domain, and Ta Feng (FENG), on the grocery domain[11]. All datasets include age and gender (except FENG) information for all users and their statistics are listed in Table 4.12, where the graph properties values regard only the training set. The Gini coefficient for each property was measured as in [48].

User nodes in INS and FENG were filtered by their number of interactions, i.e. their degree, so as to consider users with histories made up of at least 5 items. Duplicated interactions, e.g., users buying the same product twice in FENG, were removed. Based on the binary setting mentioned in Section 4.4.1, INS and FENG age labels were binarized as *Younger (Y)* and *Older (O)*, such that the *Younger* group is more represented than the *Older* one for consistency with ML-1M and LFM-1K, while gender labels were already binary, as *Males (M)* and *Females (F)*.

We also adopted the splitting strategy used in our prior study of Section 4.3 for each dataset: per each user, 20% (the most recent if a timestamp is available, randomly sampled otherwise) of the interactions forms the test set; the remaining interactions are split again, such that 10% (selected in the same way) of this interactions subset forms the validation set and the remaining 70% forms the train set. The validation set was used to select the training epoch where the model reported the best recommendation utility on the adjacency matrix A. Given the goal of finding the edges causing unfairness in the test set, the truth ground labels of the latter were extracted to optimize the fairness loss in (4.18).

Models

Recently, novel GNNs have been devised to solve the top-k recommendation task. We relied on Recbole, which includes different families of GNNs-based recommender systems. `GNNUERS` was adopted on the following models:

- GCMC [142]: this method is comprised of two components: a graph auto-encoder, which produces a node embedding matrix, and a decoder model, which predicts the relevance of the missing entries in the adjacency matrix from the node embedding matrix.

- NGCF [149]: this state-of-the-art GNN-based recommender system propagates embeddings in the user-item graph structure. In particular, it leverages high-order connectivities in the user-item integration graph, injecting the collaborative signal into the embedding process in an explicit manner.

- LigthGCN [71]: it is a simplification of a GCN, including only the most essential components for collaborative filtering, i.e., the neighborhood aggregation. It uses a single embedding as the weighted sum of the user and item embeddings propagated at all layers in the user-item interaction graph.

[11] Yelp [100] was also considered to include the business domain, but the users' gender information was predicted by their name, making questionable analyses on this dataset.

These three GNNs are trained with the default hyper-parameters defined by Recbole, specific for each model.

Explanation Baseline Methods

As mentioned in Section 4.4.1, `GNNUERS` in its base form applies a policy that prevents the algorithm from restoring previously deleted edges. Additionally, we examined an extension of `GNNUERS` by applying another policy, *Connected Nodes (CN)*: it limits the perturbation to the edges connected to the unprotected user nodes to investigate whether recommendations unfairness is only due to the interactions performed by the unprotected group. Therefore, *CN* guides the learning process to select the users' actions of the unprotected group that made f favor them.

The literature does not include baselines that explain unfairness in the form of user-item edges as `GNNUERS`. The works proposing unfairness explainability methods in recommendation [39, 59] select relevant user/item features as explanations, which cannot be compared with the ones generated by our framework. Approaches proposed to explain unfairness in GNNs [44] were devised for classification tasks: although it is not clear if this method could be extended to recommendation tasks, such an engineering adaptation goes beyond the goal of our work. Other alternative counterfactual explainability algorithms in GNNs [97, 82] generate explanations at the instance level, which cannot be adapted to envision the unfairness task at the model level.

To this end, we adopted CASPER [117] for comparison, a model-agnostic method that causes the highest instability in the recommendations by perturbing a single interaction, i.e. an edge of the graph. The instability induced by CASPER could alter the recommendations, and, as a result, re-distributing the utility levels over the demographic groups and positively affecting unfairness. At inference step, our models generate the recommendation lists by using the training network perturbed by CASPER, then fairness and utility metrics are measured. CASPER uses the timestamp of each interaction to generate a directed acyclic graph of the interactions of each user. INS does not include the time information, so CASPER was not applied on this dataset.

We also introduce *RND-P* as sanity check, a baseline algorithm that at each iteration randomly perturbs edges with a probability ρ, such that it mimics the `GNNUERS` edges selection process, but based on a random choice. Given the size diversity of our evaluation datasets, we set $\rho = 1/(|E_{train}|/100)$, where E_{train} is the set of training edges, as the value that works best across the selected epochs, such that *RND-P* perturbs edges depending on the network size to prevent this method from deleting all the edges in a few iterations.

The explanations methods were executed on all the models and datasets over 800 epochs adopting an early stopping method when $\mathcal{L}_{fair}$ does not improve with a delta higher than 0.001 for at least 15 consecutive epochs.

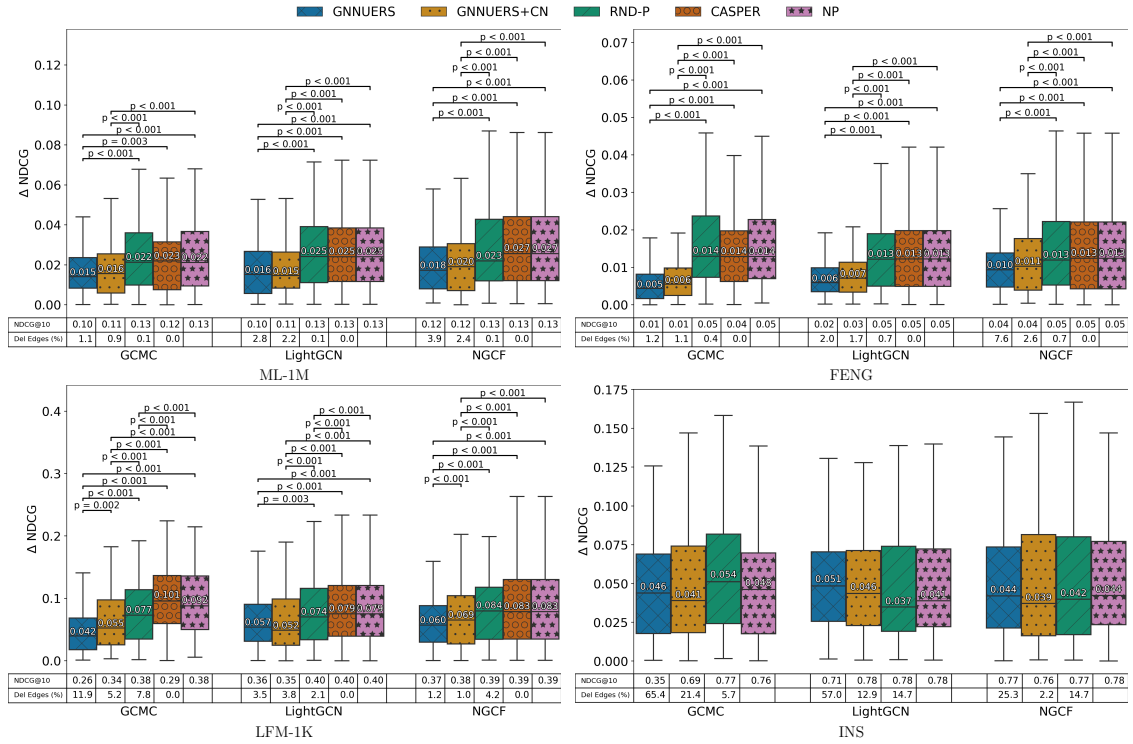


Figure 4.7: Distribution of ΔNDCG between younger and older users subgroups, randomly sampled 100 times. A Wilcoxon signed-rank test is performed between each pair of boxes and the respective p-value is shown if it is lower than $\frac{0.05}{m}$ according to the Bonferroni correction, where m is the number of pairwise comparisons performed for each model.

4.4.3 Results

Unfairness Explainability Benchmark

We first investigated the capability of `GNNUERS` to select counterfactual explanations that effectively optimize (4.18). `GNNUERS` learning process selects users coming from both demographic groups, stores them in fixed size batches according to their distribution in the dataset and, optimizes the loss to minimize disparity in NDCG@10 (average) between the protected and unprotected group. The evaluation follows an analogous process: we randomly sample 100 subgroups with the same demographic groups distribution, with sample size equal to the batch size. This choice is also due to reduce the sampling bias present in the datasets, i.e. the evaluation is not affected by the different sample size of unprotected and protected groups. The batch size for each dataset is selected such that it splits the users in at least five partitions, guaranteeing a low probability of picking the same users in the randomly sampled subgroups. We measure the differences in each subgroup with ΔNDCG, i.e. the differences in performance between the two user groups, related to the unfairness

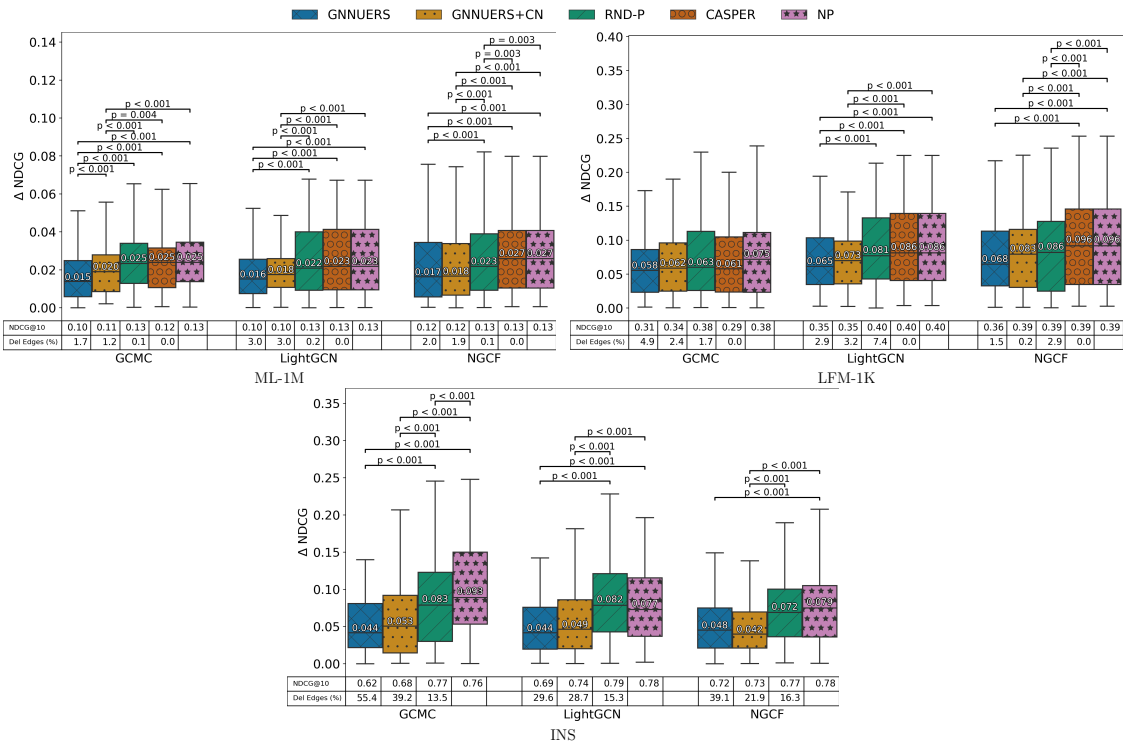


Figure 4.8: Distribution of ΔNDCG between males and females users subgroups, randomly sampled 100 times. A Wilcoxon signed-rank test is performed between each pair of boxes and the respective p-value is shown if it is lower than $\frac{0.05}{m}$ according to the Bonferroni correction, where m is the number of pairwise comparisons performed for each model.

level measured by (4.18) (without the modulus operator). We compare our proposed solution with CASPER, RND-P and the original values of ΔNDCG.

Age. The Figure 4.7 shows the ΔNDCG distribution across the subgroups by the demographic groups of "age". For each boxplot, on top of it, we include the Wilcoxon signed-rank test p-value significance of the difference between the means of each distributions pair (it is included IFF the p-values are lower than $\frac{0.05}{m}$ according to the Bonferroni correction, where m is the number of pairwise comparisons performed for each model). On the bottom of the plots, we include the NDCG@10 after the perturbation and the percentage of deleted edges. First, we can see how in ML-1M and FENG, our methods significantly narrow the ΔNDCG distribution of the age subgroups compared with `NP` by perturbing just 1% of the edges in some settings. The single edge deleted by `CASPER` minimally affects the recommendations of all the experiments, with more prominent effects on GCMC under several datasets, e.g., ML-1M, FENG. On the other hand, the perturbation applied by `RND-P`, in some cases, can generate a decrease in unfairness, but the early stopping prevents it from removing too many edges, which highlights its inefficiency in reducing ΔNDCG.

Among our methods, GNNUERS+CN reports the least perturbations and generates competitive decreases in ΔNDCG.

On INS, no explanation method is consistent in reducing ΔNDCG. This means that the differences in performance between the demographics groups derived from "age" are not affected by the perturbation of individual interactions, but other aspects, e.g., specific nodes, caused the original disparity. Generally, we can see how both GNNUERS and GNNUERS+CN, are able to significantly reduce ΔNDCG in most of the settings, selecting small subsets of deleted edges. Interestingly, in some cases, the NDCG@10 drops significantly (LFM-1K and FENG), while in others remain consistent (ML-1M). This means that our algorithms are able to generate explanations of the unfairness by detecting edges that contribute significantly to increase unfairness and to improve performances for only one subgroup.

