FACULTY OF
COMPUTER SCIENCE

Human-Centred
Artificial Intelligence
Otto von Guericke University - Magdeburg

Otto von Guericke University Magdeburg

Ph.D. Program in Computer Science

Thesis for the Degree of Doctor of Philosophy

Fairness Analysis of Graph Neural Networks for Behavioral User Modeling

by

Erasmo Purificato

Advisor: Prof. Dr.-Ing. Ernesto William De Luca

Faculty of Computer Science

Institute of Technical and Business Information Systems

Human-Centred Artificial Intelligence Research Group

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

Fairness Analysis of Graph Neural Networks for Behavioral User Modeling

DISSERTATION

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

angenommen durch die Fakultät für Informatik
der Otto-von-Guericke-Universität Magdeburg

von M.Sc., Erasmo, Purificato

geb. am 11.05.1991 in Formia (Italien)

Gutachterinnen/Gutachter

Prof. Dr.-Ing. Ernesto William De Luca
Prof. Dr. Federica Cena
Prof. Dr. Marco De Gemmis

Magdeburg, den 16.12.2024

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Candidate's declaration

I hereby declare that this thesis, submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Computer Science, is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references. Automated tools (e.g., Large Language Models) have only been used as grammar and spelling checkers.

In particular, I did not knowingly invent any results or conceal contradictory results, intentionally misuse statistical procedures to interpret data in an unjustified manner, plagiarize other people's results or publications, or distort the results of other research.

I am aware that violations of copyright law may result in injunctive relief, claims for damages by the author, and criminal prosecution by law-enforcing authorities. The work has not yet been submitted as a dissertation in the same or a similar form either in Germany or abroad and has not yet been published as a whole.

Parts of this dissertation have been published in international journals and/or conference articles (see the list of the author's publications at the end of the thesis).

Magdeburg, March 30, 2025

Erasmo Purificato

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Abstract

Artificial Intelligence (AI) systems have become integral to modern life, driving the functionality of Information Retrieval and Recommender Systems to provide tailored content efficiently. As these technologies advance and embed themselves into various sectors such as healthcare, finance, and social media, addressing their ethical and societal impacts is increasingly critical. Ensuring that AI systems are transparent, equitable, and sustainable is essential. This highlights the need for research in fields like Human-Centered AI (HCAI) and Responsible AI that promote respect for societal values. The ethical implications of AI have drawn significant attention, leading to the establishment of several European regulations that emphasize principles such as transparency, accountability, and fairness. These regulations aim to protect individuals' rights by ensuring clear explanations for AI decisions, obtaining explicit consent before automated decisions, and preventing discrimination.

In response to these ethical challenges, this dissertation begins by developing a Responsible AI framework, focusing on transparency and fairness to build trust and reliability among domain experts. This foundational work sets the stage for a detailed exploration of algorithmic fairness in user modeling applications, particularly those involving Graph Neural Networks (GNNs). While GNNs have shown significant potential in converting user interaction data derived from graph structures into actionable insights, traditional evaluations often neglect fairness considerations, focusing primarily on accuracy metrics.

In our research, we assessed fairness in state-of-the-art GNNs for behavioral user modeling, detecting potential discrimination based on different user modeling paradigms. Recognizing the limitations of binary fairness metrics, which oversimplify real-world scenarios by forcing non-binary attributes into binary categories, we extended four existing fairness metrics to adapt to multiclass scenarios. This expansion enables a more detailed and accurate evaluation of fairness, enhancing comprehension of AI systems' influence on various user demographics.

To tackle the potential discrimination detected in GNN-based user modeling, we introduced FAME (short for Fairness-Aware MEssages), an in-processing bias mitigation strategy that modifies the message-passing algorithm during GNN training. Additionally, we developed FairUP, a framework that standardizes input processing and integrates comprehensive fairness analysis components, supporting both pre-processing and post-processing fairness techniques. We have also developed GNNFairViz, an interactive visual analytics tool designed to facilitate the practical implementation of these methods.

This tool offers interactive visualizations to help users examine and identify biases in GNN models.

Overall, this dissertation addresses critical challenges in fairness analysis for behavioral user modeling, focusing on ethical AI development. Initially, we designed decision-making systems aligned with HCAI and Responsible AI principles, complying with European regulations. The core contribution is the innovative evaluation of fairness in user modeling using GNNs, introducing methods to assess and mitigate biases for more equitable AI outcomes. Additionally, we extended binary fairness metrics to multiclass and multigroup scenarios for a more accurate evaluation. The development of unified frameworks for standardized fairness evaluation and visualization enhances consistency, transparency, and understanding of the impact of the models. These contributions emphasize the importance of integrating fairness into systems' design, laying a foundation for future responsible AI research and applications.

Keywords: User Modeling, Algorithmic Fairness, Human-Centered AI, Responsible AI, Graph Neural Networks.

Zusammenfassung

Künstliche Intelligenz (KI)-Systeme sind zu einem integralen Bestandteil des modernen Lebens geworden und treiben die Funktionalität von Information Retrieval und Empfehlungssystemen an, um maßgeschneiderte Inhalte effizient bereitzustellen. Da sich diese Technologien weiterentwickeln und in verschiedene Sektoren wie Gesundheitswesen, Finanzen und soziale Medien einbetten, wird es immer wichtiger, ihre ethischen und gesellschaftlichen Auswirkungen zu berücksichtigen. Es ist deshalb essentiell, dass KI-Systeme transparent, gerecht und nachhaltig sind. Dies unterstreicht die Notwendigkeit von Forschung in Bereichen wie menschenzentrierte KI und verantwortungsbewusste KI, die den Respekt vor gesellschaftlichen Werten fördern.

Die ethischen Implikationen von KI haben erhebliche Aufmerksamkeit erregt und zur Etablierung mehrerer europäischer Vorschriften geführt, die Prinzipien wie Transparenz, Verantwortlichkeit und Fairness betonen. Diese Vorschriften zielen darauf ab, die Rechte der Individuen zu schützen, indem sie klare Erklärungen für KI-Entscheidungen gewährleisten, eine explizite Zustimmung vor automatisierten Entscheidungen einholen und Diskriminierung verhindern.

Als Antwort auf diese ethischen Herausforderungen entwickelt diese Dissertation zunächst ein Framework für verantwortungsbewusste KI, das sich auf Transparenz und Fairness konzentriert, um Vertrauen und Zuverlässigkeit unter Fachexperten aufzubauen. Diese Grundlagenarbeit bildet die Basis für eine detaillierte Exploration der algorithmischen Fairness in Anwendungen zur Benutzermodellierung, insbesondere solchen, die Graph Neural Networks (GNNs) beinhalten. Während GNNs ein erhebliches Potenzial bei der Umwandlung von Benutzerinteraktionsdaten aus Graphstrukturen in umsetzbare Erkenntnisse gezeigt haben, vernachlässigen traditionelle Bewertungen oft Fairnessaspekte und konzentrieren sich hauptsächlich auf Genauigkeitsmetriken.

In unserer Forschung haben wir die Fairness in hochmodernen GNNs für die Verhaltensbenutzermodellierung bewertet und mögliche Diskriminierungen basierend auf verschiedenen Benutzermodellierungsparadigmen entdeckt. Angesichts der Einschränkungen binärer Fairnessmetriken, die reale Szenarien durch das Erzwingen nicht-binärer Attribute in binäre Kategorien vereinfachen, haben wir vier bestehende Fairnessmetriken erweitert, um sie an Mehrklassen-Szenarien anzupassen. Diese Erweiterung ermöglicht eine detailliertere und genauere Bewertung der Fairness und verbessert das Verständnis des Einflusses von KI-Systemen auf verschiedene Benutzerdemografien.

Um die in GNN-basierten Benutzermodellierungen erkannte potenzielle Diskriminierung zu bekämpfen, haben wir FAME (Fairness-Aware Messages) eingeführt, eine

In-Processing-Bias-Mitigation-Strategie, die den Nachrichtenübermittlungsalgorithmus während des GNN-Trainings modifiziert. Darüber hinaus haben wir FairUP entwickelt, ein Rahmenwerk, das die Eingangsverarbeitung standardisiert und umfassende Fairnessanalysekomponenten integriert und sowohl Pre-Processing- als auch Post-Processing-Fairnesstechniken unterstützt. Wir haben auch GNNFairViz entwickelt, ein interaktives visuelles Analysetool, das die praktische Implementierung dieser Methoden erleichtert. Dieses Tool bietet interaktive Visualisierungen, um Benutzern zu helfen, Verzerrungen in GNN-Modellen zu untersuchen und zu identifizieren.

Insgesamt befasst sich diese Dissertation mit kritischen Herausforderungen in der Fairness-Analyse für Behavioral User Modeling und konzentriert sich auf die ethische Entwicklung von KI. Zunächst wurden Entscheidungssysteme entworfen, die mit den Prinzipien von menschenzentrierte KI und verantwortungsvoller KI übereinstimmen und den europäischen Vorschriften entsprechen. Der Kernbeitrag ist die innovative Bewertung der Fairness in der Benutzermodellierung mittels GNNs, wobei Methoden zur Bewertung und Minderung von Verzerrungen eingeführt werden, um gerechtere KI-Ergebnisse zu erzielen. Darüber hinaus werden binäre Fairness-Metriken auf Mehrklassen- und Mehrgruppenszenarien erweitert, um eine genauere Bewertung zu ermöglichen. Die Entwicklung einheitlicher Frameworks für standardisierte Fairness-Bewertung und -Visualisierung verbessert die Konsistenz, Transparenz und das Verständnis der Auswirkungen der Modelle. Diese Beiträge betonen die Bedeutung der Integration von Fairness in das Systemdesign und legen eine Grundlage für zukünftige Forschung und Anwendungen im Bereich der verantwortungsvollen KI.

Schlüsselwörter: Benutzermodellierung, Algorithmischen Fairness, Menschenzentrierte KI, Verantwortungsbewusste KI, Graph Neural Networks.

Sommario

I sistemi di Intelligenza Artificiale (IA) sono diventati parte integrante della vita moderna, guidando il funzionamento dei sistemi di recupero delle informazioni e dei sistemi di raccomandazione al fine di fornire contenuti personalizzati in modo efficiente. Con l'avanzamento e l'integrazione di queste tecnologie in vari settori quali la sanità, la finanza e i social media, diventa essenziale trattare i loro impatti etici e sociali per garantire che i sistemi di IA siano trasparenti, equi e sostenibili. Ciò comporta la necessità di approfondire la ricerca in campi come l'IA Responsabile che promuovano il rispetto di valori etici. Le implicazioni etiche dell'IA hanno attirato notevole attenzione, portando all'istituzione di diverse normative europee che mirano a proteggere i diritti degli individui assicurando spiegazioni chiare per le decisioni dell'IA, ottenere il consenso esplicito degli utenti soggetti a decisioni di sistemi automatici e prevenire che gli utenti vengano discriminati da tali sistemi.

In risposta a questi requisiti etici, il progetto di dottorato inizia con lo sviluppo di un sistema di IA Responsabile che, focalizzandosi su trasparenza ed equità, punta a trasmettere fiducia e affidabilità negli esperti del settore. Questo lavoro fondativo prepara il terreno per un'esplorazione dettagliata dell'equità algoritmica nelle applicazioni di modellazione degli utenti, in particolare quelle che coinvolgono le reti neurali a grafo (abbreviate con GNN, dal termine inglese). Sebbene le GNN abbiano mostrato un enorme potenziale nel convertire i dati utente derivati da strutture a grafo in informazioni utili, le valutazioni tradizionali spesso trascurano le considerazioni di equità, concentrandosi principalmente su metriche di accuratezza.