Gender. GNNUERS generates a significant decrease in ΔNDCG also for the subgroups generated by the attribute "gender", as shown in Figure 4.8. In ML-1M and LFM-1K, GNNUERS significantly explains unfairness for all the models, reducing ΔNDCG by a relatively lower amount compared to the same experiments on age groups, in particular on LFM-1K. Perturbing a single edge (CASPER) or more in a random way (RND-P) does not decrease ΔNDCG in these cases (except for GCMC on LFM-1K), while our method has proven to be effective regardless of the sensitive attribute that defines the demographic groups. On INS, differently from what seen before, GNNUERS can reduce unfairness between gender groups, by deleting a relevant higher number of edges compared to other experiments. This result emphasizes how crucial is to select the right demographic attribute affecting the results.

Impact on Recommendation Utility

GNNUERS is devised to minimize the gap in recommendation utility between the demographic groups, without or minimally affecting the utility for the protected group. We empirically evaluate this aspect, by examining the edges deletion impact on the utility for each group. The NDCG@10 was measured individually for both demographic groups to then averaging it by groups. The impact on recommendation utility was measured as the change in utility after applying the perturbation. To estimate this change significance, a Wilcoxon signed-rank was performed between the 100 NDCG@10 averages measured on the recommendations altered from each explanation method and the ones generated from the non-perturbed network. For this analysis we consider GNNUERS and its extended version GNNUERS+CN.

The Table 4.13 shows: the average utility (highlighted for the unprotected group) after perturbing the edges and its relative change from the original one between brackets; for each value, the symbol (*) denotes the significance of the statistical test with the 95% of confidence interval. We can see how, for any dataset and model, the NDCG change for the unprotected group is greater than the protected group one. This confirms that our algorithms can select the edges responsible for an higher utility for the unprotected group. However, also the NDCG for the protected

Table 4.13: For both protected and unprotected groups each column include the value of NDCG after applying GNNUERS and in the brackets its relative change from the original NDCG. Unproctected group values are highlighted and in italic.

	Model	Policy	Age		Gender	
			Younger	Older	Males	Females
ML-1M	GCMC	GNNUERS	*0.11* (-21.6%)*	0.11* (-11.6%)	*0.10* (-24.2%)*	0.10* (-14.6%)
		GNNUERS+CN	*0.12* (-14.0%)*	0.11* (-06.7%)	*0.11* (-16.3%)*	0.10* (-10.2%)
	LightGCN	GNNUERS	*0.11* (-21.6%)*	0.10* (-15.4%)	*0.10* (-22.8%)*	0.09* (-17.6%)
		GNNUERS+CN	*0.11* (-16.7%)*	0.10* (-10.6%)	*0.11* (-19.4%)*	0.09* (-14.1%)
	NGCF	GNNUERS	*0.12* (-11.6%)*	0.11* (-02.7%)	*0.13* (-07.8%)*	0.11* (-01.8%)
		GNNUERS+CN	*0.13* (-09.2%)*	0.12 (-00.4%)	*0.13* (-07.1%)*	0.11 (-01.0%)
FENG	GCMC	GNNUERS	0.01* (-67.0%)	*0.01* (-78.9%)*	-	-
		GNNUERS+CN	0.02* (-62.4%)	*0.01* (-77.6%)*	-	-
	LightGCN	GNNUERS	0.02* (-51.3%)	*0.02* (-61.2%)*	-	-
		GNNUERS+CN	0.03* (-36.7%)	*0.03* (-51.2%)*	-	-
	NGCF	GNNUERS	0.04* (-10.4%)	*0.04* (-28.1%)*	-	-
		GNNUERS+CN	0.04* (-01.9%)	*0.05* (-13.2%)*	-	-
LFM-1K	GCMC	GNNUERS	0.27* (-22.8%)	*0.26* (-40.4%)*	0.33* (-12.0%)	*0.31* (-26.4%)*
		GNNUERS+CN	0.32* (-08.0%)	*0.36* (-16.4%)*	0.35* (-05.9%)	*0.35* (-17.4%)*
	LightGCN	GNNUERS	0.34* (-05.9%)	*0.37* (-15.5%)*	0.34* (-07.9%)	*0.37* (-16.7%)*
		GNNUERS+CN	0.33* (-09.5%)	*0.37* (-15.8%)*	0.35* (-07.4%)	*0.38* (-13.9%)*
	NGCF	GNNUERS	0.35* (-01.5%)	*0.39* (-09.6%)*	0.36 (-00.9%)	*0.38* (-12.4%)*
		GNNUERS+CN	0.35 (00.2%)	*0.40* (-05.9%)*	0.37* (02.5%)	*0.42* (-04.1%)*
INS	GCMC	GNNUERS	*0.37* (-52.7%)*	0.33* (-56.8%)	*0.62* (-20.0%)*	0.60* (-13.3%)
		GNNUERS+CN	*0.69* (-10.6%)*	0.70* (-08.0%)	*0.68* (-12.4%)*	0.67* (-02.7%)
	LightGCN	GNNUERS	*0.71* (-09.9%)*	0.71* (-08.4%)	*0.69* (-12.8%)*	0.69* (-05.2%)
		GNNUERS+CN	*0.79 (00.1%)*	0.78 (00.3%)	*0.74* (-06.7%)*	0.73* (01.5%)
	NGCF	GNNUERS	0.78 (-00.3%)	*0.77* (-02.4%)*	*0.72* (-09.6%)*	0.73* (-01.2%)
		GNNUERS+CN	0.77* (-01.1%)	*0.76* (-03.8%)*	*0.73* (-08.7%)*	0.74 (00.7%)

group is affected in most of the experiments. This is because the results are model dependent, and removing edges reduces the connectivity, and then the information propagation through the GNN. Also, higher NDCG losses for the unprotected groups reflect a better unfairness explanation, as seen for all the models in FENG[12]. Based on this observation, since GNNUERS perturbations for NGCF result in the least faithful unfairness explanation w.r.t. the other models in the previous RQ, for the same model it reports the lowest loss in utility for both demographic groups. As a matter of fact, not only for NGCF the NDCG for the protected group is minimally affected, but it also increases in some settings, e.g., for males users in LFM-1K. The GNNUERS+CN policy exhibits this behavior slightly more than GNNUERS, except for FENG, where the NDCG is equal between deomgraphic groups, but GNNUERS reports an additional 10% loss in utility. Using GNNUERS+CN edges selection is then

[12]GNNUERS learning process could be stopped once a desired level of fairness or utility is reached for the explanation, depending on the application requirements.

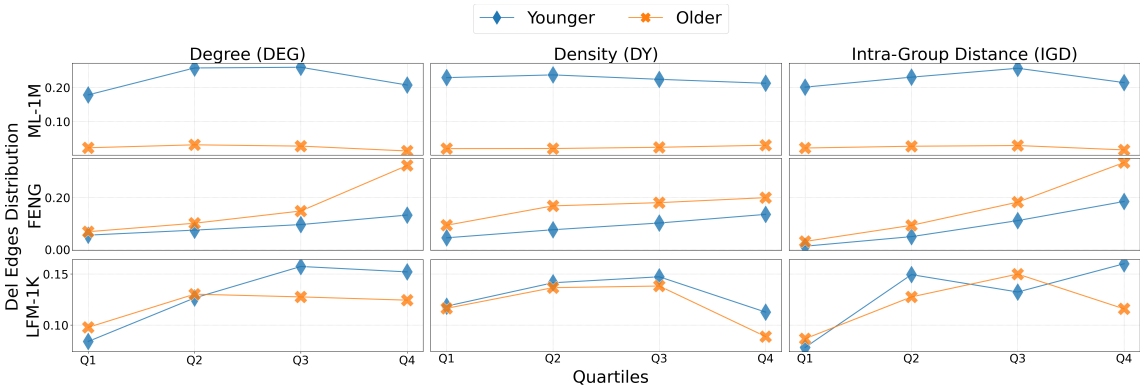


Figure 4.9: Deleted edges distribution (*Del Edges Distribution*) over the quartiles (Q1-Q2-Q3-Q4) defined for each *age* group through sorting the nodes by each graph property. The edges were deleted applying `GNNUERS` on NGCF.

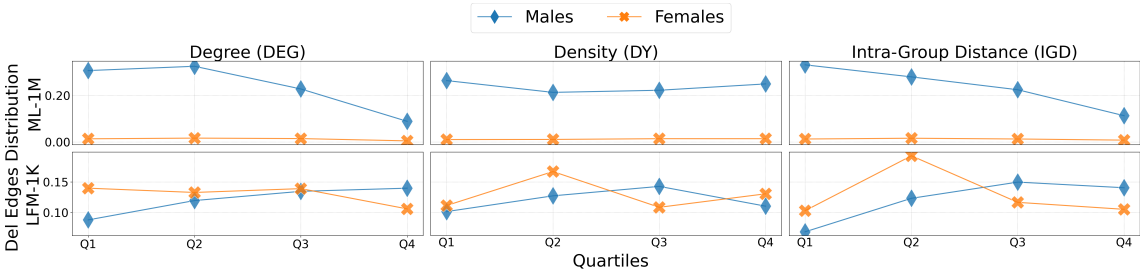


Figure 4.10: Deleted edges distribution (*Del Edges Distribution*) over the quartiles (Q1-Q2-Q3-Q4) defined for each *gender* group through sorting the nodes by each graph property. The edges were deleted applying `GNNUERS` on NGCF.

beneficial to reduce the impact on the NDCG for the protected group.

Edges Selection Process

Despite the edges selected by `GNNUERS` represent themselves the counterfactual explanations of the utility disparity across demographic groups, the edges taken as they are do not provide sufficient information to fully understand the graph features that contribute to the unfairness. To this purpose, we leverage the user nodes connected to such edges and categorize them by the properties defined in Section 4.4.2. Specifically, user nodes were distinguished by demographic groups (*Males*, *Females*, *Younger*, *Older*) and the nodes in each group were partitioned in quartiles: the order of the data points was defined by the value of each graph property that characterizes the users in each quartile, e.g., low-DEG nodes (Q1). Thus, the graph properties were individually measured for each demographic group by defining the subgroups *Males*, *Females*, *Younger*, *Older* as the set Z one at a time, e.g., if Z is the subgroup of *Males*, IGD_z represents how the male $z \in Z$ is close to the other

males $z' \in Z/\{z\}$.

For each demographic group and each quartile, the number of perturbed edges was taken and normalized by the total number of edges perturbed in each experiment. Therefore, the perturbed edges can be categorized as well based on the extent such edges are distributed over the quartiles for each graph property. A higher distribution over nodes with low or high DEG, DY, IGD levels highlights the aspects that drove our explanation method to perturb the given edges, and, therefore, highlights the aspects that led to the unfairness in the original recommendations.

`GNNUERS+CN` was not included in the following experiments because the edges deletion constraint could only reflect the distribution over quartiles defined for the unprotected group. RQ1 reported FENG, ML-1M, LFM-1K as the datasets on which our method consistently explain unfairness by reducing the disparity in NDCG between the demographic groups derived by both "gender" and "age". RQ2 highlighted NGCF as the model where GNNUERS can explain unfairness while affecting the recommendation utility of the protected group the least. `GNNUERS` selection process will be analyzed only on the experiments with these settings, as they would better reflect the properties characterizing biased recommendations.

Age. For each graph property, the Figure 4.9 reports the distribution of the edges deleted by `GNNUERS` over the quartiles, which are defined for each demographic group derived from "age". Except for LFM-1K, the distribution of deleted edges is significantly higher for the unprotected groups (ML-1M: Younger, FENG: Older), which highlights that the unfair recommendations are highly dependent on the unprotected users' interactions. On ML-1M, the over-representation and over-DEG of the unprotected group is enough to deviate the recommendations in their favor. No relevant pattern is reported across quartiles, but, on the basis of the unbalanced groups' representation, the nodes themselves could be related to the unfairness: `GNNUERS` explains it by consistently pruning more their edges compared to the protected group ones. For both FENG and LFM-1K, `GNNUERS` deletes more edges connected to high-DEG and high-IGD unprotected users (LFM-1K: Younger, FENG: Older), i.e. at the last quartile (Q4), which represents those unprotected users with the most interactions and the closest ones to other unprotected user nodes. According to Table 4.12, Older users in FENG are less represented, but their higher average DEG and IGD reflects the observations derived from Figure 4.9. On LFM-1K, the edges of the unprotected user nodes selected by `GNNUERS` are also characterized by a high DY (Q4), that is regarding users connected with more popular artists (items) on average compared to the protected users. Indeed, this category (high DEG, DY, IGD on Table 4.12) of Younger users could have increased the popularity of such connected artists, causing them to be more recommended to Younger in lower quartiles and increasing their recommendations utility.

Gender. Figure 4.10 reports the distribution of the deleted edges over the graph properties quartiles, defined for each gender group. FENG does not include gender labels, so it could not be included in this figure. On ML-1M, the significant difference of deleted edges between Males and Females denotes the extent the un-

fairness is caused by an over-representation of the unprotected group (Males), in terms of nodes and edges, i.e. over-DEG, according to Table 4.12. Additionally, unfairness seems related to the Males nodes characterized by a low DEG and IGD (Q1). Hence, `GNNUERS` uncovers unfairness as the connections regarding isolated (low IGD) Males who prefer not to watch many movies (low DEG). The deleted edges distribution over the latter is slightly higher towards the lowest (DY Q1) and highest (DY Q4) popularity levels, restricting the `GNNUERS` edges selection to isolated Males connected with just a few of mostly niche or mainstream movies. On LFM-1K, `GNNUERS` perturbes the interactions applying a dual effect. Except for DY, lower values (Q1) of the graph properties report a higher amount of deleted edges connected to the unprotected (Females) nodes, while the nodes with higher values (Q4) that lost more edges are the protected (Males) ones. Unfairness explanation is then accomplished by the simultaneous perturbation of low-DEG, low-IGD Females and high-DEG, high-IGD Males. Removing the interactions of isolated (low IGD) Females listening to a few artists (low DEG) could drastically impact the fairness levels, causing `GNNUERS` to inevitably perturb edges connected to nodes with the opposite properties, i.e. high-IGD, high-DEG Males. The fact that Males report a higher Gini on DY compared to Females suggests the former tend more to interact with artists of diverse popularity, which emphasizes how the modification of the artists' popularity impacts the recommendations.

4.5 Unfairness Mitigation via Graph Augmentation

This section describes our work aimed to mitigate recommendation unfairness across demographic groups in GNN-based systems by augmenting the graph data. The previous section introduced `GNNUERS`, that perturbs the bipartite graph representing the user-item interactions network to detect a set of edges as an explanation of the unfairness in recommendation utility across demographic groups. Here we re-formulate the method such that it does not perturb existing edges in the graph, but it generates a new augmented graph by adding user-item interactions not previously included in the dataset. These new edges added to the graph are selected with the goal of mitigating unfairness when the augmented graph is fed in input to the GNN-based system during inference. Even though not specifically defined in the original work, we denote this method as `BiGA`, which stands for **Bi**partite **G**raph **A**ugmenter. We present `BiGA` through the main modification with respect to `GNNUERS` and an additional set of policies aimed to focus on a specific subset of interactions on the basis of relevant aspects regarding recommendation, GNNs and behaviors characterizing each demographic group. Specifically:

1. We describe the re-formulated *perturbation mechanism* that finds new user-item interactions to add to the dataset in the form of graph edges. Driven

by the goal of reducing the disparity in recommendation utility across demographic groups, the perturbation mechanism, along with the optimization function, leverages the users preference information from a subset of data not used for evaluation to guarantee the edges are added under a realistic setting.

2. We devised a set of *sampling policies* to guide the edges selection process towards a smaller set of user-item interactions to add. Grounded to demographic groups behavioral aspects, popular issues studies in recommendation and the aggregation mechanism used in GNNs, the policies enable our method to follow a clearer path in the augmentation phase, obstructed by the vast set of missing interactions in common sparse datasets.

4.5.1 Methodology

Perturbation Mechanism for Graph Augmentation

Similarly to what defined in Section 4.4.1, we leverage a perturbation vector p of size B to perturb the original adjacency matrix A. Entries in A are replaced by the entries in p through a function $h : \mathbb{N}^{|U|} \times \mathbb{N}^{|I|} \to \mathbb{N}_{<B}$ (given p as a 0-indexed vector), that maps the 2D indices $(u,\ i)$ of A into a 1D index for p, such that an edge $\tilde{A}_{u,i}$ is added if $p_{h(u,i)} = 1$. Formally:

$$\tilde{A}_{u,i} = \begin{cases} p_{h(u,i)} & if\ h(u,i) \in \mathbb{N}_{<B} \\ A_{u,i} & otherwise \end{cases} \tag{4.23}$$

p is derived from a real valued vector $\hat{p}$, as done in [97, 135], by applying a sigmoid transformation before rounding values ≥ 0.5 to 1 and values < 0.5 to 0. We initialize $\hat{p}_i = -5, \forall i \in [0, B)$, such that $\hat{p}_i \approx 0$ after the sigmoid transformation and it is guaranteed $\tilde{A} = A$.

Ground Truth Information for Unfairness Mitigation

The augmented graph generation follows the goal of the loss function introduced for `GNNUERS` in Section 4.4.1, i.e. reducing the disparity in NDCG across demographic groups (ΔNDCG) and leveraging the approximated version $\widehat{NDCG}$ to optimize the perturbation mechanism towards mitigating the unfairness ($\Delta\widehat{NDCG}$). Additionally, we denote the set (train or validation) from which the ground truth labels are taken to measure $\widehat{NDCG}$ during the perturbed graph generation as the *perturbation* set. Focused again on a binary settings as prior studies [9, 81, 92] and ours, we define the subsets $U_D = \{u \in U \,|\, u \in \mathcal{G}_D\}$ and $U_A = \{u \in U \,|\, u \in \mathcal{G}_A\}$, where $\mathcal{G}_D, \mathcal{G}_A$ are the *disadvantaged* and *advantaged* groups respectively. The group with lower (higher) utility on the *perturbation* set is denoted as disadvantaged (advantaged). The graph augmentation aims at increasing the utility of such group, and not reducing the advantaged group one. Thus, edges are only added to user nodes in U_D.

Sampling Policies

Even if the edges selection is guided by a fairness-aware loss function, the set of edges to perturb could be vast. The user and item nodes of this set are described by several properties, which could support or obstruct our method. For instance, in a setting affected by popularity bias, adding edges to connect disadvantaged users to popular items could positively affect their recommendation utility, but it could also increase it for the advantaged group, hence, not mitigating the bias towards the latter. Thus, we applied several sampling policies to narrow the set of edges (connected to user nodes in U_D) to be perturbed:

- **B (Base)**: the base algorithm with no sampling applied.
- **ZN (Zero NDCG)**: selects the users with no relevant items in their top-k recommendation lists, i.e. $NDCG@k = 0$.
- **LD (Low Degree)**: selects the $\Psi_U\%$ of user nodes with the lowest degree, i.e. fewest interactions in the training set.
- **S (Sparse)**: denoting a user u's *density* as the average popularity of the items u interacted with in the training set, it selects the $\Psi_U\%$ of users with the lowest *density* (highest *sparsity*), i.e. mostly interacting with niche items.
- **F (Furthest)**: selects the $\Psi_U\%$ of furthest user nodes from U_A, where the distance from $u_D \in U_D$ is computed as the shortest paths lengths average between u_D and all $u_A \in U_A$.
- **IP (Item Preference)**: following [13], we estimate the extent an item is preferred by U_D, thus, I is reduced by selecting the $\Psi_I\%$ most preferred items by the same group.

where $\Psi_U\%$, $\Psi_I\%$ denote parameters to sample the user set U or the item set I respectively. We fix $\Psi_U\% = 35\%$, $\Psi_I\% = 20\%$.

These policies were selected factoring in the way each demographic group interacts with the items (IP, S), common phenomena described in recommendation literature (ZN, LD), the aggregation operation in GNNs models (LD, F). We can distinguish between policies of type U (ZN, LD, S, F) or I (IP) if the sampling is applied on the user or item set respectively. We also contemplated inter-group combinations between policies U and I, but intra-group ones are excluded not to lead to excessive reduction of the user or item set.

4.5.2 Experimental Setup

Our experiments are based on the artifacts used to evaluate `GNNUERS`, but we reduce the set of datasets for fairness assessment to two corpora: MovieLens 1M (ML-1M) [69], and Last.FM 1K (LFM-1K) [25]. The advantaged groups, their representation

Table 4.14: Mitigation performance of our method's policies: the relative difference in ΔNDCG between the scores measured on the *perturbation* set before and after applying each policy is reported. Unfairness mitigation is represented by negative values (highlighted).

	Policy Type	Policy	GCMC		LightGCN		NGCF	
			Gender	Age	Gender	Age	Gender	Age
ML-1M		B	14.5%	15.3%	-100.0%	-74.1%	6.9%	10.0%
	U	ZN	0.0%	-98.3%	-99.3%	-33.3%	0.8%	5.5%
		LD	14.5%	15.3%	-100.0%	-74.1%	6.9%	10.0%
		S	1.7%	13.6%	-92.4%	-80.2%	11.5%	10.0%
		F	5.1%	15.3%	-95.2%	-81.5%	5.4%	9.1%
	I	IP	9.4%	544.1%	-93.1%	-92.6%	18.5%	9.1%
	U+I	ZN+IP	-53.8%	-11.9%	-88.3%	-81.5%	-100.0%	8.2%
		LD+IP	9.4%	-18.6%	-93.1%	-92.6%	18.5%	9.1%
		S+IP	7.7%	10.2%	-97.2%	-80.2%	7.7%	15.5%
		F+IP	9.4%	-50.8%	-97.9%	-66.7%	4.6%	12.7%
LFM-1K		B	-93.7%	-89.9%	164.3%	271.0%	-0.7%	-32.6%
	U	ZN	-99.7%	-92.4%	-40.9%	-97.2%	-0.7%	-49.9%
		LD	-59.2%	-84.4%	164.3%	271.0%	-0.7%	-32.6%
		S	5.8%	-93.3%	-69.3%	-80.9%	0.2%	-32.6%
		F	-95.5%	-97.3%	-58.6%	-79.4%	-0.7%	-32.6%
	I	IP	-88.4%	-94.3%	-6.4%	-0.4%	-2.4%	-34.6%
	U+I	ZN+IP	-0.3%	-96.4%	-3.5%	0.2%	0.7%	-32.6%
		LD+IP	-88.4%	-94.3%	-6.4%	-0.4%	-2.4%	-32.4%
		S+IP	-8.9%	-0.2%	-3.8%	0.9%	0.7%	-34.6%
		F+IP	1.1%	-0.4%	-3.8%	1.3%	0.7%	-34.8%

w.r.t. to the related sensitive attribute and other relevant information are depicted in Table 4.12. We also adopt the same splitting strategy, that is for each dataset we arranged the interactions list of each user in ascending order of recency, and split the sorted lists with a ratio 7:1:2 to include each subset in the train, validation, test set respectively. The validation set was used (i) to select the training epoch where the non-perturbed model reached the highest NDCG, (ii) as the *perturbation* set for our method. During the evaluation step, the edges selected by our method were added to the training set, and, if present, removed from the other two sets.

As done for `GNNUERS`, we relied on Recbole [173] and adopted the same GNNs-based models (GCMC, LigthGCN, NGCF) to solve the top-k recommendation task. We optimized the hyper-parameters under a grid search strategy.

4.5.3 Results

Edges Augmentation Analysis

If the addition of the edges selected by our method successfully mitigates the model unfairness, the characteristics of such edges, i.e. the nodes composing such edges,

could describe a possible cause of the original recommendations unfairness (before the graph was perturbed). Under a given policy, the features of the sampled nodes characterize the edges selected by our method (Section 4.5.1). To this end, Table 4.14 depicts the unfairness mitigation performance of all the policies in each setting. Such performance is the relative difference in $|\Delta\text{NDCG}|$ between the scores measured on the *perturbation* set before and after a policy was applied, where $\Delta\text{NDCG} = \overline{\text{NDCG}}_{U_D} - \overline{\text{NDCG}}_{U_A}$ is the difference between the utility mean of U_A and U_D.

Some settings are positively affected by our procedure, regardless of the given policy, though other ones can successfully be perturbed only by one or a few policies. In fact, this aspect is highlighted by `ZN+IP`, the only policy mitigating unfairness across gender groups on ML-1M for GCMC and NGCF, with a noteworthy change of -100% for the latter. While the individual policies `ZN` (the disadvantaged users provided with no relevant items out of the 10 recommended, i.e. $NDCG@10 = 0$) and `IP` (items mostly preferred by the disadvantaged users) could not report a similar result, the interactions added to the females, i.e. U_D, by these policies combination were able to reduce the gap in NDCG between gender groups.

Some policies systematically excel more than other ones under the same settings, such as `ZN` across age groups on ML-1M (GCMC), and on LFM-1K (LightGCN, NGCF). It is not obvious that adding interactions to the users selected by `ZN` could mitigate the unfairness. In particular, with an in-depth inspection, we observed that the policies successfully reducing ΔNDCG have a minimal or negligible effect on the recommendation utility of U_A, highlighting that the added edges are selected to only improve the utility of U_D.

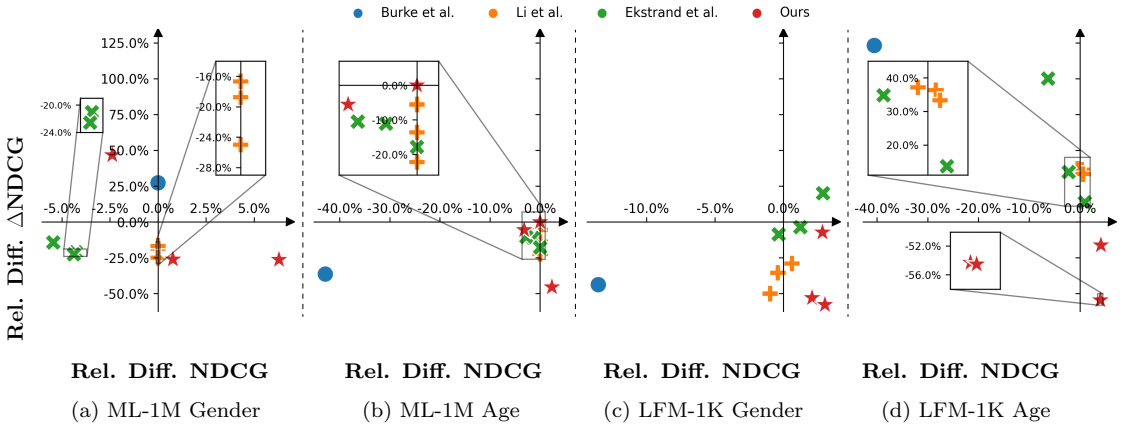
While some policies can consistently reduce ΔNDCG under the same settings (`ZN+IP` on ML-1M for gender ; `ZN`, `S`, `F` on LFM-1K for age), thus, suggesting the unfairness originates at the dataset level, other experiments underline that the bias is also included at the model level and a given policy does not work for different models (`ZN`, `F+IP` on ML-1M for age ; `LD` on LFM-1K for age).