Nella nostra ricerca, abbiamo valutato l'equità nelle GNN allo stato dell'arte per la modellazione comportamentale degli utenti, rilevando potenziali discriminazioni basate su diversi paradigmi di profilazione utente. Dalla successiva analisi dei limiti di utilizzo di metriche binarie, che semplificano eccessivamente gli scenari del mondo reale forzando attributi multi-valore in categorie binarie, abbiamo esteso quattro metriche di equità esistenti per adattarle a scenari multi-classe. Questa espansione consente una valutazione più dettagliata e accurata dell'equità, migliorando la comprensione dell'influenza che i sistemi di IA hanno su diversi gruppi demografici.

Infine, per affrontare le potenziale discriminazioni rilevate nella modellazione degli utenti basata su GNN, abbiamo introdotto FAME (acronimo di Fairness-Aware MEssages), una innovativa strategia di mitigazione dei bias che si applica direttamente durante l'addestramento delle reti neurali considerate. Inoltre, abbiamo sviluppato due sistemi denominati FairUP e GNNFairViz. Il primo standardizza la valutazione delle pre-

stazioni delle GNN per la modellazione utente, mentre il secondo offre visualizzazioni interattive per aiutare gli utenti a esaminare e identificare i bias nelle architetture neurali.

Nel complesso, questa dissertazione affronta sfide critiche nell'analisi dell'equità per la modellazione comportamentale degli utenti, concentrandosi sullo sviluppo etico dell'IA. Inizialmente, abbiamo progettato sistemi decisionali allineati ai principi dell'IA Responsabile, rispettando le normative europee. Il contributo principale è rappresentato dalla valutazione dell'equità nella modellazione degli utenti utilizzando le GNN, introducendo metodi per valutare e mitigare i bias. Inoltre, abbiamo esteso le metriche binarie di equità a scenari multi-valore per fornire una valutazione più accurata. Lo sviluppo di strumenti standardizzati per la valutazione e la visualizzazione delle prestazioni di equità dei modelli di GNN migliora la coerenza, la trasparenza e la comprensione dell'impatto dei modelli stessi. Tali contributi evidenziano l'importanza di integrare l'equità nella progettazione dei sistemi di IA, ponendo le basi per future ricerche e applicazioni dell'IA Responsabile.

Parole chiave: Modellazione degli Utenti, Equità algoritmica, Intelligenza Artificiale Responsabile, Reti Neurali a Grafo.

Contents

List of Acronyms

The following acronyms are used throughout the doctoral dissertation.

A-FAME	Attention-based Fairness-Aware Messages
AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CatGCN	Categorical Graph Convolutional Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
EO	Equal Opportunity
FAME	Fairness-Aware Messages
FMP	Fair Message Passing
FNR	False Negative Rate
FPR	False Positive Rate
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HCAI	Human-Centered Artificial Intelligence
HCI	Human-Computer Interaction
HGAT	Heterogeneous Graph Attention Network

HGN	Heterogeneous Graph Network
IR	Information Retrieval
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
ML	Machine Learning
OAE	Overall Accuracy Equality
RHGN	Relation-aware Heterogeneous Graph Network
RNN	Recurrent Neural Network
RQ	Research Question
RS	Recommender System
SP	Statistical Parity
TE	Treatment Equality
TNR	True Negative Rate
TPR	True Positive Rate
UI	User Interface
XAI	Explainable Artificial Intelligence
XUI	Explainable User Interface

List of Figures

List of Tables

Chapter 1

Introduction

All you need is the plan, the road map, and the courage to press on to your destination.

Earl Nightingale

In today's digital age, whether intentional or not, people's lives are inevitably intertwined with Artificial Intelligence (AI) systems. Among the most widespread and utilized tools, Information Retrieval (IR) systems and Recommender Systems (RSs) effectively and efficiently provide relevant information to end-users based on their needs, personality traits, and context. Defined broadly, AI comprises any technique that enables computers to mimic human behavior and reproduce or excel over human decision-making to solve complex tasks independently or with minimal human intervention [61]. As these systems become increasingly sophisticated and pervasive, from online shopping [107, 131, 339] and social networks [164, 256, 387] to healthcare [67, 229] and finance [59, 148], the ethical considerations and the related implications on society surrounding their implementation and decision-making processes have gained significant prominence. At this point, ensuring transparency, equity, and sustainability is essential in guiding the development of AI or, more specifically, Machine Learning (ML) systems [235].

The rising importance of research fields like Responsible AI [95] and Human-Centered AI (HCAI) [61, 305] further underscores the need for automated systems that not only perform well but also align with societal values and human rights.

Contextually, **algorithmic fairness** [63, 77, 104, 197, 231, 248, 255, 383] has emerged as a pivotal aspect of this human-centered perspective. It aims to design technologies that are reliable and beneficial to all users and prevent perpetuating or exacerbating existing inequalities. Therefore, developing comprehensive and inclusive fairness approaches is crucial to guaranteeing ML systems operate equitably. Establishing fairness in AI involves not only detecting and mitigating biases but also fostering trust and accountability in these systems. The ultimate goal is to implement frameworks and models that support social good, respect user autonomy, and promote justice across demographic groups.

In this scenario, the European Union has established a series of guidelines and laws to ensure the ethical development and deployment of AI systems. The *Ethics Guidelines for*

Trustworthy AI[1] define ethical principles and key requirements that any system should meet in order to be deemed trustworthy. Besides, the *Digital Markets Act*[2] and *Digital Services Act*[3] aim to regulate digital services and platforms, ensuring fair competition and protecting users' rights. The forthcoming *EU AI Act*[4] seeks to provide a comprehensive regulatory framework, addressing risks and setting requirements for trustworthy AI. Adhering to these regulations is crucial to safeguarding individuals' rights, e.g., not to be subjected to automated decisions without explicit consent, to receive clear explanations for decisions made by automated decision-making systems, and to be protected against discrimination.

To address the described ethical challenges, in the presented doctoral thesis, our initial work focuses on developing a Responsible AI framework tailored to loan approval processes. This framework aligns with the principles set out by the European regulations, emphasizing transparency and equity. By incorporating these principles, our research aims to build trust and reliance among domain experts in AI systems, showcasing how explainability and fairness can enhance decision-making processes. A novel scale is proposed to evaluate the system's effectiveness and explainability, and usability tests are conducted to measure satisfaction among diverse stakeholders, including loan officers, data scientists, and researchers.

The continuous interaction with the aforementioned systems generates vast amounts of personal data. To harness this data, **user modeling** (or user profiling) techniques are employed to construct accurate representations of users [103, 267]. These representations are essential for personalizing and enhancing users' experiences by predicting their preferences and behaviors. User modeling holds considerable value in various applications, such as IR [30, 324], RSs [278, 344, 359], e-commerce [107, 131], and social networks [295, 307], where understanding user behavior is fundamental for delivering relevant and engaging content. In this light, our preliminary and foundational work on Responsible AI set the stage for a more focused investigation into algorithmic fairness and paved the way for the core research of the presented doctoral project, which centers on addressing algorithmic fairness in behavioral user modeling tasks.

Focusing on the technological perspective, users' behaviors can be naturally shaped with graph structures, considering the nodes as users or items and the edges as the relationships or interactions between them. **Graph Neural Networks** (GNNs) [69, 138, 147, 180, 281, 333, 375, 378, 386] have shown significant potential in modeling graph data and transforming user interactions into actionable knowledge. However, traditional evaluations of GNNs have primarily focused on accuracy-based metrics, overlooking critical fairness considerations. Our research aims to move beyond accuracy, assessing the presence of biases in applications where users characterized by certain sensitive personal characteristics (e.g., age or gender) may be systematically disadvantaged.

In analyzing fairness in user modeling, we recognized the limitations of traditional

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`. Accessed March 30, 2025.

[2] `https://tinyurl.com/eu-digital-markets-act-en`. Accessed March 30, 2025.

[3] `https://tinyurl.com/eu-digital-services-act-en`. Accessed March 30, 2025.

[4] `https://artificialintelligenceact.eu/`. Accessed March 30, 2025.

binary fairness metrics. The adoption of these metrics often distorts real contexts by forcing non-binary attributes into binary categories, leading to incorrect evaluation. To address this challenge, we extended four existing algorithmic fairness metrics to accommodate multiclass and multigroup scenarios. This approach provides a more nuanced and accurate fairness assessment, allowing for a better understanding of how different user groups are impacted by AI systems. Our experiments with real-world datasets and state-of-the-art GNN models demonstrated that these generalized fairness metrics offer a more detailed and effective analysis of disadvantaged sensitive groups, ultimately bridging the gap between theoretical definitions and practical applications in bias detection.

Bias mitigation is the key topic for creating equitable decision-making processes by reducing existing discrimination [137, 212, 232, 342, 384]. To tackle the potential biases present in GNNs, we developed a specific approach for binary user modeling tasks. We introduced FAME (Fairness-Aware MEssages), an in-processing method that modifies the GNN training's message-passing algorithm to promote fairness. This technique, along with its attention-based variant, A-FAME, reduces biases by adjusting messages based on the sensitive attributes of connected nodes, thereby improving fairness in node classifications.

Recognizing the need for standardized evaluation in GNN-based user modeling, we developed FairUP, a framework that harmonizes input processing and integrates comprehensive fairness analysis components. It supports both pre-processing and post-processing fairness techniques, providing a holistic approach to evaluating and mitigating biases in datasets. In the same research direction, to facilitate the practical application of our fairness techniques, we created GNNFairViz, a visual analytics tool designed to analyze biases in GNNs. This tool offers interactive visualizations that help users inspect and diagnose biases, providing deeper insights into how attribute and structural biases influence model outcomes.

In summary, this dissertation embarks on a journey from the initial development of a Responsible AI framework to a comprehensive exploration of fairness in GNN-based user modeling. By advancing fairness metrics, introducing novel bias mitigation techniques, and developing standardized evaluation frameworks and visual analytics tools, this work contributes to the ethical deployment of AI systems, ensuring they operate transparently, equitably, and sustainably.

Research challenges The challenges addressed in the presented doctoral project are outlined below:

1. Developing automated decision-making systems that reflect HCAI and Responsible AI principles and respect European regulations;

2. Assessing and mitigating fairness in behavioral user modeling applications that employ GNNs;

3. Extending existing algorithmic fairness metrics from binary to multiclass and multigroup scenarios to tackle the observed limitations of binary metrics, which often distort real-world contexts;

4. Implementing unified frameworks for a standardized fairness evaluation and visualization.

1.1 Dissertation Structure

The doctoral work on fairness analysis of GNNs in behavioral user modeling presented in this manuscript is organized into thematic chapters covering all stages of the research. While the chapters are not ordered chronologically, the primary aim is to convey the storyline of the thesis in a logical and coherent fashion.

Before diving into the core part of the research, Chapter 2 delineates the foundational concepts necessary for the subsequent discussions in this dissertation. This chapter serves as an in-depth exploration of the four main topics integral to our study: *Human-Centered Artificial Intelligence* (Section 2.1), *Algorithmic Fairness* (Section 2.2), *Graph Neural Networks* (Section 2.3), and *User Modeling* (Section 2.4). This background is essential for understanding the specific challenges and methodologies of the presented research and for appreciating the broader implications of our proposed solutions and related findings in the fields.