Mitigation Procedures Comparison

In this section, we evaluate the trade-off between the recommendation utility and the unfairness mitigation performance of our method in comparison with SOTA algorithms. Selecting mitigation procedures for the recommendation unfairness on the end-user side is a non-trivial task, over-complicated by the multitude of fairness notions and evaluation protocols. Thus, based on the similarity to our evaluation protocol, we relied on the framework of our prior reproducibility study[13] presented in Section 4.3 and compared the mitigation procedures used for top-k recommendation with our method. Given our focus on the mitigation task, we only considered models reporting high utility levels[14], which could effectively solve the recommen-

[13]Experiments on LFM-1K were re-run to match our splitting strategy.

[14]We discarded LBM, STAMP, FunkSVD since they reported a NDCG lower than half of the best models one (ItemKNN for ML-1M, UserKNN for LFM-1K) for both datasets.



(a) ML-1M Gender (b) ML-1M Age (c) LFM-1K Gender (d) LFM-1K Age

Figure 4.11: On the x-axis (y-axis) the relative difference, denoted as `Rel. Diff.`, in recommendation utility NDCG (utility disparity ΔNDCG) between the scores measured before and after each method was applied on *gender* and *age* groups. Multiple points per method indicate the use of multiple models for each method. Positive (negative) values on the x-axis (y-axis) denote an increment (decrement) in NDCG (ΔNDCG).

dation task and reflect existing biases as in real-world scenarios. The GNN-based recommendation systems used with our algorithm satisfy this property. The following results regarding our method pertain to the policies that reported the lowest |ΔNDCG| on the *perturbation* set (Table 4.14).

Figure 4.11 highlights the extent to which each method affected the recommendation utility (x-axis) and the disparity in the latter between user groups (y-axis). Our approach reports the best mitigation performance on all the settings, given by the points labeled as `Ours` being systematically the lowest ones on the y-axis. Moreover, all the algorithms affect (often negatively) the recommendation utility at least for one setting, except ours, whose points never move along the x-axis. In Table 4.15 we report the resulting levels of utility (NDCG) and fairness (ΔNDCG) after each algorithm was applied. Determining which setting is the best one depends on the requirements of a specific application and to what extent the fairness is relevant. In terms of recommendation utility, the GNN-based models are slightly less effective than the best ones, but the mitigation produced by our method results in utility disparity levels significantly lower than the systems reporting the highest NDCG, e.g., LFM-1K on gender groups.

4.6 Findings and Discussion

We gather the findings of our studies in recommendation unfairness regarding the systematic reproducibility study in Section 4.3, the unfairness explainer (`GNNUERS`) in GNN-based recommender systems in Section 4.4, the mitigation procedure driven

Table 4.15: Recommendation utility (NDCG) and utility disparity (ΔNDCG) after applying each method on user groups. For each setting, the best and second best scores are in bold and italic respectively.

			NDCG ↑		ΔNDCG ↓	
	Mitigation	Model	Gender	Age	Gender	Age
ML-1M	Burke et al. [21]	SLIM-U	0.084	0.048	^0.028	^*0.014*
	Li et al. [92]	BiasedMF	0.112	0.112	^**0.013**	^0.017
		NCF	0.120	0.120	^0.015	^0.019
		PMF	0.123	0.123	^0.015	^0.021
	Ekstrand et al. [47]	ItemKNN	**0.134**	**0.138**	^0.030	^0.024
		TopPopular	0.104	0.107	^0.030	^0.034
		UserKNN	*0.131*	*0.137*	^0.024	^0.023
	Ours	GCMC	0.126	0.126	^0.022	^0.017
		LightGCN	0.127	0.127	-*0.014*	^**0.012**
		NGCF	0.129	0.129	^0.017	^0.023
LFM-1K	Burke et al. [21]	SLIM-U	0.301	0.207	^0.072	^-0.145
	Li et al. [92]	BiasedMF	0.245	0.247	*-0.049	*-0.060
		NCF	0.202	0.203	**-0.023**	-0.048
		PMF	0.164	0.164	*-0.049	^-0.044
	Ekstrand et al. [47]	ItemKNN	0.286	0.269	*-0.116	**0.020**
		TopPopular	0.321	0.315	*-0.102	-0.050
		UserKNN	**0.411**	**0.397**	^-0.106	-0.031
	Ours	GCMC	0.384	0.384	-*0.024*	-0.039
		LightGCN	*0.397*	**0.397**	-0.030	-*0.030*
		NGCF	0.387	*0.387*	*-0.077	-0.062

by a bipartite graph augmenter in Section 4.5:

- In general, the reproduced mitigation procedures did not substantially impact on recommendation utility, regardless of the sensitive attribute, dataset, task. The impact is larger in LFM 1K than ML 1M.

- Unfairness depends on the mitigation, model, and fairness notion. Often the mitigation impact is small. Lowering DP does not imply lowering KS, and vice versa. Unfairness is higher in LFM 1K than ML 1M.

- The disparate impact does not always harm the minority group. The latter was advantaged for both attributes in LFM 1K (TR), in both datasets for age and in LFM 1K for gender (RP).

- Except extreme cases, `GNNUERS` selects edges that systematically and significantly explain unfairness, regardless of the data, models and demographic groups on which is applied.

- `GNNUERS` and `GNNUERS+CN` reduce the utility for unprotected groups, detecting

Table 4.16: Summary categorization of the considered mitigation procedures with respect to the proposed properties.

Paper	Applicability	Coherence	Consistency	Data Robustness	Reproducibility	Scalability	Trade-off	Transferability
Ekstrand et al. [47]	Higher	Higher	Higher	Higher	Higher	Higher	Higher	Lower
Li et al. (A) [92]	Higher	Lower	Higher	Higher	Higher	Lower	Higher	Higher
Frisch et al. [58]	Lower	Higher	Higher	Lower	Higher	Lower	Lower	-
Burke et al. [21]	Lower	Lower	Lower	Lower	Higher	Lower	Lower	-
Tsintzou et al. [141]	Higher	-	-	-	Lower	Higher	-	-
Li et al. (B) [93]	Higher	-	-	-	Higher	Lower	-	-
Wu et al. (A) [154]	Higher	-	-	-	Higher	Lower	-	-
Wu et al. (B) [155]	Higher	-	-	-	Lower	Lower	-	-

edges that generated a disparity in performance, while reporting a negligible loss for the protected one.

- `GNNUERS` edges selection process is significantly affected by the dataset domain. Experiments uncovered an unfairness mainly related to differences between demographic groups in interest (DEG) and closeness (IGD).
- Some sampling policies suggest the unfairness originates at the dataset level by consistently reducing the disparity in recommendation utility, while other experiments underline that the bias is also included at the model level and a given policy does not work for different models.
- Our unfairness mitigation algorithm offers a strong balance between utility and fairness and demonstrates greater reliability in mitigating unfairness than the other approaches.

Despite our work aimed to assess, explain and mitigate unfairness in recommendation, several issues are still open in the literature. In particular, several challenges emerged while conducting the systematic reproducibility study in Section 4.3. In particular, (i) the code base modularity should be improved to easily accommodate different datasets as an input, (ii) many procedures required extensive computational resources to treat the recommendation models, (iii) the evaluation settings were often different. Our reproducibility study shows the first attempt of comparing a wide range of unfairness mitigation procedures under the same evaluation protocol, considering two relevant yet transferable fairness notions. As a summary, Table 4.16 provides a categorization of the considered mitigation procedures with respect to the proposed properties. For each property and mitigation procedure, we assigned one of the two following labels: `Higher` when the corresponding work was better than the others on average for the selected property, `Lower` otherwise. A blank entry was left for studies that could not be evaluated in terms of the corresponding property. For papers whose source code was not available, the corresponding mitigation procedure was only analyzed in terms of *applicability* and *scalability*. Our findings are expected to represent a guideline for researchers working on mitigating consumer unfairness.

Chapter 5

Conclusions

In this thesis, we investigated the assessment, mitigation and explanation of unfairness in artificial intelligence technologies to prevent such an issue from real-world applications. In particular, we contextualized existing methods in the literature, devised novel methods to mitigate unfairness and frameworks to highlight specific data patterns as an explanation of the issue under consideration. *Data Balancing*, *Counterfactuality*, and *Graph Neural Networks Explainability* methods have made it possible to better understand the unfairness issue and efficiently counteract it.

5.1 Contributions Summary

The research under this thesis has been proved to provide the following contributions:

- The fairness-aware speaker recognition framework proposed in Chapter 3 provides tools to experiment with fairness in the field. To our knowledge, it is the first framework providing functionalities to mitigate and explain unfairness issues across users' groups in speaker recognition.

- The reproducibility study conducted in Chapter 4 and the practical perspectives defined for unfairness mitigation methods lay the foundation for fairness-aware analysis in recommendation under a common protocol. Researchers can leverage the reproduced works and the experimental settings of our studies to benchmark their methods with state-of-the-art algorithms and to adopt a comprehensive evaluation protocol for fairness analysis.

- The explainability framework proposed in Chapter 4 can identify the relationship between user-item interactions and unfairness. This information can be leveraged by system designers not only to better understand the aspects characterizing this issue, but also to mitigate it.

5.2 Limitations and Open Issues

The contributions provided by the research work undertaken in this thesis represent significant advancements for the literature, but they still present several limitations, which give rise to new open issues:

- The fairness-aware framework proposed in Chapter 3 for speaker recognition includes an unfairness mitigation method based on a data balancing approach. However, several studies in machine learning proved that the models can affect the unfairness measured in the outcomes, highlighting how counteracting unfairness from the sole *data viewpoint* is not enough to guarantee fair results.

- The same framework for speaker recognition includes a surrogate model to examine the influence of vocal characteristics on the outcomes unfairness. The speech covariates discovered as relevant to explain the estimated unfairness are related to personal and sensitive characteristics, which highlight the need to devise unfairness mitigation methods accounting for causally-related features.

- The reproducibility study conducted in Chapter 4 for unfairness mitigation methods in recommendation provides an extended overview of the state-of-the-art results. However, several techniques require tailored experiments and the common protocol in our study does not provide a thorough analysis of their performance. Additionally, the limited datasets with sensitive attributes do not offer a comprehensive examination of each method capabilities.

- The explainability framework introduced in Chapter 4 is based on GNNs for recommendation to identify specific edges of the graph as negatively influential on the recommendation unfairness. The literature includes several families of recommender systems, and only a few are based on GNNs, which limits the adoption of our method to other models. Additionally, the mitigation technique based on our explanation algorithm could be hindered by models deeper than the ones tested in our study, given that the modifications applied on the graph topology could have a minor influence on the prediction process.

5.3 Future Works

The limitations described in the previous section shed light on new and additional works for future investigations:

- **Large and Multi-Domain Datasets**. Available datasets in speaker recognition and recommendation include a good amount of users and sensitive attributes to work with. Still, systems running in production have been trained with an extreme amount of data, typically proprietary. Research in artificial

intelligence, in particular on fairness, is then limited by the size of the available corpora, but also by their nature. Indeed, as far as we know, a dataset providing speakers' utterances and recommended items based on a vocal query are not publicly available, but it is highly probable that companies producing voice assistants could benefit from data of this nature. It is then fundamental to gather data to support additional and specific tasks as recommending items personalized on the speaker identity and query utterance.

- **Transferability across Models**. Works proposing novel methods to mitigate unfairness are often devised or tested only a limited set of models. Their transferability to systems with a different architecture or more powerful is not well examined. It is then unclear if such methods could be adopted in practice due to the incompatibility with the models, insufficient execution times, unreliability on large amount of data. Novel mitigation algorithms should then be devised taking into account several properties, among which fairness remains the main one.

- **Enhanced Unfairness Mitigation**. Several methods proposed to mitigate unfairness in speaker recognition and recommendation overlook the model influence on the unfairness, Indeed, several algorithms focus on modifying the data fed in input to the AI systems, but such modifications are performed without accounting for their impact on the model. Even though the unfairness could originate from the data characteristics, the model is significantly responsible for the prediction process and different models could report different levels of unfairness with the same data. It follows that novel mitigation methods should learn to generate fairer outcomes by also learning the systems influence on the resulting fairness.

- **Generalized Explainability**. The research into AI explainability, focused on improving the interpretability of deeper and sophisticated systems, has been increasing in recent years. Particularly on recommendation, explanations tools are mainly proposed to clarify why a certain item is recommended to a specific users. Meanwhile, several works study properties that go beyond the mere predictions process, such as novelty, coverage, serendipity, and fairness. Future research will aim to devise new explanations tools that could generalize to multiple properties, instead of providing explanations for an individual aspect.

Based on the rapid evolution of speaker recognition and recommendation over these years, we are hopeful that more attention will be dedicated to systems combining the power of both technologies to solve more specialized tasks. Although the single technologies are individually and thoroughly studied in the literature, the analyzed process of vocal assistants requires to consider multiple aspects during its design, and unfairness should be counteracted on each step of the pipeline.

Appendix A

Fairness in Therapeutic Counseling

A.1 Introduction

The recent pandemic has forced people to confine themselves to four walls and has limited human interactions. Psychological issues such as anxiety and depression have surfaced due to this prolonged situation [12]. The pandemic also emphasized the importance of psychologists and remote therapy access. These services are cost-intensive, and people are forced to make sacrifices or wait for bonus payments issued by the governments. Furthermore, common people do not have familiarity and awareness of mental health issues, which puts them at risk of incurring misinformative content publicly available online [112, 113].

The motivation of this study is to facilitate individuals in autonomously evaluating therapeutical content personalized for their psychological issues. This objective is inspired by a publication over 20 years old [113] that emphasized the potential benefits that information retrieval (IR) methods could offer to both patients and therapists. IR can help in addressing mental health issues by efficiently providing relevant and reliable information, personalizing treatment options based on specific needs and preferences, and monitoring mental health trends in the population.

In this chapter, we aim to examine the application of IR tools to rank relevant therapy counseling sessions according to the psychological disease a patient aims to solve. Nonetheless, applying automatic decision-making systems to data that includes such sensitive attributes raises the need to account for unfairness issues, being fairness one of the key requirements artificial intelligence (AI) systems should meet according to the European Commission [23]. AI responsibility towards the entity receiving the ranked content is especially studied in the IR subfield of recommendation systems. Researchers have faced beyond-accuracy issues, such as explainability [10] and fairness [18, 19, 102], to make these systems more trustworthy. Along the same lines, ranking methods for therapeutical content should be examined to assess that the IR task is performed fairly across psychological diseases.

A.2 Literature Review

Literature in healthcare focused on mental health issues and IR applied on such domain is limited and it presents several limitations. First, despite the massive amount of data that humanity produces, ironically, there is a scarcity of publicly accessible data in healthcare and its sub-domains, e.g., mental health. Second, publications regarding IR in healthcare do not analyze psychological diseases in detail, but IR models applied to ranking tasks with large datasets, e.g., PubMed [98], could also handle data related to mental health issues.

To bridge the former gap, researchers have recently worked towards releasing freely accessible datasets in the psychological domain. The work in [123] proposed a new benchmark for empathetic dialogue generation and released *EMPATHETICDIALOGUES*, a dataset containing 224,850 conversations grounded in emotional situations and gathered from 810 different participants. [158] released AnnoMI, a dataset of conversational therapy sessions annotated by experts, composed of 110 high-quality and 23 low-quality Motivational Interviewing (MI) dialogues.

The second issue arises because psychological diseases are mentioned as part of the wider and general healthcare field. The community started to recognize healthcare disparities in the early 2000s [28]. Driven by the fairness key requirements set up by the European Commission [23], recent works in healthcare [109, 35, 40, 96] addressed unfairness issues of AI outcomes, e.g., in neural medicine and kidney function estimation. In the mental health domain, [132, 152] examined unfair gaps in access and questionable diagnostic practices against racial and ethnic groups, e.g., African Americans receiving a medical prescription less likely than Whites.

A.3 Methodology

Problem Formulation

We model an IR task where documents are represented as therapy sessions and queries as psychological topics. Let S be the set of queries, i.e., topics, T the set of documents, i.e. therapy sessions, $Y = \{0, 1\}$ the set of relevance labels, which represent the extent to which a document is relevant for a specific query, where higher values denote higher relevance. In particular, given $s_i \in S$, a subset of documents $T_i = \{t_{i,1}, t_{i,2}, ..., t_{i,n}\}$, sorted in descending order by relevance, is retrieved to satisfy s_i. The relevance of each document in T_i is denoted by $y_i = \{y_{i,1}, y_{i,2}, ..., y_{i,n}\}$, such that $y_{i,j}$ is the relevance degree of $t_{i,j}$ for s_i.

A ranking model $\mathcal{F} : S \times T \rightarrow \hat{Y}$ takes queries and documents as input and predicts the relevance score $\hat{y}$ for each query-document pair. Therefore, training a ranking model becomes an optimization problem. Given relevance-graded query-document pairs, this means finding the model hyper-parameters θ that minimize

the following objective function:

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\mathcal{F}, s_i, t_{i,j}, y_{i,j}) \tag{A.1}$$

Fairness Definition

A relevant amount of the IR literature analyzes fairness regarding the entities being ranked [122, 62, 63]. Here we focus on the impact of ranking utility disparity on psychological topics, which are considered proxies of psychological diseases. The therapeutical content being retrieved should help and support any query with the same level of therapy quality, regardless of the psychological disease being searched. To this purpose, we operationalized the fairness definition of *Demographic/Statistical Parity*. This definition is satisfied if the likelihood of relevant content is the same for any psychological disease. Following other works focusing on the entity receiving the ranked lists in IR [155, 18], unfairness is assessed as the disparity of ranking utility across consumers. Let the Normalized Discounted Cumulative Gain (NDCG@k) be the ranking utility metric, the disparity across topics D_S is measured as the average pairwise absolute difference between NDCG@k values and defined as follows:

$$D_S = \frac{1}{\binom{|S|}{2}} \sum_{1 \leq i < j \leq |S|} \|NDCG_i@k - NDCG_j@k\|_2^2 \tag{A.2}$$

A.4 Experimental Settings

Dataset

For our work, we have used AnnoMI[1] [158], a high-quality dataset of expert-annotated MI transcripts of 133 therapy sessions distributed over 44 topics, e.g., smoking cessation, anxiety management, and 9695 utterances. Dialogues comprise several utterances of the therapists and the patients, labeled according to the therapist's approach, e.g., question, reflection, and the patient's reaction, e.g., neutral, change. AnnoMI is the first dataset of its kind to be publicly accessible in the psychology domain, which is suitable for several tasks.

Our work focuses on providing therapeutical content in response to a particular psychological disease, then we have only considered the therapist dialogue sessions and, thus, the AnnoMI therapist's utterances. However, due to the low number of therapy sessions per topic, we reduced the ranking task to use only the therapist's utterances text in each session. Nonetheless, the quality of each therapist's utterance text is considered to be the same as the therapy session it belongs to, given that the quality is dependent on the utterances each session is made up of.

[1]Data available at `https://github.com/vsrana-ai/AnnoMI`

We also re-labeled the topics of the dataset to reduce the original set. As a result of an exploratory analysis supported by a psychology researcher, we aggregated the topics based on (i) similarities between the psychological diseases being treated and (ii) families of topics that could be envisioned in a more general group.

(i) several topics regard an equal or similar problem, but, in some cases, they are associated with other aspects mainly related to the central topic, e.g., the topics *weight loss; diet* and *weight loss* were merged into the single topic *diet and weight loss management*

(ii) some new groups can be defined to include several topics to represent a broader psychological problem that connects all the considered topics, e.g., the new topic *motivated towards adopting better life style* envisions the same issue depicted by *increasing self-confidence* and *reducing violence*

The two AnnoMI topics *birth control* and *opening up* were instead discarded because they did not fit the new topics labeling. 51 utterances were then removed from the dataset, which does not affect the integrity of the whole corpus.

Data Preparation

To make each utterance text reflect as much as possible the quality of the respective therapy session, we filtered out the utterance with less than 5 tokens, leading to a total of 3984 utterances. The value was selected empirically to remove most of the classic expressions not related to the respective topic, e.g., "Okay, all right.".

Given the thorny type of sensitive data, it is fundamental to assess the outcomes fairness of the rankers used in our experiments. Based on the definition in Section A.3, the dataset is split into training and testing sets following the ratio 80:20, respectively, such that the distribution of each psychological disease follows the same ratio across the sets. Furthermore, the topics not representing both the low and high MI dialogues quality were dropped, otherwise the disparities in ranking utility could be misinterpreted as an unfairness issue.

Table A.1 lists the remaining topics and their distribution over the high (1) and low (0) *MI quality* classes characterizing the 3984 utterances.

Information Retrieval models

A simple IR system takes a query as input and retrieves a list of documents ranked by the predicted relevance. We use as a query the name of the psychological disease being treated in a therapeutical counseling session, e.g., "diabetes management."

To analyze the applicability of IR tasks in the psychological domain, we use a high diversity of neural rankers, selected to cover different levels of network complexity. Arc-I [73], Arc-II [73], DRMMTKS [163], DUET [110], Dense Baseline [66],

Table A.1: Distribution of topics over each *MI quality* class.

	MI quality	
Topic (Psychological Disease)	Low (0)	High (1)
adhering to medical procedure (AMP)	2.36%	7.83%
asthma management (AM)	0.66%	3.12%
compliance with rules (CR)	0.43%	9.50%
diabetes management (DM)	0.18%	11.45%
managing life (MF)	0.86%	12.18%
reducing alcohol consumption (RAC)	7.14%	24.95%
reducing alcohol consumption—smoking cessation (RAC—SC)	0.51%	0.66%
smoking cessation (SC)	3.70%	14.49%

HBMP [139], KNRM [160], Naive [66] are some of the models provided in the MatchZoo Python library[2]. TFR BERT [68] is a ranker based on the popular BERT proposed by TensorFlow and represents the most complex network in this study. All the models were optimized with a Softmax cross entropy loss and the selected hyperparameters can be found in the linked source code. We selected the best predictions generated by each model based on the NDCG@5 on the training set.

A.5 Results

Ranking in Psychological Domain

Research in IR applied in the psychological domain, more specifically in therapeutical counseling, is hard to be found in healthcare literature. As far as our knowledge, this is the first paper addressing the problem of ranking therapeutical documents to support users with psychological diseases. To this end, we first investigated whether the rankings produced by IR techniques guarantee a good level of utility on therapeutical counseling data.

Analyzing the ranking utility over different relevance levels could provide insights into the decision-making process of each ranker. A model reporting utility values quite different across relevance levels could be affected by the similarity between the utterance text and the queried topic, instead of finding a pattern that connects high-quality utterances to the relative psychological disease. Under this viewpoint, Figure A.1 depicts the distribution of NDCG@k, $k \in \{3, 5, 10\}$ over topics, i.e. queries, measured on the rankings generated by the trained models. The performance of most of the models is positively affected as the relevance level increases, with higher medians and narrower ranges. Conversely, two of the most performing rankers, *DRMMTKS* and *TFR BERT*, are robust against the k used to measure the NDCG,

[2] `https://github.com/NTMC-Community/MatchZoo`

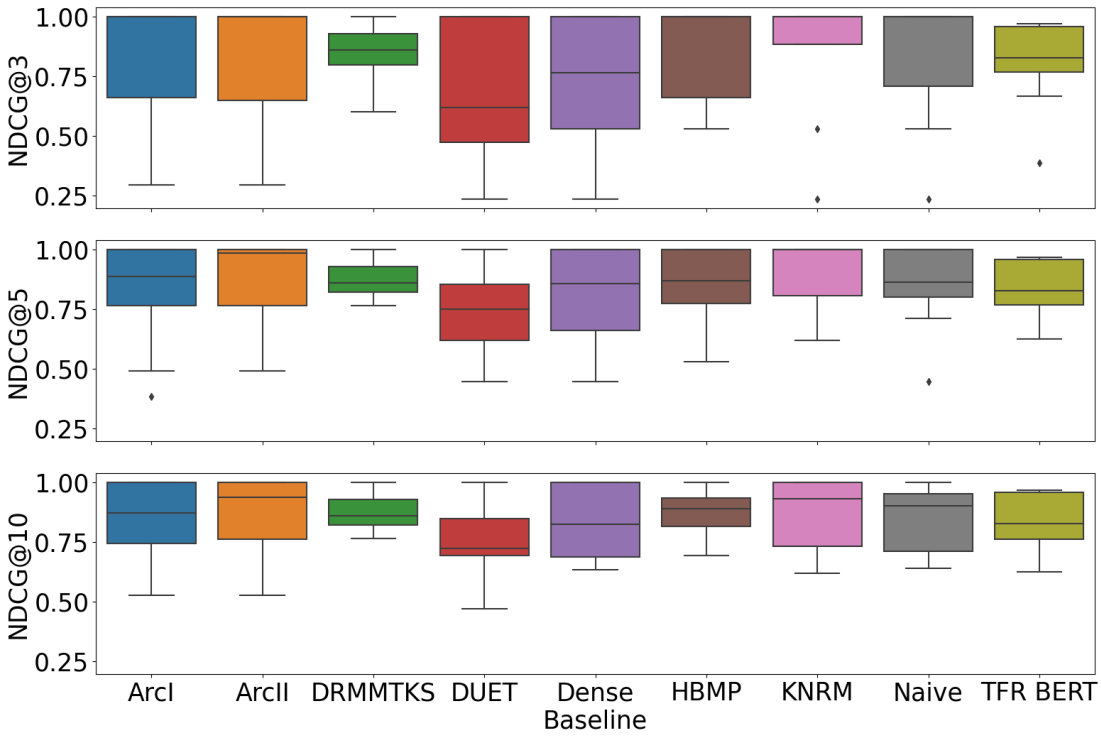


Figure A.1: Distribution of NDCG@k, $k \in \{3, 5, 10\}$ across topics for each ranker. Higher NDCG denotes higher utility.

except for one outlier when the top 3 results are considered. The rankings generation by these two models seems then more reliable because their predictions better map the relation between the queried diseases and the *MI quality* of therapeutical content.