The core of the thesis is structured into several chapters, each focusing on a different stage of our research. Every chapter starts with a motivation section that introduces the study and outlines its specific objectives; thereafter, a review of relevant literature is provided. These sections set the scene for a comprehensive exploration of the topic, emphasizing its relevance and significance within the field. After the motivation, the chapter presents the methodology, explaining the theoretical frameworks and empirical strategies used, explicitly referring to the main topics described in Chapter 2, and resources used in the study. This is followed by a description of the experimental settings, detailing the operational conditions. Subsequently, the chapter delves into data evaluation, integrating statistical analysis and critical interpretation to rigorously assess the findings. The results are analyzed at the conclusion of each chapter, including a comparison of insights with relevant literature. Furthermore, a summary is presented to encapsulate the innovative proposals and the resulting findings. This iterative chapter format ensures a thorough and systematic exploration of each research phase, allowing for a deep and coherent understanding of the contributions of this doctoral work to the field.

In Chapter 3, the initial study of this doctoral research is discussed. In particular, we illustrate the implementation of a Responsible AI framework equipped with XAI and fairness components. The latter will receive more emphasis in the description, given the primary focus of the dissertation.

Chapter 4 presents the fairness assessment of state-of-the-art GNNs designed for behavioral user modeling tasks in real-world binary scenarios. Moreover, in the same chapter, we discuss the existing challenges for fairness assessment in such a scenario, paving the way for our research in multiclass and multigroup bias detection.

Chapter 5 introduces a novel approach for assessing algorithmic fairness in real-world scenarios by proposing multigroup and multiclass fairness metrics. By extending fairness assessment beyond binary scenarios, we provide a more comprehensive understanding of

model biases and improve the detection and mitigation of discrimination.

In Chapter 6, we present FAME (short for Fairness-Aware MEssages), an innovative bias mitigation algorithm for GNNs that adjusts the message-passing procedure during the model training based on sensitive attribute differences to promote fairness in binary user modeling tasks. Two variants of this approach are proposed: FAME, suitable for GCN-based models, and A-FAME (short for Attention-FAME), for GAT-based models.

Chapter 7 describes the implementation of two novel tools for standardizing algorithmic fairness computation and visualization. In particular, FairUP empowers researchers and practitioners to simultaneously examine classification performance and algorithmic fairness metrics scores of GNN models, while GNNFairViz, developed in collaboration with the Fudan University (Shanghai, China), allows them to gain insights into how model biases occur through a visual analytics framework.

In Chapter 8, we summarize the contributions presented in the dissertation, including scientific works not specifically related to the core topic, underline potential limitations detected during the studies, and discuss future research directions.

1.2 Research Contributions

This section outlines the significant research achievements of this thesis, each corresponding to a core chapter of the study. Underneath, each subsection offers a brief overview of the research carried out in the respective phase of this study, summarizing the key contributions and progress made in the field. After the summary, the formal citation of the resulting article (being already published, under review, or planned to be submitted) is included to provide a comprehensive academic context. This format enables us to showcase our research contributions, making it simple for readers to contextualize the specific academic impact of your work.

1.2.1 Survey on User Modeling and User Profiling

Chapter 2 establishes the basis for the dissertation by examining the crucial subjects discussed in the doctoral project. In particular, to identify the existing research gaps in the core topics' literature, we provide a comprehensive and in-depth review of the user modeling field (Section 2.4), with a specific focus on modern paradigm shifts and recent advances in this area. The resulting survey and the three tutorials we gave at international scientific conferences aim to set a milestone in user modeling research by proposing novel definitions and redefining the taxonomy of the field.

Resulting publication The content of the aforementioned background section on user modeling is included in the article "*User Modeling and User Profiling: A Comprehensive Survey*" under review at User Modeling and User-Adapted Interaction (UMUAI) Journal. A pre-print version of this article is already available [267]. The related tutorial summaries are the following: "*Paradigm Shifts in User Modeling: A Journey from Historical Foundations to Emerging Trends*" [265] published at the 32nd ACM Conference

on User Modeling, Adaptation and Personalization (UMAP 2024); "*Leveraging Graph Neural Networks for User Profiling: Recent Advances and Open Challenges*" [262] published at the 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023); "*Tutorial on User Profiling with Graph Neural Networks and Related Beyond-Accuracy Perspectives*" [264] published at the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2023).

1.2.2 Fairness Analysis in Generic Machine Learning Models

The study presented in Chapter 3 describes the Responsible AI framework developed for managing the ML model life cycle, emphasizing explainability methods, bias detection approaches, and pre-processing bias mitigation techniques in generic ML models. This framework functions as a decision-making system specifically designed for loan approval processes. It includes an Explainable AI component that enables users to understand the rationale behind model decisions and a fairness component that implements a bias detection method followed by a pre-processing bias mitigation strategy to ensure the unbiased nature of data used in model training.

Resulting publication The content of the chapter is included in the article "*The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes*" [269] published at the International Journal of Human-Computer Interaction (IJHCI).

1.2.3 Fairness Assessment of Graph Neural Networks in Binary User Modeling Scenarios

The research presented in Chapter 4 deals with evaluating the fairness of GNN models in behavioral user profiling tasks using real-world datasets. Two state-of-the-art GNN models, CatGCN and RHGN, are assessed for bias using fairness metrics, including statistical parity, equal opportunity, overall accuracy equality, and treatment equality. The study highlights the differences in computing fairness scores due to different user modeling paradigms and the need for debiasing processes in GNN models to ensure fairness. It also suggests considering disparate impact and disparate mistreatment metrics for a comprehensive assessment.

Additionally, this chapter discusses challenges in algorithmic fairness within user modeling using GNNs, emphasizing limitations in evaluating fairness in binary scenarios and using absolute difference scores. Through a case study, we underline how these practices can lead to misleading evaluations and hinder the identification of disadvantaged groups. We thus advocate for multiclass assessments for a deeper understanding of disadvantaged groups to enable more effective interventions.

Resulting publications The content of the chapter is included in the following papers: "*Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact*

and Mistreatment in Behavioural User Profiling" [261] published at the 31st ACM International Conference on Information & Knowledge Management (CIKM 2022); "*What Are We Missing in Algorithmic Fairness? Discussing Open Challenges for Fairness Analysis in User Profiling with Graph Neural Networks*" [268] published at Advances in Bias and Fairness in Information Retrieval, Proceedings of the 4th International Workshop on Algorithmic Bias in Search and Recommendation (BIAS 2023), co-located with the 45th European Conference on Information Retrieval (ECIR 2023); "*Recent Advances in Fairness Analysis of User Profiling Approaches in E-Commerce with Graph Neural Networks*" [263] published at Discussion Papers Proceedings of the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023).

1.2.4 Multiclass and Multigroup Fairness Assessment

The study described in Chapter 5 presents a new method for evaluating algorithmic fairness in practical situations by suggesting fairness metrics for multiple groups and multiple classes. By expanding the assessment of fairness beyond binary situations, the research aims to offer a more thorough insight into model biases and enhance the identification and reduction of discrimination. Through the evaluation of GNN-based models on diverse datasets, the proposed multigroup and multiclass fairness metrics offer a more nuanced perspective on fairness, highlighting the importance of considering the distribution of classes and groups in fairness evaluations. The contribution of these novel metrics lies in their ability to enhance fairness assessments in machine learning models, particularly in user profiling tasks, by addressing the limitations of traditional binary fairness metrics.

Resulting publication The chapter's content is included in the following paper: "*Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks*" [266] published at Minds and Machines Journal for Artificial Intelligence, Philosophy and Cognitive Science (Special Issue on "*Interdisciplinary Perspectives on the (Un)fairness of Artificial Intelligence*").

1.2.5 Bias Mitigation for Graph Neural Networks in Binary User Modeling Scenarios

In the study described in Chapter 6, we propose FAME (Fairness-Aware MEssages), a novel in-processing bias mitigation algorithm designed specifically for GNNs. The standard FAME is designed for compatibility with GCN-based models. Along with its variant A-FAME (Attention-FAME) tailored for GAT-based models, it operates by adjusting the message-passing process within GNNs based on differences in sensitive attributes. This adjustment aims to reduce biases present in the network and promote fairness in binary user modeling tasks. By modifying the message-passing algorithm to account for sensitive attribute discrepancies between connected nodes, FAME offers a promising approach to addressing bias in GNNs. Through experimental validation on three datasets compared with six baselines, we demonstrate the efficacy of FAME and underscore the

effectiveness of in-processing bias mitigation techniques for enhancing fairness in GNNs. By producing accurate and fair node classifications, FAME lays a robust foundation for further exploration and development of bias mitigation strategies in this context.

Resulting publication The content of the chapter is included in the following paper: "*GNN's FAME: Fairness-Aware MEssages for Graph Neural Networks*" [270] accepted at the 33rd ACM ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2025).

1.2.6 Frameworks for Fairness Analysis of Graph Neural Network-based Models

The research presented in Chapter 7 illustrates the implementation of two innovative frameworks designed to assess algorithmic fairness in GNN-based models. With the described work, we aim to develop novel tools that are able to standardize fairness scores computation and visualization. We provide a comprehensive analysis of the frameworks, namely FAIRUP and GNNFAIRVIZ. We describe in detail their architectures, along with the specific components and functionalities of each tool. Our discussion emphasizes the contributions they make to the literature on algorithmic fairness, particularly in the context of GNNs. We highlight the advantages these frameworks offer in advancing the evaluation and understanding of GNN fairness to enhance reproducibility and transparency. Additionally, we critically assess the limitations inherent in the developed tools and propose potential areas for future improvements based on expert feedback.

Resulting publications The content of the chapter is included in the following papers: "*FairUP: A Framework for Fairness Analysis of Graph Neural Network-Based User Profiling Models*" [1] published at the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023); "*GNNFairViz: Visual Analysis for Graph Neural Network Fairness*" [376] published at IEEE Transactions on Visualization and Computer Graphics (TVCG) Journal.

Chapter 2

Background

If you don't know where you come from, then you don't know where you are; and if you don't know where you are, then you don't know where you're going.

Terry Pratchett

This chapter lays the foundation for the dissertation by exploring the four pivotal topics addressed in the presented work: *human-centered artificial intelligence* (Section 2.1), *algorithmic fairness* (Section 2.2), *graph neural networks* (Section 2.3), and *user modeling* (Section 2.4). Each section of this chapter begins with a historical overview, tracing the evolution of the respective fields and highlighting key developments that have shaped their current state. Following this, the discussion focuses on the essential terminology and fundamental concepts of each topic. Furthermore, to guide the reader, the outset of every section includes a brief outline of its structure, ensuring a coherent flow of information and facilitating a deeper understanding of the domains.

2.1 Human-Centered Artificial Intelligence

The use of Artificial Intelligence (AI) systems in various fields creates high expectations for their benefits, as well as concerns for possible misuse [327]. These systems are widely implemented in daily activities, making decisions that have an impact on society, particularly using our personal data and trace data [55]. The models used in AI systems and their outcomes are often difficult to understand, and they may have embedded bias, especially against minority groups, due to historical data and algorithmic choices [63]. Concerns also arise from privacy issues, human rights challenges, and the generation of illusions of meaning [55, 304]. Although such systems have been largely driven by technology-centered design, and due to the mentioned potential societal consequences, researchers in Human-Computer Interaction (HCI) and AI are now focusing on exploring a novel area called **Human-Centered Artificial Intelligence** (HCAI). The use of the term HCAI is growing, reflecting a desire for decision-making systems to serve people and

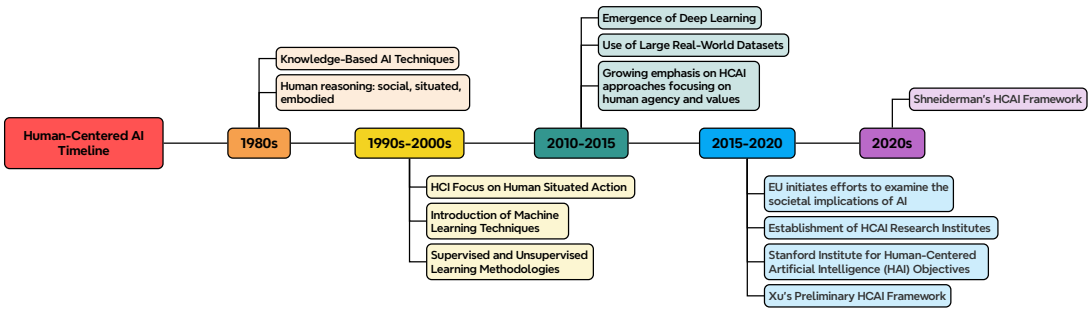


Figure 2.1. Timeline reporting the major events of the HCAI history.

address concerns about AI's potential to exploit and mislead. However, recent studies reveal how the concept of HCAI has different meanings for different people [61]. For instance, the "*human*" might be the subject of AI algorithmic study, the user of AI products, or an agent in the design of the AI system itself. Moreover, HCAI is sometimes used aspirationally, similar to "sustainable mining" or "trusted autonomy," with considerable debate about whether and how it can be achieved.

This background section is dedicated to examining the progression and nuances of the HCAI field. The analysis starts with a historical overview, mapping the evolution of HCAI from its initial concepts to the sophisticated frameworks of today, emphasizing the shift toward augmenting human capabilities and improving user interactions (Section 2.1.1). Following the historical discourse, we will explore key concepts related to HCAI, which include interdisciplinary approaches integrating cognitive sciences, ethics, technology, and regulations (Section 2.1.2). This part aims to establish a foundational understanding of the principles and methodologies that form the basis of HCAI, facilitating a deeper discussion on its applications and implications.

2.1.1 Historical overview

To clearly understand how the contributions presented in this thesis stem into the context of HCAI, it is worth starting with a historical overview of the evolution from AI and HCI to HCAI research, which is based on the work of Capel and Brereton [61]. Figure 2.1 displays the major milestones in this field.

During the 1980s, the initial AI techniques depended on a knowledge-based strategy. This allowed computers to reason using logical inference rules and knowledge statements that were automatically coded into formal languages [124]. However, this approach had its limitations, such as the difficulty of explaining knowledge in detail, especially implicit knowledge [53]. Several studies in HCI have argued that human reasoning occurs in a social context, is situated, and is embodied. It involves actions that are often improvised in the real world, making it too complex to be replicated by formal models [99, 318]. For many years, HCI researchers have focused on studying how machines could be designed as effective tools for human-situated action rather than trying to automate

human reasoning [143].

The introduction of *machine learning* (ML) techniques has significantly transformed the field of AI. The goal of ML is to uncover meaningful relationships and patterns from data without explicitly programming computers with knowledge and instructions [165]. This facilitates the automation of constructing analytical models using inferential statistics [39]. The advancement of machine learning has been driven by the development of new programming frameworks, increased accessibility to the necessary computational resources, and the availability of vast amounts of data. Through the analysis of historical data and the identification of patterns within large and complex datasets, machine learning has the potential to generate consistent and trustworthy classifications that can inform decision-making processes [165]. Supervised and unsupervised learning are the two main methodologies for machine learning algorithms.

Supervised learning entails the process of training models using labeled datasets in order to make predictions for new data, with particular emphasis on tasks such as classification and regression [177]. The precision of this approach is heavily contingent upon the quality of the training data.

Unsupervised learning, in contrast, operates on unlabelled data to elucidate latent patterns, predominantly employed for clustering and data dimensionality reduction, thereby enabling the identification of inherent data structures and streamlining analysis [177]. Human involvement is essential in supervised learning to label datasets and train algorithms for accurate predictions. Individuals are involved in various actions, such as categorizing training data, assessing algorithms, modifying models, and validating ML processes in human-in-the-loop and interactive ML paradigms [106]. Conversely, unsupervised learning algorithms do not rely on labeled data and instead discover patterns and absorb information from the input data through their own processes [124].

In the last decade, *deep learning* (DL) emerged as a specialized subset of ML, employing artificial neural networks with multiple layers, thus leading to the term "deep". This advancement over classical machine learning models increases the ability to process and learn from raw data directly without the need for manual feature extraction. The capability of DL to handle and analyze data through these intricate, multi-layered networks allows it to address complex and large-scale problems in fields like computer vision [64] and natural language processing [202].

However, the complexity of these models also leads to challenges in interpretability, often obscuring the understanding of how decisions are made within these deep networks [388]. The ambiguity surrounding the importance of the features and the reasoning behind the decision-making process makes it challenging to understand neural networks' behavior. The problem grows even more complex due to the lack of a clear connection between the network's parameters and concrete physical meanings. Hence, the black-box nature of DL raises concerns about interpretability and bias.

Through the analysis of large real-world sources, including e-commerce platforms, video data, and social media, ML and DL methods benefit from accessing more comprehensive contextual information compared to earlier symbolic AI approaches, which facilitate the extraction of meaningful insights from the data. Using extensive datasets

with digitally recorded contextual information makes it possible to mimic and simulate human reasoning, interaction, and language. According to Blackwell [40], this procedure reduces the individual with contextual understanding to a mechanical collector of interaction information. Our investigation should move away from non-situated cognition and towards non-human interaction. The continuous effort in AI research to imitate, reproduce, and replace human abilities and behaviors, such as conversational agents and humanoid robots, has prompted a growing fascination with HCAI, which focuses on the agency and values of human users.

In light of the increasing focus on HCAI approaches, global organizations such as the European Union have initiated efforts to examine the societal implications of AI and promote the development of decision-making systems that prioritize human well-being, with specific guidelines (e.g., *EU Ethics Guidelines for Trustworthy AI*[1]) and regulations (e.g., *EU Artificial Intelligence Act*[2]). The worldwide establishment of research institutes focusing on HCAI, e.g., at Stanford University, University of California Berkeley, and the Massachusetts Institute of Technology (MIT), reflects a shared commitment to advancing ethical AI technologies. The goal of these institutions is to build AI systems that can enhance human capabilities without replacing them. The three primary objectives proposed by the *Stanford Institute for Human-centered Artificial Intelligence* (HAI)[3] to guide research and design in the field are: (1) to technically reflect the depth characterized by human intelligence; (2) to improve human capabilities rather than replace them; (3) to focus on AI's impact on humans.

In accordance with these ideals, Xu [369] put forward a preliminary HCAI framework consisting of three fundamental components: (1) ethically aligned design, aimed at generating AI solutions that eschew discrimination, uphold fairness and justice, and do not substitute for human agency; (2) technology that faithfully replicates human intelligence, thereby refining AI technology to embody the complexity associated with human cognition; (3) human factors design, ensuring that AI solutions are explainable, comprehensible, useful, and usable. Xu's framework [369] is designed to offer individuals secure, efficient, healthy, and satisfying solutions for HCAI through a comprehensive approach. Significant definitions in the discipline have been offered by Shneiderman [303, 304, 305], asserting that HCAI is responsible for "*amplifying, augmenting and enhancing human performance in ways that make systems reliable, safe and trustworthy*". He posits an HCAI framework to encourage designers and researchers to carefully consider and discuss the key elements of automation and autonomy: (1) design for high levels of human control and high levels of computer automation so as to increase human performance; (2) understand the situations in which full human control or full computer control are necessary; (3) avoid the dangers of excessive human control or excessive computer control. Shneiderman's definition of HCAI [303, 304, 305] stresses the importance of user experience design and emphasizes placing humans at the center of the design process.

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`. Accessed March 30, 2025.

[2] `https://artificialintelligenceact.eu/the-act/`. Accessed March 30, 2025.

[3] `https://hai.stanford.edu/`. Accessed March 30, 2025.

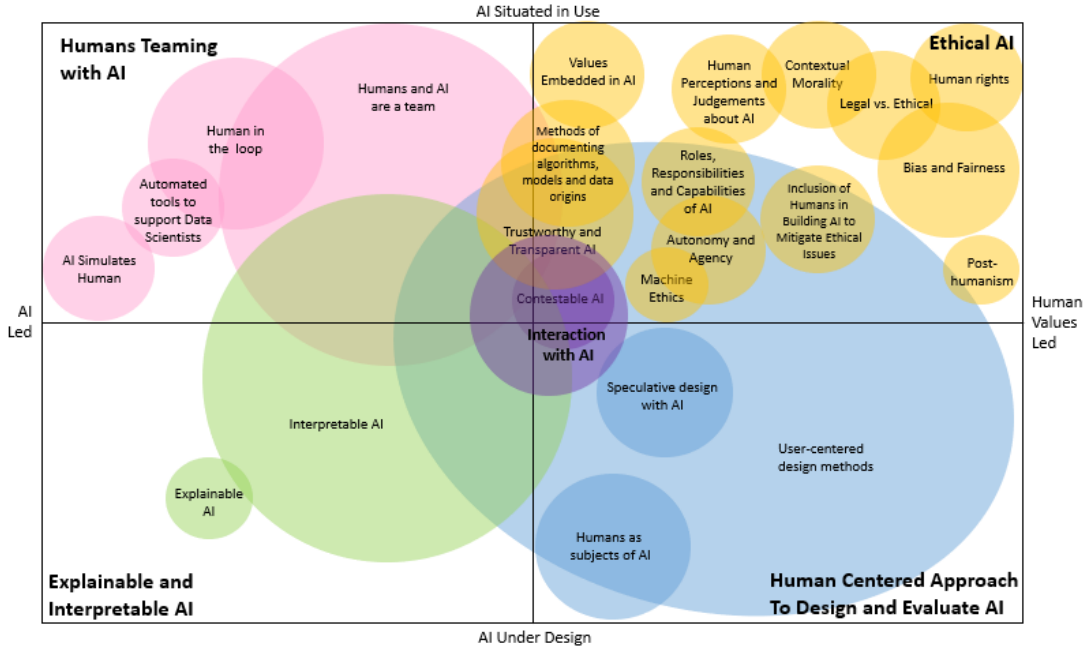


Figure 2.2. Map of the Human-Centered Artificial Intelligence research field by Capel and Brereton [61].

This involves prioritizing the measurement of human performance and satisfaction, as well as valuing customer needs while ensuring meaningful human control in the design and implementation of HCAI systems.

2.1.2 Related concepts

The current landscape of HCAI research has been drawn in a recently published survey [61]. A graphical representation of the map reporting the identified areas and corresponding subfields is displayed in Figure 2.2. Four principal areas of study have been outlined: *Explainable and Interpretable AI*, *Human-Centered Design Methods*, *Human-AI Teaming*, and *Ethical AI*. A fifth domain, referred to as *Interaction with AI*, is emerging at the intersection with all the others. Below we will provide a concise overview of the concepts associated with HCAI (not necessarily belonging to the above-mentioned mapping) that have been examined and discussed within the context of our research.

Explainable and Interpretable AI These areas of research encompass a variety of tools, methodologies, and frameworks designed to assist individuals in comprehending the decisions or forecasts generated by AI systems [6]. The development of explainable and interpretable AI is a direct response to the opaque nature of AI models, which often leave it ambiguous as to how or why an AI reached a specific decision or prediction. The concept of *explainability* serves the dual purpose of facilitating comprehension of a model's behavior as well as enhancing its overall performance [294]. Explainable AI (XAI)

has been firstly defined within the DARPA[4]'s XAI Program [133]: "*XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners*". While explainability mirrors the capability of a model to *explain* its outcomes, the concept of *interpretability* reflects a trend toward HCAI, indicating the capacity of humans to *interpret* and make sense of a system's decision-making processes or predictions [259].