Other than that, even if the complexity of the underlying structure is significantly different across models, there is no evidence of a relationship between higher model complexity and higher NDCG average. For instance, *TFR-BERT*, one of the most complex rankers, reports a slightly lower performance than the simpler *Naive*. Nonetheless, the NDCG@10 distribution across topics for all models spans ranges higher than 0.5, and, considering the over-representation of the high *MI quality* class, it is reasonable to obtain such measures.

We can then positively answer our first research question, with all the models reporting rankings of high utility on therapeutical counseling data.

Unfairness Levels across Topics

In our evaluation protocol, a query represents a psychological disease, which directly reflects the patients that suffer from it and that look for support to treat it. Hence, the decision-making process that generates a ranked list of therapeutical content affects all the patients that could be helped by the retrieved information. To ensure

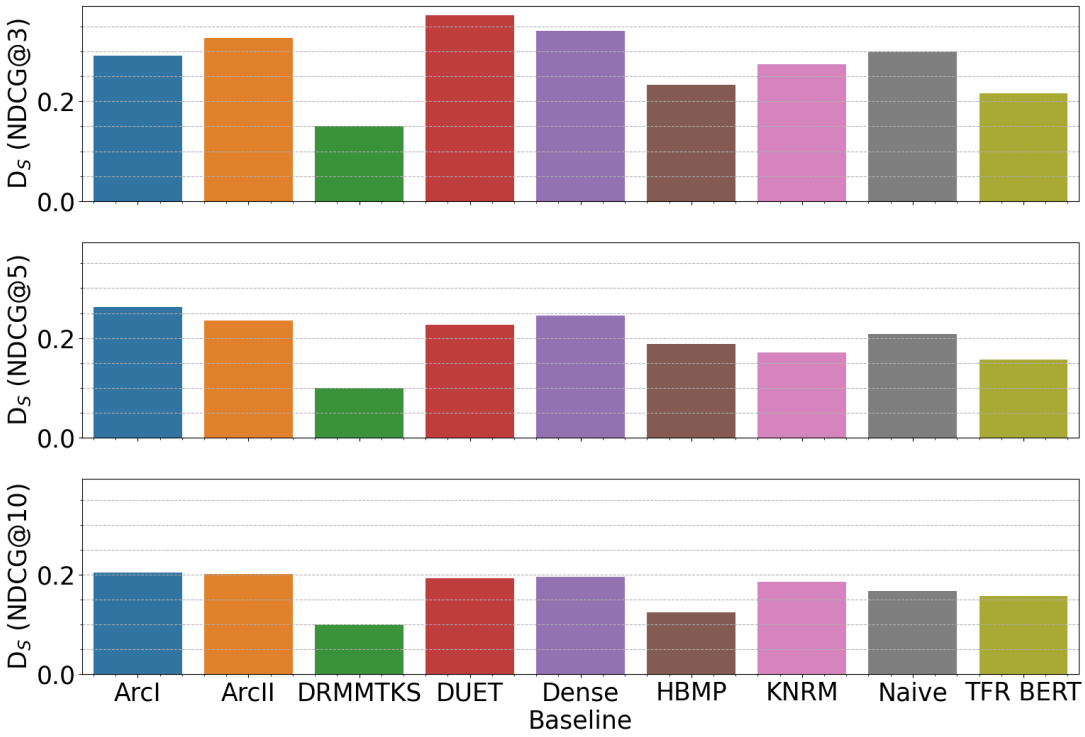


Figure A.2: D_S, the average pairwise absolute difference between utility values, measured on the rankings provided for each psychological topic at different relevance levels. Lower values are fairer.

that therapeutic content is accessible and effective for patients with diverse mental health needs, it is crucial to investigate whether ranking algorithms exhibit bias towards specific psychological diseases. Such bias may result in negative impacts on the reliability of ranking outcomes, and limit the quality of therapeutical content provided to patients.

The operationalized fairness notion in (A.2) is then used to measure the extent to which the ranking utility for each topic differs from the others in a pairwise fashion. Figure A.2 shows the unfairness level of each ranker in terms of NDCG disparity at different k of relevance. *DRMMTKS*, *TFR BERT*, *HBMP* report the best fairness degree for each top-k, while *DUET*, *ArcI* trade places as the unfairest model at different relevance levels.

Even though these rankers are suitable to be applied to psychological data in terms of utility, they do not seem to be reliable to provide fair rankings. Among all the models, *DRMMTKS* reports the lowest and most stable D_S across the different relevance levels. Still, the NDCG disparity between each pair of topics is close to 0.1 on average, which systematically highlights a different distribution of high- and low-quality therapeutical content in the top-k lists across queries. Besides,

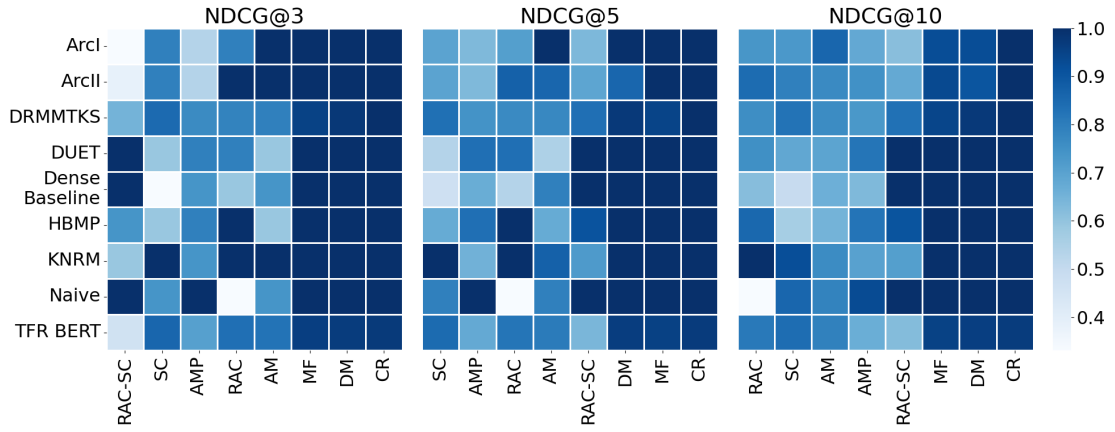


Figure A.3: Each cell represents the NDCG measured on the ranking generated for any queried topic (column) and any ranker (row) at different relevance levels. The columns are sorted by NDCG average over the columns. Darker values denote higher utility.

the unfairness levels of the top-3 ranking lists uncover a significant utility variance. Being the relevance binary, this means that some queries result in at least 1 low-quality document out of 3 in the list, which could potentially harm the patients suffering from the queried mental issue.

These observations remark on the importance of examining the model outcomes under a fairness definition and taking countermeasures to mitigate these issues, especially in domains with such sensitive data.

Systematic Negative Impact

The previous research question uncovered a significant unfairness issue, reporting most of the rankers as not being able to provide similar utility values for all the topics. However, we have no insights into the utility of the rankings provided for each individual topic and we do not know whether querying specific psychological diseases leads to higher utility rankings w.r.t. to other ones. It is then important to conduct individualized analyses of each query to uncover potential factors that may be correlated with ranking performance.

Figure A.3 aggregates all the models and psychological topics in three heatmaps (one for each relevance degree), where each cell represents the NDCG measured on the ranked list provided by a ranker (row) for an individual queried topic (column). Being the columns sorted in ascending order from left to right by NDCG average across models, it is straightforward to notice how the darker cells concentrate on the few topics at the right of each heatmap. In particular, the high-quality therapeutical content retrieved for the queried topics *Compliance with Rules (CR)*, *Diabetes Management (DM)*, and *Managing Life (MF)* is systematically ranked higher compared

to the other models.

Such observation is probably related to the higher representation of high-quality utterances for the just mentioned topics. However, what seems to affect the ranking utility the most is the high representation of low-quality utterances. Indeed, the topics *SC*, *RAC*, *AMP* are composed of more than 1% of low-quality documents and the related columns are closer to the left side of the heatmap for each relevance level. A balanced representation of low- and high-quality utterances, as for the *RAC-SC* mental health issue, results in models with ranking utility divergent from each other.

Though these results are affected by the representation of *MI quality* classes, it does not seem evident a systematic impact on specific diseases. At least one model is able to provide optimal utility for a query and different models exhibit high and low-ranking utility for the same topics, except for the ones with an over-representation of high-quality utterances.

A.6 Conclusions and Future Works

In this paper, we investigated whether IR could be instrumental in supporting patients with their own mental health issues. Our methodology was based on a ranking task to provide high-quality therapeutic content in higher positions than low-quality ones. For such purpose, we used nine ranking models of a wide diversity of network complexities, and our results are indicative that conversational therapeutic data is suitable for ranking tasks, reporting high average utility.

Future works will focus on augmentation techniques to extend AnnoMI to generate a higher number of therapy sessions per topic to work with. Unfairness mitigation procedures will also be employed to reduce the ranking utility disparity reported across psychological topics. We also aim to incorporate world knowledge in the form of triples to address the domain adaptation challenges in mental health [87].

Appendix B

Generalized Explainer of Global Issues

Recently, we have been investigating whether our unfairness explainer in GNN-based recommender systems (GNNUERS) could be extended not only to explain unfairness, but also other issues. Indeed, unfairness is an issue at the global level because it regards the whole system, not a single consumer or provider. It follows that switching the objective and adjusting the manipulation of the graph topology could make the method generalizable towards explaining other global issues such as provider unfairness [59], instability [117], user coverage [103].

B.1 Proposed Method

Hence, we introduce a novel approach, named GENIUS-RS (**G**eneralized **E**xplai**N**er of Global **I**ss**U**e**S** in GNN-based **R**ecommender **S**ystems), which finds and adopts a set of user-item interactions as an explanation of a global issue affecting a GNN-based recommender system. In other words, the manipulated graph fed in input to the GNN, treated as a black-box, aims to make the latter generate a set of altered recommendations with an alleviated level of the global issue. The edges differing between the manipulated graph and the original one represent, then, an explanation of the global issue. As GNNUERS, this procedure is driven by counterfactual reasoning, given that the manipulated graph represents a distorted version of the original graph in the counterfactual world. GENIUS-RS is still an ongoing work, given that we aim to extend the experiments suite to account for other global issues, e.g., instability and user coverage, to reflect the purpose of this approach. We also plan to extend the explanation tools to not only consider the removal of the edges, denoted as GENIUS-RS^{-}, but also the addition, denoted as GENIUS-RS^{+}, and the edge rewiring, still a work in progress.

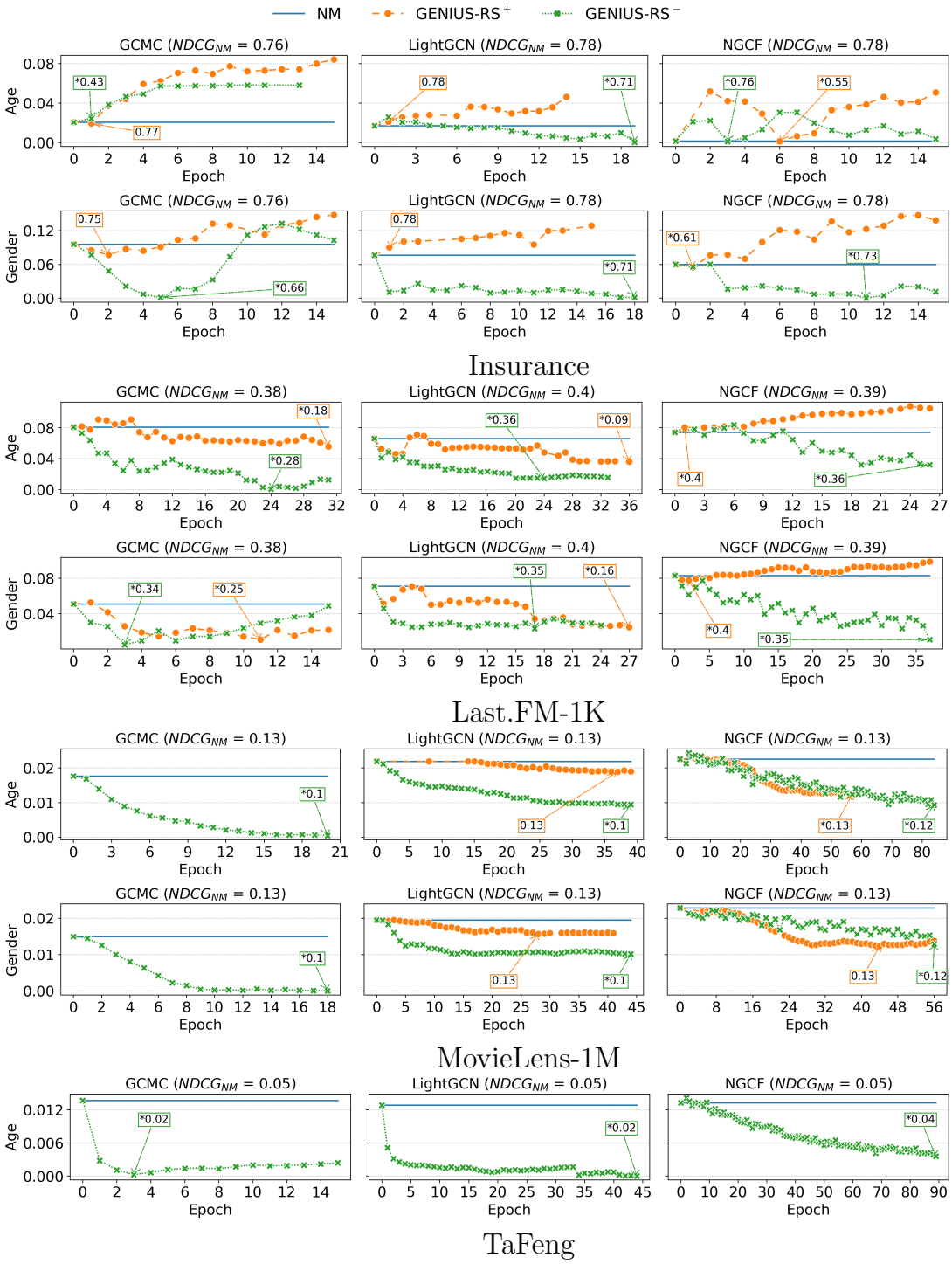


Figure B.1: ΔNDCG over a minimum of 15 epochs and a maximum equal to the epoch where one of `GENIUS-RS`$^{+}$ or `GENIUS-RS`$^{-}$ reported the most effective explanation (lower ΔNDCG). The horizontal line represents the ΔNDCG measured on the original recommendations (without manipulation). The annotated values are the NDCG measures with the manipulated graph (NDCG of `NM` is reported in the title) at the epoch (pointed by the arrow) with the lowest ΔNDCG for each manipulation strategy, and (*) denotes the statistical difference from the NDCG distribution of `NM` under a Wilcoxon signed-rank test with a 95% confidence interval.