Ethical AI The field of Ethical AI is known for its dedication to preserving fundamental human values and rights [298], as well as advocating for increased transparency in the advancement of AI technologies [169]. This concept is fundamentally concerned with safeguarding the rights and principles of individuals involved in the development and implementation of artificial intelligence, especially in sensitive scenarios.

Trustworthy AI As AI continues to proliferate and become integrated into human-serving operations across a range of application domains, important factors must be considered in developing systems that can be seen as *trustworthy* and *reliable*. The methodology used for collecting system data, as well as the design of algorithms and models derived from this data, must be disclosed [206]. Another important factor to consider is the adherence to design principles established by authoritative guidelines or enforced regulations, such as, respectively, the *EU Ethics Guidelines for Trustworthy AI* and the *EU Artificial Intelligence Act*, already referenced in Section 2.1.1. The guidelines emphasize that the concept of *trustworthy AI* consists of three primary elements: adherence to existing laws and regulations (*lawful AI*); compliance with society's ethical principles, even in cases where no regulations have been established yet (*ethical AI*); and resilience from both a technical and social standpoint to prevent unintended harm (*robust AI*). Four ethical principles are outlined by the authors that must be met for an AI system to be deemed trustworthy: *respect for human autonomy*, *preventing harm* to others, ensuring *fairness* in the AI system's predictions, and providing *explainability* of the AI system's outcomes.

Bias and Fairness The study of *fairness* in AI systems examines the potential consequences of *biases* within such systems, as well as how AI has the capability to perpetuate societal discrimination through its utilization of data and algorithms [63]. The ultimate objective of this research field is to develop AI systems that are not only more accurate but also fairer in their operations. This area also includes investigating the root causes of bias, beyond data and algorithms. Research on fairness constitutes one of the core topics of this thesis. A detailed overview of this field is provided in Section 2.2.

Responsible AI Not directly linked with the HCAI map, *Responsible AI* [95] refers to the development and use of AI technologies in ways that are ethical, transparent, and accountable, thus aligning closely with the general principles of HCAI. This approach emphasizes creating AI systems that not only support fairness, interpretability, and respect

[4]Defense Science Research Projects Agency.

for privacy and human rights but also prioritize human welfare and collaboration. By considering the societal impacts of decision-making systems, Responsible AI strives for inclusivity, actively mitigating biases in AI algorithms and data. It improves interactions between humans and machines by ensuring that AI deployments are aligned with human needs and contexts. Responsible AI reinforces the HCAI mandate to design, develop, and deploy AI technologies in a manner that is participatory and sensitive to human values, fostering trust and sustainability in AI applications.

2.2 Algorithmic Fairness

Fairness is a fundamental principle that governs human interactions, emphasizing the need for justice, equality, and impartiality. In societal terms, fairness often involves distributing resources, opportunities, and treatment so as not to favor certain individuals or groups over others unless there is a justified reason to do so [283]. In the research conducted by the linguist Wierzbicka [357], the word "fairness" is identified as a representation of cultural norms that dictate the regulation of human activities through explicit and implicit rules of engagement. Most participants commonly perceive these rules as universally applicable and justifiable. As societies evolve, the concept of fairness has been continually reinterpreted and applied in various contexts, reflecting the complexities and diversities of human values and ethical standards.

Contributions from various academic disciplines have significantly enhanced the discussion on the topic of fairness. Philosophers like Rawls in "A Theory of Justice" [283] have laid down principles defining justice in terms of fairness, emphasizing equal liberty and opportunity. Behavioral scientists, including Kahneman and Tversky [171, 326], have explored how biases influence perceptions of fairness, revealing the psychological support that alters human judgment. Contemporary humanities researchers have focused on the intersections of culture, gender, and historical contexts of fairness. In particular, Nussbaum addressed how societal structures and cultural norms influence women's capabilities and rights, offering a profound critique of gender bias and inequality [244]. Similarly, Spielhaus has examined how gender and religious identities shape interactions and policies in multicultural societies. Her insights into the dynamics of religious bias and discrimination provide a critical perspective on the historical and cultural dimensions of fairness, especially concerning the treatment of religious (i.e., Islamic) minorities [313].

With the advancements in technology, the principles of fairness extend to the AI and ML fields. In AI, fairness ensures that automated systems operate without inherent biases, promoting equal treatment across all user interactions [33, 101]. This is particularly critical as AI systems are increasingly employed in decision-making roles, from personalized advertising to automated customer service. The challenges of fairness become more pronounced in ML, where models learn from data to make predictions or decisions. ML models are indeed susceptible to reflecting or amplifying biases present in their training data. Addressing fairness in this context involves developing methods to detect, quantify, and mitigate biases, ensuring that such applications do not perpetuate historical discrimination [117]. This is essential in sensitive domains such as credit scoring [161, 190], job

recruitment [194, 195, 289], and healthcare [67, 229], where biased decisions can have profound consequences on individuals' lives. However, it also significantly impacts other domains, such as e-commerce [230, 339] and social networks [142, 295]. In particular, unfairness has a substantial impact on user experience and service delivery in these areas. The effects can manifest as less effective services for a considerable segment of users, including receiving inappropriate recommendations [383]. Such disparities can accumulate substantial economic and social consequences over time and across populations [33, 63, 77, 335].

From an ethical perspective, and as mandated by regulations such as the GDPR (specifically, Article 22[5]), digital platforms are obliged to avoid discrimination stemming from automated decision-making systems, particularly concerning sensitive attributes. This legal framework necessitates a nuanced understanding of fairness and discrimination, often calling for rigorous scrutiny of even minimal differences to prevent systemic biases.

Why *algorithmic*? The term "algorithmic" in **algorithmic fairness** highlights the role of algorithms in the broader debates of fairness within AI and ML. The origin of the word "algorithm" can be traced back to the late medieval Latin "algorismus", which derives from Al-Khwārizmī, the Islamic mathematician who authored the Arabic manuscripts describing the Indian arithmetic system [317]. The meaning of the term has evolved into the technical definition used in modern computer science: "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output" [78]. Algorithms are essentially sequences of instructions or sets of rules that enable computers to perform tasks, solve problems, or make decisions. These range from simple calculative formulas to complex models that process and analyze large datasets.

Discussing "algorithmic fairness" means focusing on these algorithms to ensure they operate without bias. In this context, the emphasis is given to how algorithms are applied in decision-making processes that significantly impact human lives across various domains like hiring, lending, criminal justice, healthcare, e-commerce, social networks, and recommendations. The primary concern is whether decisions made by algorithms are fair, unbiased, and equitable, particularly toward individuals or groups based on protected characteristics such as race, gender, or age. The "algorithmic" aspect underlines that issues of fairness arise from these computational processes and automated decision systems [104]. Algorithms have the potential to perpetuate or even amplify societal biases if they are not carefully designed and monitored.

Thus, ensuring algorithmic fairness involves a critical examination of both the input data and the algorithmic processes that analyze this data. This analysis is necessary to identify and mitigate biases, ensuring the algorithms promote fairness systematically. Addressing these issues involves not only adjusting the algorithms but also modifying the data handling practices and sometimes the objectives of the models themselves. A comprehensive algorithmic fairness approach helps build trust in AI applications that increasingly influence many aspects of our lives, and ensures that technological advance-

[5] `https://gdpr-info.eu/art-22-gdpr/`. Accessed March 30, 2025.

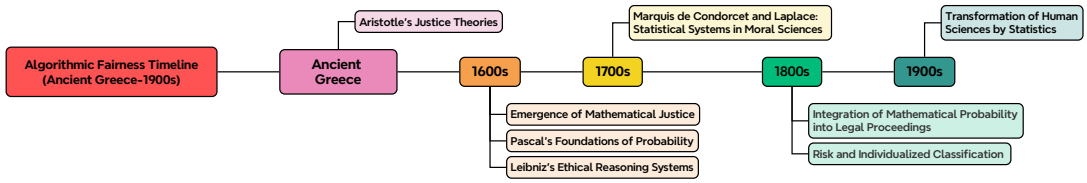


Figure 2.3. Timeline reporting the major events of the algorithmic fairness history (from Ancient Greece to 1900s).

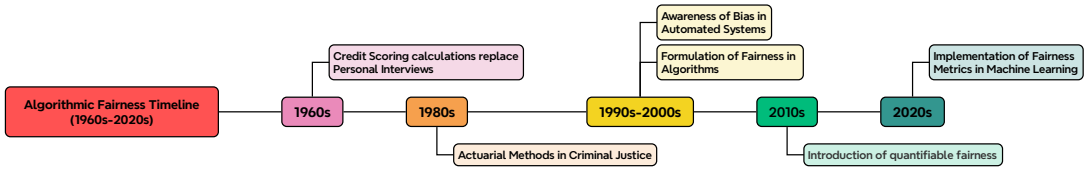


Figure 2.4. Timeline reporting the major events of the algorithmic fairness history (from 1960s to 2020s).

ments contribute positively to promote a fair and inclusive society by adhering to our collective ethical standards.

In this section, we will systematically explore the algorithmic fairness field, beginning with a historical overview, which traces the development of fairness concepts from philosophical origins to modern machine learning applications (Section 2.2.1). Following, we will introduce basic definitions of fairness to establish foundational terminology (Section 2.2.2), examine the causes of bias in machine learning (Section 2.2.3), and discuss types of fairness and their legal definitions (Section 2.2.4). We will finally delve into the metrics used to measure and detect algorithmic bias (Section 2.2.5) and the existing bias mitigation strategies (Section 2.2.6). The content of the last three sections is based on surveys and books published in the algorithmic fairness literature (i.e., [37, 63, 231, 238, 335]).

2.2.1 Historical overview

The concept of fair algorithms has a long-standing historical foundation interconnected with the adoption of mathematics and statistics within moral sciences, as accurately portrayed in Ochigame's article [245].

Figures 2.3 and 2.4 displays the major milestones in algorithmic fairness over the centuries and recent decades.

Over time, moral philosophers have often formulated notions of justice inspired by mathematical concepts. Aristotle investigated the principles of distributive and corrective justice using the notions of geometrical and arithmetical proportions [54]. However, it was not until the early modern period that more organized attempts to utilize mathematical

calculations in resolving political disputes concerning justice and fairness began to arise.

During the seventeenth century in England, mathematicians constructed tables for calculating "present value" in order to establish equitable conditions for specific agricultural leases, particularly for land under the ownership of the Church of England [92].

During the Enlightenment era, ethical inquiries were a focal point in the initial development of probability theory and in the early conceptualization of computational devices [91]. Pascal's calculations of equivalent expectations were driven by concerns regarding equitable distribution, particularly in the context of apportionment in gambling and uncertain business contracts such as insurance and shipping [136]. Pascal's work laid the foundation for subsequent developments in the field of probability. Leibniz, the creator of an early calculating device, also aimed to devise a comprehensive system of reasoning grounded in a clear and unambiguous formal language to address ethical disagreements [22].

During the eighteenth century, probabilists like Marquis de Condorcet and Laplace developed a statistical system for the moral sciences, specifically in jurisprudence [83], based on the ideas of Pascal and Leibniz. One noteworthy instance of the incorporation of mathematical probability into legal proceedings occurred with the evolution of the concept of "contractual fairness" within English law. As a result of these advancements, a rule was implemented in 1810, allowing contracts for the sale of reversions to be canceled if the price was deemed unfair [178].

During the mid-nineteenth century, the probabilistic approach to the moral sciences had declined in popularity. However, probabilistic and statistical calculations remained fundamental in establishing various normative claims about society. In the nineteenth century, the concept of "risk" became part of everyday language in the United States with the rise of corporate risk management [204]. This led to the practice of individualized risk classification, such as in life insurance, which later resulted in controversies over racial discrimination, with corporations charging differential rates based on race [44]. Mathematical statistics transformed the human sciences in the twentieth century. Optimization models influenced by the theory of "expected utility" expanded the use of statistical methods and risk classification systems (often called "actuarial" because of their origins in insurance) in capitalist institutions [105]. Actuarial methods became pervasive in the second half of the century. Credit scoring calculations became popular in the 1960s as a replacement for personal interviews [201].

During the 1980s, the criminal justice system, especially in the United States, began to increasingly rely on actuarial methods for risk assessment [157]. These methods utilized statistical models to predict the likelihood of reoffending based on data gathered from past criminal records. These can be considered the predecessors of decision support tools, such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [21]. While these methods were praised for their objectivity compared to human judgment, they also faced criticism for potentially reinforcing existing biases present in the criminal data, such as racial disparities in arrest and conviction rates.

Progressively, the 1990s and early 2000s saw growing awareness about the potential biases these automated systems might carry. By the mid-2000s, formal studies began

to propose mathematical formulations of fairness. The landmark papers by Dwork et al. [101, 102] introduced notions of fairness that could be mathematically quantified, setting the stage for a plethora of research into fair machine learning practices. These efforts have grown to include an array of disciplines, including law, ethics, sociology, computer science, and statistics.

In more recent years, the focus has shifted toward the development and implementation of fairness metrics in ML [33, 63, 231, 255, 341]. The community has proposed a variety of methods and procedures to quantify and address fairness, including concepts such as *disparate impact* and *disparate treatment*, and specific metrics like *statistical parity*, *equal opportunity*, *equalized odds*, and *overall accuracy equality*. These notions, which will be described in detail in the next sections, have been critical in guiding the development of algorithms that are not only efficient but also equitable. The integration of these fairness metrics into practical applications signifies a crucial step forward in making machine learning tools that can be used responsibly in society.

2.2.2 Basic definitions

There is often confusion about basic terminology in discussions regarding fairness, particularly in systems governed by algorithms. The terms *bias*, *unfairness*, and *discrimination* are commonly used interchangeably in algorithmic fairness literature [63, 77], even though they have distinct linguistic meanings. We will provide below the three different definitions for a clear and complete understanding of the field. Within this thesis, we align with the vision of considering these concepts as synonyms.

Bias Refers to a systematic inclination or predisposition that results in judgments that are consistently distorted and prejudicial in favor of or against an individual or group compared to another. This can occur either consciously or unconsciously and affects human reasoning, behaviors, and decision-making processes [326].

Unfairness Describes actions or situations where equitable treatment is not provided, resulting in unequal outcomes for individuals or groups. It encompasses situations where decisions or treatments are not based on merit or relevance but on prejudice, leading to injustice or partiality [283].

Discrimination Involves treating someone differently or less favorably because of certain characteristics, such as race, gender, age, or religion. It manifests in actions that exclude or restrict individuals based on group-based attributes, leading to the denial of opportunities or rights [17].

2.2.3 Causes of bias in machine learning

The current literature has uncovered many possible causes of unfairness in ML. As outlined by Mehrabi et al. [231] and Pessach and Shmueli [255], these causes can be categorized into four main groups:

- The datasets utilized for learning may already contain biases stemming from biased device measurements, historically biased human decisions, erroneous reports, or other sources. ML algorithms are designed to mimic the inherent biases that are already present in the data they are trained on.

- The presence of biases induced by incomplete data, including missing values, can lead to the creation of datasets that do not accurately reflect the characteristics of the intended population.

- Algorithms that prioritize reducing overall prediction errors may result in biases that favor majority groups over minorities. Additionally, unreliable algorithms, such as those with weak generalization abilities, can contribute to unfairness.

- Biases may arise due to "proxy" attributes that are connected to sensitive attributes (e.g., race, gender, and age) that are generally not acceptable for use in decision-making systems. Proxy attributes are not sensitive but can be utilized to infer sensitive attributes. In case the dataset contains such attributes, the ML algorithm may implicitly make decisions based on sensitive attributes while appearing to use legitimate attributes.

2.2.4 Fairness types and legal definitions

Depending on the specific context and the ethical considerations involved, fairness can be categorized into different types, each addressing particular aspects of equity. Understanding these distinctions is crucial for designing and deploying algorithms that adhere to desired ethical standards and societal norms. In Section 2.2.5, we will mathematically define the fairness metrics associated with each category shown below.

Individual fairness Focuses on the principle that similar individuals should be treated similarly by an algorithm. The goal here is to ensure consistent and equitable treatment for individuals who are alike in relevant respects. The underlying challenge is defining what makes individuals "similar" in a specific context and determining how to measure this similarity in a way that aligns with ethical guidelines. Metrics to measure this bias type are *fairness through unawareness*, *fairness through awareness*, and *counterfactual fairness*.

Group fairness Concerns the equitable treatment of different defined groups based on sensitive attributes such as race, gender, or age. The aim is to ensure that no group is systematically advantaged or disadvantaged by an algorithm. In particular, splitting the population into several groups and computing a statistical measure for each group, the selected measure should be equal across all groups. Group fairness metrics include, among others, *statistical parity*, *equal opportunity*, *equalized odds*, *overall accuracy equality*, *predictive parity*, and *treatment equality*.

Subgroup fairness This concept seeks to maximize the advantageous properties associated with both group and individual fairness. The approach contrasts with the two standard notions but utilizes them to optimize results. It selects a group fairness constraint and investigates the extent to which this constraint is upheld across a wide range of subgroups.

Beyond the above general categorization, discrimination has been depicted in the legal domain through the introduction of two primary definitions: *disparate impact* and *disparate treatment*. Along with these definitions, a third notion, namely *disparate mistreatment*, has been introduced to consider aspects overlooked by the others, as illustrated below.

Disparate impact Also referred to as *adverse impact*, it is the indirect (i.e., unintentional) discrimination that occurs when practices or systems appear to treat all individuals uniformly. This applies to scenarios in which a model exhibits disproportionate discrimination against specific demographic groups, even when the model does not directly utilize the sensitive attribute in predicting outcomes but rather relies on proxy attributes.

Disparate treatment Disparate treatment is the deliberate act of treating an individual in a differential manner based on their membership in a protected class, constituting a form of direct discrimination. Disparate treatment focuses on unfair treatment aimed at individual people, unlike disparate impact, which addresses discrimination at the group level.

Disparate mistreatment This concept involves the assessment of misclassification rates within user groups characterized by varying values of a sensitive attribute, as opposed to the evaluation of corrected predictions. Furthermore, the concept of disparate mistreatment holds importance in situations where the misclassification cost varies based on the specific demographic group impacted by the error.

2.2.5 Fairness metrics

This section aims to present a comprehensive overview of the predominant fairness metrics utilized in ML classification tasks. It will include detailed descriptions and mathematical definitions of these metrics. As is customary in algorithmic fairness literature, the depicted metrics pertain to the scenario in which both the target class and the sensitive attribute are *binary*. The corresponding notation is outlined in Table 2.1.

Throughout the presented thesis, the specific metrics used in bias detection will refer to the following definitions.

Fairness through unawareness An algorithm is deemed to be fair if it does not explicitly incorporate any sensitive attributes into its decision-making processes.

Symbol	Description
$P(w \mid z)$	Probability of the event w given the event z
$f : x \to y$	Algorithm (or ML model) with input x and output y
$y \in \{0, 1\}$	Actual outcome
$\hat{y} \in \{0, 1\}$	Predicted outcome (by the algorithm or ML model)
$s \in \{0, 1\}$	Sensitive attribute (e.g., race, gender)
$y_i,\ \hat{y}_i,\ s_i$	Generic values of the actual outcome, predicted outcome, and sensitive attribute (either 0 or 1)
TPR	True positive rate
FPR	False positive rate
TNR	True negative rate
FNR	False negative rate

Table 2.1. Notation used in the description of fairness metrics

Fairness through awareness An algorithm must provide equitable predictions for individuals who exhibit similar characteristics in order to be considered fair. Essentially, when people have similar measurements for a specific task, they should experience similar outcomes.

Counterfactual fairness An algorithm is fair if it remains unchanged in a counterfactual scenario where the value of an individual's sensitive attribute is different while all other conditions remain the same.

$$P(\hat{y}_{s\leftarrow 1} = \hat{y}_i \mid s = 0) = P(\hat{y}_{s\leftarrow 0} = \hat{y}_i \mid s = 0) \tag{2.1}$$

where $\hat{y}_{s\leftarrow 1}$ denotes the predicted outcome in the counterfactual scenario in which the sensitive attribute s is changed.

Statistical parity Also known as *demographic parity*, *group fairness*, or *equal acceptance rate*, defines fairness as an equal likelihood for every group to be assigned to the positive class. The predictions should be independent of sensitive attributes.

$$P(\hat{y} = 1 \mid s = 0) = P(\hat{y} = 1 \mid s = 1) \tag{2.2}$$

Disparate impact metric Often equated to statistical parity, this metric is designed to mathematically represent the legal notion of the disparate impact definition. It requires a high ratio between the positive prediction rates of both sensitive groups.

$$\frac{P(\hat{y} = 1 \mid s = 0)}{P(\hat{y} = 1 \mid s = 1)} \geq 1 - \varepsilon \tag{2.3}$$

where ε commonly takes the value of 0.2. In this situation, the metric corresponds to the "*80%-rule*". Also known as "*four-fifth rule*", it is defined by the *U.S. Equal Employment*

Opportunity Commission (Title VII, 29 CFR Part 1607[6]) and prescribes that any group (categorized by race, orientation, or ethnicity) with a selection rate of less than four-fifths (i.e., 80%) of the group with the highest rate is indicative of *disparate impact*, resulting in discriminatory effects on a protected group.

Equal opportunity Demands that the likelihood of a subject in a positive class receiving a positive prediction be the same for each sensitive group, i.e., the TPR of the ML model should be equal across groups.

$$P(\hat{y} = 1 \mid y = 1, s = 0) = P(\hat{y} = 1 \mid y = 1, s = 1) \tag{2.4}$$

Equalized odds Also referred to as *conditional procedure accuracy equality*, it is similar to *equal opportunity*, but instead of considering only the TPR, equalized odds simultaneously take into account the FPR.

$$\begin{cases} P(\hat{y} = 1 \mid y = 1, s = 0) = P(\hat{y} = 1 \mid y = 1, s = 1) \\ P(\hat{y} = 1 \mid y = 0, s = 0) = P(\hat{y} = 1 \mid y = 0, s = 1) \end{cases} \tag{2.5}$$

Overall accuracy equality Defines fairness as the equivalent likelihood of a subject from either the positive or negative class to receive a correct prediction. This means that the classification accuracy should be the same for every sensitive group.

$$\begin{aligned} &P(\hat{y} = 0 \mid y = 0, s = 0) + P(\hat{y} = 1 \mid y = 1, s = 0) = \\ &= P(\hat{y} = 0 \mid y = 0, s = 1) + P(\hat{y} = 1 \mid y = 1, s = 1) \end{aligned} \tag{2.6}$$

Predictive parity A classifier is deemed to satisfy this criterion when both protected and unprotected groups exhibit an equivalent *positive predictive value* (PPV), representing the probability that a positive result truly corresponds to the positive category.

$$P(y = 1 \mid \hat{y} = 1, s = 0) = P(y = 1 \mid \hat{y} = 1, s = 1) \tag{2.7}$$

Treatment equality Defines fairness as the equal error rate made by the classifier for each sensitive group. This means that both groups should have both the same FNR and FPR.

$$\frac{P(\hat{y} = 1 \mid y = 0, s = 0)}{P(\hat{y} = 0 \mid y = 1, s = 0)} = \frac{P(\hat{y} = 1 \mid y = 0, s = 1)}{P(\hat{y} = 0 \mid y = 1, s = 1)} \tag{2.8}$$

2.2.6 Bias mitigation approaches

Mitigating bias is a crucial phase for developing fair and reliable models. There are three primary approaches to bias mitigation in machine learning: *pre-processing*,

[6] `https://tinyurl.com/eeoc-vii-29cfr-part1607`. Accessed March 30, 2025.

in-processing, and *post-processing*. Each method targets a different stage of the model development process, from handling initial data to adjusting final model outputs.