B.2 Experimental Evaluation

We present here a partial experimental evaluation based on the same experimental setup used for GNNUERS in Section 4.4.2, presenting only the manipulation strategies of edge removal (GENIUS-RS$^-$) and addition (GENIUS-RS$^+$) adopted for explaining unfairness across consumer groups. The two manipulation strategies GENIUS-RS$^-$ and GENIUS-RS$^+$ operate on two significantly different set of predefined edges. The high sparsity of the selected datasets bounds GENIUS-RS$^-$ to manipulate a maximum of 1,000,209 user-item interactions (MovieLens-1M), while GENIUS-RS$^+$ can manipulate more than 600,000,000 graph edges (TaFeng). It follows that our preliminary results do not include the experiments on TaFeng for all the settings, and on MovieLens-1M for GCMC, due to out-of-memory problems raised with GENIUS-RS$^+$.

Figure B.1 depicts the monitoring of the ΔNDCG over the epochs along which GENIUS-RS is executed in its edge deletion (GENIUS-RS$^-$) and addition (GENIUS-RS$^+$) manipulation strategies. Compared with the consumer unfairness measured with the non-manipulated graph (NM), most of the experiments report a significantly lower ΔNDCG along the epochs. Except for GCMC and NGCF on Insurance (the smallest dataset) across age groups, GENIUS-RS$^-$ can consistently explain consumer unfairness across demographic groups, reaching quasi-optimal explanation levels (ΔNDCG $\simeq$ 0) in several experiments. Conversely, GENIUS-RS$^+$ seems dependent on the dataset size, given that it cannot explain the consumer unfairness on Insurance. It also reports different trends among the models, possibly due to the GNNs being distinctly influenced by the unseen interactions added by GENIUS-RS$^+$.

In summary, except for extreme cases and despite the limited set of experiments, GENIUS-RS, in particular GENIUS-RS$^-$, systematically provides explanations of the given global issue in a specific way, thanks to the moderate losses in NDCG. We plan to carry out further analyses on additional global issues to confirm such preliminary findings and the GENIUS-RS's generalizability level, potentially presenting our results to forthcoming conferences or journals.

Bibliography

[1] Acm artifact review and badging. `https://www.acm.org/publications/policies/artifact-review-and-badging-current`, 2021. [Online; accessed 25-September-2021].

[2] The artificial intelligence act. https://artificialintelligenceact.eu/, 2023.

[3] Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. Perceptions and opinions of patients about mental health chatbots: Scoping review. *J. Med. Internet Res.*, 23(1):e17828, January 2021.

[4] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.*, 30(1):127–158, 2020.

[5] Carlo Abrate and Francesco Bonchi. Counterfactual graphs for explainable classification of brain networks. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2495–2504. ACM, 2021.

[6] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021*, volume 161, pages 2114–2124. AUAI Press, 2021.

[7] Mark Przybocki Alvin and Alvin Martin. Nist speaker recognition evaluation chronicles. In *Proc. of the IEEE Odyssey - The Speaker and Language Recognition Workshop*, pages 12–22, 2004.

[8] Marcelo G Armentano, Ariel Monteserin, Franco Berdun, Emilio Bongiorno, and Luis María Coussirat. User recommendation in low degree networks with a learning-based approach. In *Mexican International Conference on Artificial Intelligence*, pages 286–298. Springer, 2018.

[9] Ashwathy Ashokan and Christian Haas. Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.*, 58(5):102646, 2021.

[10] Giacomo Balloccu, Ludovico Boratto, Gianni Fenu, and Mirko Marras. Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 646–656. ACM, 2022.

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

[12] Aakash Bhandari, Vivek Kumar, Pham Thi Thien Huong, and Dang NH Thanh. Sentiment analysis of covid-19 tweets: Leveraging stacked word embedding representation for identifying distinct classes within a sentiment. In *Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021*, pages 341–352. Springer, 2022.

[13] Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, and Fernando Ortega. Deepfair: Deep learning for improving fairness in recommender systems. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(6):86–94, 06/2021 2021.

[14] Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program], 2022.

[15] Jean-François Bonastre, Frédéric Bimbot, Louis-Jean Boë, Joseph P Campbell, Douglas A Reynolds, and Ivan Magrin-Chagnolleau. Person authentication by voice: A need for caution. In *Proc. of the European Conference on Speech Communication and Technology (ECSCT)*, 2003.

[16] Ludovico Boratto, Francesco Fabbri, Gianni Fenu, Mirko Marras, and Giacomo Medda. Counterfactual graph augmentation for consumer unfairness mitigation in recommender systems. In *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management, Birmingham, UK, October 21-25, 2023*, 2023.

[17] Ludovico Boratto, Gianni Fenu, and Mirko Marras. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model. User Adapt. Interact.*, 31(3):421–455, 2021.

[18] Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. Consumer fairness in recommender systems: Contextualizing definitions and mitigations. In *Advances in Information Retrieval - 44th European Conference on*

IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I, pages 552–566, 2022.

[19] Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. Practical perspectives of consumer fairness in recommendation. *Inf. Process. Manag.*, 60(2):103208, 2023.

[20] Léo Brunot, Nicolas Canovas, Alexandre Chanson, Nicolas Labroche, and Willeme Verdeaux. Preference-based and local post-hoc explanations for recommender systems. *Inf. Syst.*, 108:102021, 2022.

[21] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR, 2018.

[22] Tom Bäckström, Okko Räsänen, Abraham Zewoudie, Pablo Pérez Zarazaga, and Liisa Koivusalo. Introduction to speech processing, 2019.

[23] Federico Cabitza, Davide Ciucci, Gabriella Pasi, and Marco Viviani. Responsible AI in healthcare. *CoRR*, abs/2203.03616, 2022.

[24] Joseph P Campbell, Wade Shen, William M Campbell, Reva Schwartz, Jean-Francois Bonastre, and Driss Matrouf. Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2):95–103, 2009.

[25] Òscar Celma. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

[26] Chong Chen, Min Zhang, Yongfeng Zhang, Yiqun Liu, and Shaoping Ma. Efficient neural matrix factorization without sampling for recommendation. *ACM Trans. Inf. Syst.*, 38(2):14:1–14:28, 2020.

[27] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *CoRR*, abs/2010.03240, 2020.

[28] Richard J. Chen, Tiffany Y. Chen, Jana Lipková, Judy J. Wang, Drew F. K. Williamson, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in AI for medicine and healthcare. *CoRR*, abs/2110.00603, 2021.

[29] Weijian Chen, Fuli Feng, Qifan Wang, Xiangnan He, Chonggang Song, Guohui Ling, and Yongdong Zhang. Catgcn: Graph convolutional networks with categorical node features. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[30] Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. Semi-supervised user profiling with heterogeneous graph attention networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2116–2122, 2019.

[31] Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. Measuring "why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation. *CoRR*, abs/2202.06466, 2022.

[32] Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, Zhenhua Huang, Hongshik Ahn, and Gabriele Tolomei. GREASE: generate factual and counterfactual explanations for gnn-based recommendations. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2022, Proceedings*. Springer, 2022.

[33] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.

[34] Hejie Cui, Jiaying Lu, Yao Ge, and Carl Yang. How can graph neural networks help document retrieval: A case study on cord19 with concept map generation. In *European Conference on Information Retrieval*, pages 75–83. Springer, 2022.

[35] Geoffrey Currie and K Elizabeth Hawk. Ethical and legal challenges of artificial intelligence in nuclear medicine. *Semin Nucl Med*, 51(2):120–125, September 2020.

[36] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 101–109. ACM, 2019.

[37] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems, 2020.

[38] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Speech Audio Process.*, 19(4):788–798, 2011.

[39] Yashar Deldjoo, Alejandro Bellogín, and Tommaso Di Noia. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Inf. Process. Manag.*, 58(5):102662, 2021.

[40] James A Diao, Gloria J Wu, Herman A Taylor, John K Tucker, Neil R Powe, Isaac S Kohane, and Arjun K Manrai. Clinical implications of removing race from estimates of kidney function. *JAMA*, 325(2):184–186, January 2021.

[41] Karlijn Dinnissen and Christine Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in Big Data*, page 63.

[42] George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Technical report, National Inst of Standards and Technology Gaithersburg Md, 1998.

[43] Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li. Interpreting unfairness in graph neural networks via training node attribution. *CoRR*, abs/2211.14383, 2022.

[44] Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li. Interpreting unfairness in graph neural networks via training node attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[45] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[46] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness and discrimination in information access systems. *CoRR*, abs/2105.05779, 2021.

[47] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency, FAT 2018*, volume 81, pages 172–186. PMLR, 2018.

[48] Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. Exposure inequality in people recommender systems: The long-term effects. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022*, pages 194–204. AAAI Press, 2022.

[49] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor. A fairness-aware hybrid recommender system. *CoRR*, abs/1809.09030, 2018.

[50] David R Feinberg. Parselmouth praat scripts in python, Jan 2022.

[51] Gianni Fenu, Hicham Lafhouli, and Mirko Marras. Exploring algorithmic fairness in deep speaker verification. In *Proc. of the International Conference eon Computational Science and Its Applications (ICCSA)*, pages 77–93, 2020.

[52] Gianni Fenu and Mirko Marras. Demographic fairness in multimodal biometrics: A comparative analysis on audio-visual speaker recognition systems. *Procedia Computer Science*, 198:249–254, 2022.

[53] Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. Fair voice biometrics: Impact of demographic imbalance on group fairness in speaker recognition. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 1892–1896, 2021.

[54] Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. Causal reasoning for algorithmic fairness in voice controlled cyber-physical systems. *Pattern Recognit. Lett.*, 168:131–137, 2023.

[55] Gianni Fenu, Giacomo Medda, Mirko Marras, and Giacomo Meloni. Improving fairness in speaker recognition. In *ESSE 2020: 2020 European Symposium on Software Engineering, Rome, Italy, November 6-8, 2020*, pages 129–136, 2020.

[56] W T Fitch and J Giedd. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *J Acoust Soc Am*, 106(3 Pt 1):1511–1522, September 1999.

[57] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. capAI - a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. *SSRN Electronic Journal*, 2022.

[58] Gabriel Frisch, Jean-Benoist Leger, and Yves Grandvalet. Co-clustering for fair recommendation. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2021.

[59] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. Explainable fairness in recommendation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 681–691. ACM, 2022.

[60] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. PRINCE: provider-side interpretability with counterfactual explanations in recommender systems. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining*, pages 196–204. ACM, 2020.

[61] Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. ELIXIR: learning from user feedback on explanations to improve recommender models. In *WWW '21: The Web Conference 2021*. ACM / IW3C2, 2021.

[62] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1808–1812. ACM, 2021.

[63] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Guilherme Ramos. Enabling cross-continent provider fairness in educational recommender systems. *Future Gener. Comput. Syst.*, 127:435–447, 2022.

[64] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[65] Bryce Goodman and Seth R. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag.*, 38(3), 2017.

[66] Jiafeng Guo, Yixing Fan, Xiang Ji, and Xueqi Cheng. Matchzoo: A learning, practicing, and developing system for neural text matching. In Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer, editors, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1297–1300. ACM, 2019.

[67] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[68] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. Learning-to-rank with BERT in tf-ranking. *CoRR*, abs/2004.08476, 2020.

[69] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, 2016.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR 2016*, pages 770–778. IEEE Computer Society, 2016.

[71] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 639–648. ACM, 2020.