Pre-processing These methods focus on modifying the training data before it is used to train the model. The goal is to remove biases inherent in the data, either by altering the features or the labels, or by re-sampling the dataset to ensure a more balanced representation of groups across sensitive attributes. Examples of pre-processing techniques are *reweighing*, *sampling*, *disparate impact remover*, and *optimized pre-processing*.

In-processing Also referred to as *in-training*, these techniques involve incorporating bias mitigation procedures directly into the training process of the machine learning model. This can be achieved by adding constraints or regularization terms to the learning algorithm to penalize bias. Alternatively, the model's objective function can be modified to optimize both accuracy and fairness simultaneously. A common method adopted in this category is *adversarial debiasing*.

Post-processing These methods are applied after a model has been trained, focusing on adjusting the model's outputs to ensure fairness across different demographic groups. These techniques involve modifying the decision thresholds or the predictions themselves to correct for biases that the model may exhibit. Post-processing adjustments are particularly useful when it is not feasible to alter the training data or the model itself due to constraints such as time, computational resources, or external regulations. While effective in many cases, it's important to note that post-processing can sometimes lead to a reduction in the overall accuracy of the model, as the modifications are made solely based on the output rather than the underlying data or model structure. Popular post-processing techniques are *reject option classification* and *equalized odds post-processing*.

For the motivations expressed above, and as illustrated in detail in the specific chapters of the presented thesis, in our work, we will focus primarily on the adoption of *pre-processing* and *in-processing* approaches, as well as the introduction of innovative methods of these two types.

2.3 Graph Neural Networks

The advancements in neural network development have resulted in increased research on data mining and pattern recognition. In the past, many machine learning applications, such as object detection [284], machine translation [363], and speech recognition [150], required handcrafted and time-consuming feature engineering to extract useful features. Significant progress in these tasks has been made thanks to the implementation of end-to-end deep learning frameworks such as convolutional neural networks (CNNs) [203], recurrent neural networks (RNNs) [151], and autoencoders [337]. The rapid improvement of computational resources like GPUs, the availability of large training datasets, and deep learning's ability to extract hidden patterns from different types of data have all

contributed to its success in different fields.

While deep learning architectures have proven to be successful in revealing hidden patterns within Euclidean data (e.g., images, text, and videos), a growing range of applications produce data from non-Euclidean domains and present them as complex, interdependent graphs [364]. As a response to this challenge, **Graph Neural Networks** (GNNs) have emerged as a powerful method for effectively modeling and interpreting these intricate, graph-structured datasets. Despite being frequently overlooked, graphs' ubiquity and importance are crucial to understanding the success of GNNs, one of the major progresses in DL over the last decade.

Why graphs? Graphs are a prevalent way of representing data we obtain from nature, as most patterns that we observe, whether in natural or artificial systems, can be described using interconnected structures of nodes and edges. From the molecular level, where atoms are connected by chemical bonds, to the brain's connectomic structure, where neurons are connected by synapses; from transportation networks, which are composed of intersections connected by roads, to social networks, which are made up of users linked by friendship, graphs are a universal and optimal tool for describing living organisms and human-made constructs. Furthermore, from neuroscience research [114], we know that it is probable that the cognitive processes that drive our decision-making and reasoning are structured as graphs. This means that rather than imagining all the information available, we only visualize selected concepts and their relationships to represent the real system. Given this interpretation of cognition, it is highly improbable that we will be able to create a generally intelligent system without a component that relies on graph representation learning. It is worth noting that this finding does not conflict with the fact that many recent ML systems are based on the Transformer architecture [331]. In fact, as revealed in recent articles, Transformers can be viewed as a specific instance of GNNs [332].

In this section, we will explore the development of GNN architectures over time. Initially, we will present a historical perspective on their evolution (Section 2.3.1). Subsequently, inspired by several papers [332, 364, 388, 392] in literature, we will delve into technical descriptions of these models. We will begin with basic definitions of graph data (Section 2.3.2) and progress to the fundamental properties of GNNs (Section 2.3.3), the GNN framework (Section 2.3.4), popular GNN types (Section 2.3.5), and primary tasks for which they have been successfully utilized (Section 2.3.6).

2.3.1 Historical overview

The history of GNNs, whose milestones are depicted in Figure 2.5, traces its roots back to the late 1990s when researchers first began exploring the potential of applying neural network methodologies to graph-structured data. Initial efforts focused on *Recursive Neural Networks*, pioneered by Sperduti and Starita [312], and Frasconi et al. [116], which targeted directed acyclic graphs and aimed to model hierarchical structures in domains like chemistry and natural language processing. These early contributions laid

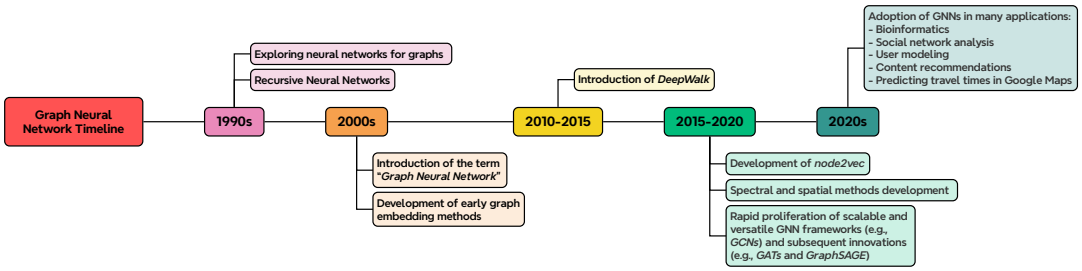


Figure 2.5. Timeline reporting the major events of the GNN history.

a foundational framework that would evolve into more sophisticated approaches.

The term "*Graph Neural Network*" was officially introduced in the early 2000s by Gori et al. [125] and Scarselli et al. [297], marking a significant evolution in the field. This new model extended neural network applications to a broader range of graph types through a general framework that utilized a fixed-point equation for learning on graphs. This period also coincided with the development of early *graph embedding* methods, which became crucial for effectively representing graph data.

A key advancement in graph representation learning came with the introduction of *DeepWalk* by Perozzi et al. [253] in 2014, which pioneered the generation of node embeddings using random walks on graphs, akin to learning word embeddings in natural language processing. Following closely in 2016, *node2vec* by Grover and Leskovec [128] built on this approach, offering a more flexible framework that allowed for biased random walks to efficiently explore and learn from diverse neighborhood structures. Both DeepWalk and node2vec demonstrated the effectiveness of embedding techniques in capturing complex relational information in graphs, setting the stage for subsequent innovations.

The development of *spectral* and *spatial* methods further enriched the GNN landscape. Spectral approaches, introduced by Bruna et al. [48], used the eigen-decomposition of the graph Laplacian to define convolutions in the Fourier space, providing a theoretical basis for graph convolutions. Despite their theoretical appeal, the computational demands of spectral methods led to the rise of spatial methods, which aggregated features from neighboring nodes directly. Introduced by researchers like Duvenaud et al. [100] and Niepert et al. [241], these spatial methods offered computational efficiency and adaptability, becoming foundational for many modern GNN architectures.

Since 2017, the GNN field has seen a rapid proliferation of scalable and versatile frameworks that have significantly broadened the practical applications of GNNs. The introduction of *Graph Convolutional Networks* (GCNs) by Kipf and Welling [180] simplified graph convolutions and became a standard in the field. Subsequent innovations such as *Graph Attention Networks* (GATs), proposed by Veličković et al. [333], and inductive learning models like GraphSAGE [138] have further pushed the boundaries by incorporating sophisticated representation learning techniques to enhance model performance and adaptability.

In the last few years, GNNs have been successfully adopted by scientific and industrial

entities in a wide range of applications, including bioinformatics [115], where they have been used to discover new antibiotics [315]. They are also used in social network analysis [164, 387], user modeling [69, 71, 140, 372], and content recommendations [226, 378]. GNNs can predict travel times in Google Maps [93] and have been instrumental in creating the latest version of machine learning hardware, the TPUv5 [237]. Moreover, systems based on GNN have assisted mathematicians in revealing the latent structure of mathematical objects [85], resulting in the formulation of novel conjectures at the forefront of representation theory [42]. The ongoing advancements in dynamic and heterogeneous graph modeling also showcase the versatility and expanding scope of GNNs, promising continued growth and innovation in handling intricate, relational data structures.

2.3.2 Basic definitions of graph data

Before diving into the core principles of GNNs, it is essential to provide precise descriptions of graph data.

Graph A *graph* is usually defined as $G = (V, E)$ and consists of a set of nodes V and a set of edges E, where each edge $e = (u, v) \in E$ exists if there is a connection between the two nodes $u \in V$ and $v \in V$.

Adjacency matrix A practical and efficient form to represent a graph is through an *adjacency matrix* $\mathbf{A} \in \mathbb{R}^{|V|\times|V|}$. The adjacency matrix $\mathbf{A}$ is a key representation in graph theory, playing a crucial role in defining the connectivity and structure of the graph in computational terms. The graph nodes are arranged in a specific order that corresponds to each row and column in the matrix. This order allows for the population of the adjacency matrix $\mathbf{A}$ with entries $a_{uv} \in \mathbf{A}$ that indicate the presence of edges within the graph, as follows:

$$a_{uv} = \begin{cases} 1 & (u, v) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

Neighborhood Moreover, the edges of the graph provide a *locality* constraint in these functions. Similar to a CNN, which operates over a small neighborhood of each pixel within an image, a GNN can process a node's neighborhood within a graph. A common definition of a neighborhood $\mathcal{N}_u$ is the following:

$$\mathcal{N}_u = \{v \in V \mid (u, v) \in E \vee (v, u) \in E\} = \{v \in V \mid a_{uv} = 1\} \tag{2.10}$$

Contextually, the *degree* of a node $u \in V$ is defined as:

$$|\mathcal{N}_u| = deg(u) = \sum_{v \in \mathcal{N}_u} a_{uv} \tag{2.11}$$

Directed/undirected graphs In an *undirected graph*, edges have no direction, indicating a bidirectional relationship between nodes. It means that if node u is connected

to node v, node v is also connected to node u. In the adjacency matrix, $a_{uv} = a_{vu} = 1$ if $(u, v) \in E$. Therefore, $\mathbf{A}$ is symmetric. In a *directed graph*, each edge has a direction, indicating a one-way relationship from one node to another. The existence of an edge from node u to node v does not imply the presence of the opposite edge, i.e., from node v to node u. Graphs may also incorporate *weighted* edges, in which the entries in the adjacency matrix can take on arbitrary real values, as opposed to being limited to $\{0, 1\}$.