[72] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[73] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2042–2050, 2014.

[74] Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan. Achieving user-side fairness in contextual bandits. *CoRR*, abs/2010.12102, 2020.

[75] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 946–956. Springer, 2019.

[76] Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in automated speaker recognition. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 230–247. ACM, 2022.

[77] Malinka Ivanova, Sushil Bhattacharjee, Sébastien Marcel, Anna Rozeva, and Mariana Durcheva. Enhancing trust in eassessment - the tesla system solution. In *Technology Enhanced Assessment Conf.*, December 2018.

[78] Muhammad Mohsin Kabir, Muhammad F. Mridha, Jungpil Shin, Israt Jahan, and Abu Quwsar Ohi. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*, 9:79236–79263, 2021.

[79] Juliette Kahn, Nicolas Audibert, Solange Rossato, and Jean-François Bonastre. Intra-speaker variability effects on speaker verification performance. In *Odyssey*, page 21, 2010.

[80] Toshihiro Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In Lise Getoor, Ted E. Senator, Pedro M. Domingos, and Christos Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 583–588. ACM, 2003.

[81] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In *Conference on Fairness, Accountability and*

Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, volume 81 of *Proceedings of Machine Learning Research*, pages 187–201. PMLR, 2018.

[82] Bo Kang, Jefrey Lijffijt, and Tijl De Bie. Explanations for network embedding-based link predictions. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021*, volume 1524, pages 473–488. Springer, 2021.

[83] Cecilia Kang. In u.s., regulating a.i. is in its "early days", Jul 2023.

[84] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, 2017.

[85] Taeyong Kong, Taeri Kim, Jinsung Jeon, Jeongwhan Choi, Yeon-Chang Lee, Noseong Park, and Sang-Wook Kim. Linear, or non-linear, that is the question! In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 517–525. ACM, 2022.

[86] Vivek Kumar, Giacomo Medda, Diego Reforgiato Recupero, Daniele Riboni, Rim Helaoui, and Gianni Fenu. How do you feel? information retrieval in psychotherapy and fair ranking assessment. In *Advances in Bias and Fairness in Information Retrieval - 4th International Workshop, BIAS 2023, Dublin, Ireland, April 2, 2023, Revised Selected Papers*, pages 119–133, 2023.

[87] Vivek Kumar, Diego Reforgiato Recupero, Rim Helaoui, and Daniele Riboni. K-lm: Knowledge augmenting in language models within the scholarly domain. *IEEE Access*, 10:91802–91815, 2022.

[88] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[89] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo SperlÃ¬. Pastle: Pivot-aided space transformation for local explanations. *Pattern Recognition Letters*, 149:67–74, 2021.

[90] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manag. Sci.*, 65(7):2966–2981, 2019.

[91] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. Analyzing item popularity bias

of music recommender systems: Are different genders equally affected? In *Fifteenth ACM Conference on Recommender Systems*, pages 601–606, 2021.

[92] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In *WWW '21: The Web Conference 2021*, pages 624–632. ACM / IW3C2, 2021.

[93] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. Towards personalized fairness based on causal notion. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1054–1063. ACM, 2021.

[94] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2320–2329. ACM, 2022.

[95] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dinggang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *CoRR*, abs/2304.01852, 2023.

[96] III Lopez, Leo, III Hart, Louis H., and Mitchell H. Katz. Racial and Ethnic Health Disparities Related to COVID-19. *JAMA*, 325(8):719–720, 02 2021.

[97] Ana Lucic, Maartje A. ter Hoeve, Gabriele Tolomei, Maarten de Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022*, volume 151, pages 4499–4511. PMLR, 2022.

[98] Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. Improving biomedical information retrieval with neural retrievers. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11038–11046. AAAI Press, 2022.

[99] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining*, pages 695–703. ACM, 2022.

[100] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019)*, volume 2440 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[101] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education*, pages 1–49, 2021.

[102] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. Equality of learning opportunity via individual fairness in personalized recommendations. *Int. J. Artif. Intell. Educ.*, 32(3):636–684, 2022.

[103] Paolo Massa and Paolo Avesani. Trust metrics in recommender systems. In Jennifer Golbeck, editor, *Computing with Social Trust*, Human-Computer Interaction Series, pages 259–285. Springer, 2009.

[104] Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000, Beijing, China, October 16-20, 2000*, pages 302–305. ISCA, 2000.

[105] Driss Matrouf, Jean-François Bonastre, and Corinne Fredouille. Effect of speech transformation on impostor acceptance. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006*, pages 933–936. IEEE, 2006.

[106] Noemi Mauro, Liliana Ardissono, Stefano Cocomazzi, and Federica Cena. Using consumer feedback from location-based services in poi recommender systems for people with autism. *Expert Systems with Applications*, 199:116972, 2022.

[107] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

[108] Yen Meng, Yi-Hui Chou, Andy T Liu, and Hung-yi Lee. Don't speak too fast: The impact of data bias on self-supervised speech models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3258–3262. IEEE, 2022.

[109] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.*, 3(8):659–666, 2021.

[110] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1291–1299. ACM, 2017.

[111] Ajili Moez, Bonastre Jean-François, Ben Kheder Waad, Rossato Solange, and Kahn Juliette. Phonetic content impact on forensic voice comparison. In *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, pages 210–217. IEEE, 2016.

[112] Janet Morahan-Martin. How internet users find, evaluate, and use online health information: A cross-cultural review. *Cyberpsychology Behav. Soc. Netw.*, 7(5):497–510, 2004.

[113] Janet Morahan-Martin and Colleen D. Anderson. Information and misinformation online: Recommendations for facilitating accurate mental health information retrieval and evaluation. *Cyberpsychology Behav. Soc. Netw.*, 3(5):731–746, 2000.

[114] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.*, 60, 2020.

[115] Mahesh Kumar Nandwana, Luciana Ferrer, Mitchell McLaren, Diego Castan, and Aaron Lawson. Analysis of critical metadata factors for the calibration of speaker recognition systems. In *Proc. of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 4325–4329, 2019.

[116] Julia Neidhardt and Mete Sertkan. Towards an approach for analyzing dynamic aspects of bias and beyond-accuracy measures. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 35–42. Springer, 2022.

[117] Sejoon Oh, Berk Ustun, Julian J. McAuley, and Srijan Kumar. Rank list sensitivity of recommender systems to interaction perturbations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1584–1594. ACM, 2022.

[118] Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. A study of bias mitigation strategies for speaker recognition. *Computer Speech & Language*, 79:101481, 2023.

[119] Mark A. Przybocki, Alvin F. Martin, and Audrey N. Le. Nist speaker recognition evaluation chronicles - part 2. In *Proc. of the IEEE Odyssey - Speaker and Language Recognition Workshop*, pages 1–6, 2006.

[120] Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Inf. Retr.*, 13(4):375–397, 2010.

[121] Hossein A. Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. Experiments on generalizability of user-oriented fairness in recommender systems. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2755–2764. ACM, 2022.

[122] Amifa Raj and Michael D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 726–736. ACM, 2022.

[123] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[124] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 231–239. ACM, 2019.

[125] Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with sincnet. In *Proc. SLT 2018*, pages 1021–1028, 2018.

[126] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digit. Signal Process.*, 10(1-3), 2000.

[127] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 97–101. The Association for Computational Linguistics, 2016.

[128] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer US, 2022.

[129] Ignacio Serna, Aythami Morales, Julian Fierrez, Manuel Cebrian, Nick Obradovich, and Iyad Rahwan. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*, 2020.

[130] Ignacio Serna, Alejandro Peña, Aythami Morales, and Julian Fierrez. Insidebias: Measuring bias in deep networks and application to face gender biometrics. In *Proc. of IAPR International Conference on Pattern Recognition (ICPR)*, 01 2021.

[131] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. Improving fairness in speaker verification via group-adapted fusion network. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7077–7081. IEEE, 2022.

[132] Lonnie R. Snowden. Bias in mental health assessment and intervention: Theory and evidence. *American Journal of Public Health*, 93(2):239–243, 2003. PMID: 12554576.

[133] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[134] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2239–2248, New York, NY, USA, 2018. Association for Computing Machinery.

[135] Suraj Srinivas, Akshayvarun Subramanya, and R. Venkatesh Babu. Training sparse neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017*, pages 455–462. IEEE Computer Society, 2017.

[136] Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3251–3257. ACM, 2019.

[137] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu, editors, *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1441–1450. ACM, 2019.

[138] Kalaivani Sundararajan and Damon L. Woodard. Deep learning for biometrics: A survey. *ACM Comput. Surv.*, 51(3):65:1–65:34, 2018.

[139] Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. Sentence embeddings in NLI with iterative refinement encoders. *Nat. Lang. Eng.*, 25(4):467–482, 2019.

[140] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. Counterfactual explainable recommendation. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 1784–1793. ACM, 2021.

[141] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. Bias disparity in recommendation systems. In Robin Burke, Himan Abdollahpouri, Edward C. Malthouse, K. P. Thai, and Yongfeng Zhang, editors, *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019*, volume 2440 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[142] Rianne van den Berg, Thomas N. Kipf, and Max Welling. Graph convolutional matrix completion. *CoRR*, abs/1706.02263, 2017.

[143] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2017.

[144] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno. Generalized end-to-end loss for speaker verification. In *Proc. ICASSP*, pages 4879–4883, 2018.

[145] Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. Towards representation alignment and uniformity in

collaborative filtering. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 1816–1825. ACM, 2022.

[146] Nan Wang, Lu Lin, Jundong Li, and Hongning Wang. Unbiased graph embedding with biased graph observations. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 1423–1433. ACM, 2022.

[147] Shoujin Wang, Xiuzhen Zhang, Yan Wang, Huan Liu, and Francesco Ricci. Trustworthy recommender systems. *CoRR*, abs/2208.06265, 2022.

[148] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 832–841. ACM, 2023.

[149] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 2019.

[150] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, jul 2022. Just Accepted.

[151] Zhiming Wang, Kaisheng Yao, Xiaolong Li, and Shuo Fang. Multi-resolution multi-head attention in deep speaker embedding. In *Proc. ICASSP 2020*, pages 6464–6468, 2020.

[152] K Wells, R Klap, A Koike, and C Sherbourne. Ethnic disparities in unmet need for alcoholism, drug abuse, and mental health care. *Am. J. Psychiatry*, 158(12):2027–2032, December 2001.

[153] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. Neural gender prediction from news browsing data. In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings*, volume 11856 of *Lecture Notes in Computer Science*, pages 664–676. Springer, 2019.

[154] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Fairness-aware news recommendation with decomposed adversarial learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 4462–4469. AAAI Press, 2021.

[155] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. *ACM Trans. Inf. Syst.*, aug 2022. Just Accepted.

[156] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 726–735. ACM, 2021.

[157] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. Learning fair representations for recommendation: A graph-based perspective. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2198–2208. ACM / IW3C2, 2021.

[158] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE, 2022.

[159] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Utterance-level aggregation for speaker recognition in the wild. In *Proc. ICASSP 2019*, pages 5791–5795, 2019.

[160] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 55–64. ACM, 2017.

[161] Sarthak Yadav and Atul Rai. Frequency and temporal convolutional attention for text-independent speaker recognition. In *Proc. ICASSP 2020*, pages 6794–6798.

[162] Qilong Yan, Yufeng Zhang, Qiang Liu, Shu Wu, and Liang Wang. Relation-aware heterogeneous graph for user profiling. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, New York, NY, USA, October 2021. Association for Computing Machinery.

[163] Zhou Yang, Qingfeng Lan, Jiafeng Guo, Yixing Fan, Xiaofei Zhu, Yanyan Lan, Yue Wang, and Xueqi Cheng. A deep top-k relevance matching model for ad-hoc retrieval. In Shichao Zhang, Tie-Yan Liu, Xianxian Li, Jiafeng Guo, and Chenliang Li, editors, *Information Retrieval - 24th China Conference, CCIR*

2018, Guilin, China, September 27-29, 2018, Proceedings, volume 11168 of *Lecture Notes in Computer Science*. Springer, 2018.

[164] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377, 2019.

[165] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983, 2018.

[166] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. Xsimgcl: Towards extremely simple graph contrastive learning for recommendation. *IEEE Trans. Knowl. Data Eng.*, 36(2):913–926, 2024.

[167] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. XGNN: towards model-level explanations of graph neural networks. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 430–438. ACM, 2020.

[168] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5):5782–5799, 2023.

[169] Shichang Zhang, Jiani Zhang, Xiang Song, Soji Adeshina, Da Zheng, Christos Faloutsos, and Yizhou Sun. Page-link: Path-based graph neural network explanation for heterogeneous link prediction. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3784–3793. ACM, 2023.

[170] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.*, 14(1):1–101, mar 2020.

[171] Yuanyuan Zhang, Yixuan Zhang, Bence Mark Halpern, Tanvina Patel, and Odette Scharenborg. Mitigating bias against non-native accents. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3168–3172. ISCA, 2022.

[172] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2022.

[173] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao

Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management*, pages 4653–4664. ACM, 2021.

[174] Jinfeng Zhong and Elsa Negre. Shap-enhanced counterfactual explanations for recommendations. In Jiman Hong, Miroslav Bures, Juw Won Park, and Tomás Cerný, editors, *SAC '22: The 37th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, April 25 - 29, 2022*, pages 1365–1372. ACM, 2022.

[175] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.