Heterogeneous graphs These graphs are characterized by the presence of nodes with distinct *types*, allowing for the categorization of nodes into separate and non-overlapping sets $V = V_1 \cup V_2 \cup \cdots \cup V_m$, where m is the number of node types and $V_i \cap V_j = \varnothing, \forall i \neq j$. In heterogeneous graphs, edges typically adhere to constraints based on the types of nodes they connect, with a common restriction being that certain edges only link nodes of specific types. *Multipartite graphs* represent a specific category of heterogeneous graphs, wherein the connectivity between nodes is restricted to pairs possessing distinct types, i.e., $(u, v) \in E$, where $u \in V_i, v \in V_j \wedge i \neq j$.

Multi-relational graphs In scenarios where graphs exhibit various types of edges $\tau \in \mathcal{T}$, the edge notation can be extended to cover the specific edge or relation types, being defined as tuples $e = (u, \tau, v) \in E$. This approach allows the creation of an adjacency matrix $\mathbf{A}_\tau$ for each edge type separately. The graphs described are classified as *multi-relational*, and can be represented and summarized by an *adjacency tensor* $\mathcal{A} \in \mathbb{R}^{|V| \times |\mathcal{R}| \times |V|}$, where $\mathcal{R}$ indicates the relation sets. Multi-relational graphs are commonly known as *knowledge graphs* due to the ability to interpret the edge tuples as denoting specific factual relationships between the nodes.

Feature information A graph often has *attribute* or *feature* data associated with it. Typically, these are attributes related to nodes and are described by a *node feature matrix* of real values, denoted as $\mathbf{X} \in \mathbb{R}^{|V| \times d}$, in which each row represents a *feature vector* $\mathbf{x}_u \in \mathbb{R}^d, \forall u \in V$. It is assumed that the node ordering is consistent with the adjacency matrix. Different types of nodes have distinct attributes when dealing with heterogeneous graphs. In some cases, graphs with real-valued edge features are considered, and occasionally, real-valued features are associated with entire graphs.

2.3.3 Fundamental properties

A key challenge in working with graphs is the non-fixed ordering of nodes, which differs from the structured data formats used in other types of neural networks, such as images in CNNs or sequences in RNNs. To address this, GNNs are designed to retain two essential properties: *permutation invariance* and *permutation equivariance*.

Permutation invariance This property is essential when the GNN's output should be the same regardless of the order of nodes in the input graph. For example, in graph classification tasks where the entire graph is assigned a label, the predicted label should not change if the nodes of the graph are reordered. This ensures that the model's output

for graph-level predictions is consistent across different permutations of the graph's nodes.

Permutation equivariance Unlike invariance, equivariance refers to situations where the output needs to change in a predictable way according to changes in the input. For GNNs, this means that if the nodes in the input graph are reordered, then the node-level outputs (e.g., features or embeddings) are rearranged correspondingly. This property is crucial for tasks like node classification, where each node's output, such as a class label, is directly tied to that node and must follow the node if the input order changes.

From a mathematical perspective, given a *permutation matrix* $\mathbf{P}$, it is desirable for any function f that operates on an adjacency matrix $\mathbf{A}$ to adhere to one of the two following constraints:

$$f(\mathbf{P}\mathbf{A}\mathbf{P}^\top) = f(\mathbf{A}) \quad \text{(Permutation invariance)} \tag{2.12}$$

$$f(\mathbf{P}\mathbf{A}\mathbf{P}^\top) = \mathbf{P}f(\mathbf{A}) \quad \text{(Permutation equivariance)} \tag{2.13}$$

2.3.4 Graph Neural Network framework

GNN models are centered around the concept of *message passing*, a process that is fundamental to every GNN variant. This mechanism allows nodes to exchange information with their neighbors, effectively aggregating local and global structural insights to update node states and eventually produce outputs for various tasks.

Initially, each node is associated with a *feature vector* $\mathbf{x}_u \in \mathbb{R}^d$. The objective is to transform these features into comprehensive embeddings $\mathbf{z}_u, \forall u \in V$, that encapsulate both the intrinsic properties of the nodes and their contextual relationships within the graph. Moreover, embeddings for both subgraphs and complete graphs can be generated. The whole procedure is composed of the following steps: *message aggregation*, *features update*, and *global pooling*.

Message aggregation Each node u exchanges messages with its neighbors over multiple iterations or layers. The process for one iteration k is described by:

$$\mathbf{m}_u^{(k)} = \text{AGGREGATE}^{(k)}(\{\mathbf{h}_v^{(k-1)} : v \in \mathcal{N}_u\}) \tag{2.14}$$

where $\mathbf{h}_v^{(k-1)}$ are the features of neighboring nodes from the previous iteration and $\mathcal{N}_u$ denotes the set of neighbors of node u. The AGGREGATE function might involve operations like summing, averaging, or even more complex neural network-based mechanisms.

Features update The node features are updated by combining the aggregated messages with the node's previous state to generate a *hidden embedding* $\mathbf{h}_u$:

$$\mathbf{h}_u^{(k)} = \text{UPDATE}^{(k)}(\mathbf{h}_u^{(k-1)}, \mathbf{m}_u^{(k)}) \tag{2.15}$$

The UPDATE function is typically expressed as a neural network:

$$\mathbf{h}_u^{(k)} = \sigma(\mathbf{W}^{(k)}\mathbf{m}_u^{(k)} + \mathbf{b}^{(k)}) \tag{2.16}$$

where σ is a non-linear activation function, such as a ReLU, $\mathbf{W}^{(k)}$ is the weight matrix, and $\mathbf{b}^{(k)}$ is the bias vector, specific to layer k.

After K iterations, the embeddings of each node can be defined by the outcome of the final layer:

$$\mathbf{z}_u = \mathbf{h}_u^{(K)}, \forall u \in V \tag{2.17}$$

Global pooling To derive a graph-level output from node-level features, GNNs operate a global pooling function, also called *readout*, that aggregates features across all nodes:

$$\mathbf{z}_G = \text{READOUT}(\mathbf{z}_u) = \text{READOUT}(\{\mathbf{h}_u^{(K)}, \forall u \in V\}) \tag{2.18}$$

This step might employ various pooling strategies like global sum, average, or even learnable methods to synthesize the information from all nodes.

2.3.5 Popular types

Different types of GNN structures have been identified in the literature, depending on the particular methods and procedures utilized for the message-passing algorithm. This section provides a formal representation of the most widely utilized models, specifically the *Graph Convolutional Network* (GCN) and the *Graph Attention Network* (GAT).

Graph Convolutional Network GCNs simplify graph convolutions by approximating spectral graph convolution. The update rule for a GCN layer can be expressed as:

$$\mathbf{h}_u^{(k)} = \sigma\left(\sum_{v \in \mathcal{N}_u \cup \{u\}} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_v|}} \mathbf{W}^{(k)}\mathbf{h}_v^{(k-1)}\right) \tag{2.19}$$

This equation combines the features of node u and its neighbors, normalized by the nodes' degrees, enhancing the stability of the learning process.

Graph Attention Network GATs introduce an attention mechanism to weigh the importance of neighbors' features dynamically. The related update function is:

$$\mathbf{h}_u^{(k)} = \sigma\left(\sum_{v \in \mathcal{N}_u} \alpha_{uv}^{(k)} \mathbf{W}^{(k)}\mathbf{h}_v^{(k-1)}\right) \tag{2.20}$$

where α_{uv} are coefficients computed by the attention mechanism, determining the influence of each neighbor's features.

2.3.6 Main tasks

There are three main purposes for which GNNs are employed:

- **Node classification**: The goal is to predict the label of a node based on its features and its neighborhood information. This task is common in social network analysis, where one might predict the role or group membership of individuals.
- **Link prediction**: This involves predicting the likelihood of a relationship between two nodes, which is useful for recommending friends in social networks or predicting interactions between proteins in biological networks.
- **Graph classification**: In this task, the entire graph (or subgraphs) is classified into different categories. This is particularly useful in chemistry for predicting the properties of molecules or in document classification, where entire graphs represent documents or sentences.

2.4 User Modeling

In a scenario where AI systems, especially IR and recommender platforms, produce a large quantity of personal data on a daily basis, it becomes important to identify individuals' interests, characteristics, and behaviors. This requirement is met by utilizing **user modeling** (or **user profiling**) techniques [103]. The main goal of these methods is to create a reliable representation of the user, commonly referred to as a **user model** (or a **user profile**), by using the data that have been generated [173]. User modeling and profiling enable organizations to understand user behavior, preferences, and interests through data analysis. This information allows organizations to deliver personalized experiences, leading to increased user satisfaction and engagement.

In this section, our aim is to thoroughly examine the various aspects of user modeling. We will begin with a historical overview tracing the evolution from early *stereotype user modeling* initiated in 1978 to today's sophisticated deep learning-based approaches (Section 2.4.1). We will proceed by presenting new encyclopedic definitions that are relevant to the domain of user modeling (Section 2.4.2), which have been refined through an extensive analysis of the field, supported by a comprehensive survey [267]. We will then examine the significant paradigm shifts that have occurred within the last decade alongside the emerging trends that have shaped the current research landscape (Section 2.4.3). Finally, we propose a formal taxonomy of the user modeling domain, which has been developed based on insights drawn from our comprehensive literature review previously mentioned (Section 2.4.4).

2.4.1 Historical overview

The fields of user modeling and user profiling have seen significant progress throughout the history of scientific literature on personalization. Figure 2.6 depicts the major milestones in these areas.

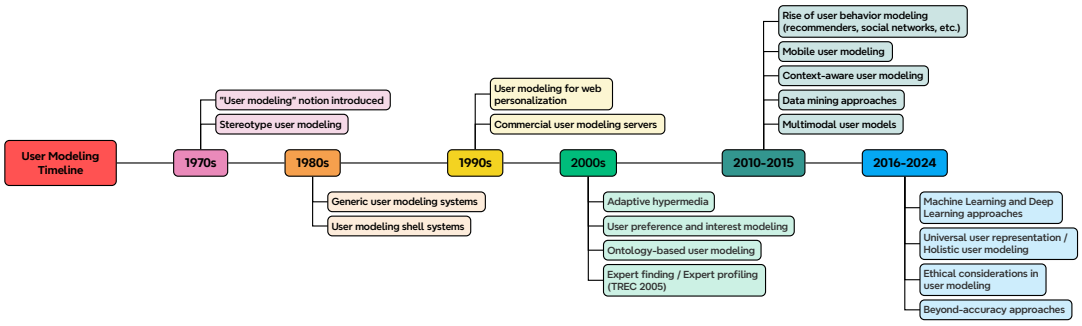


Figure 2.6. Timeline reporting the major events of the user modeling history.

The concepts of "*user model*" and "*user modeling*" were first introduced by Allen, Cohen, Perrault, and Rich (i.e., [76, 254, 288]), in the late 1970s. Their pioneering work laid the foundation for subsequent research in this field and led to the development of various application systems that collected user information and had adaptive capabilities (e.g., [16, 310, 340]). The first attempt to distinguish a user from other users was through *stereotype user modeling* [288, 299], and it inspired several future contributions (e.g., [23, 192, 396]).

From the end of the 1980s, it became clear that the user modeling component needed to be reusable to create user-adaptive systems. To achieve this, the first step taken was to develop *generic user modeling systems* [46, 112, 175, 182, 185, 246, 249], also defined as *user modeling shell systems* [181]. Toward the end of the 1990s, web personalization gained significance in electronic commerce [15, 152]. User modeling was recognized as crucial [113], and many systems were developed and published during this period (e.g., [51, 56, 176, 183, 186]). Research on user profiling and user modeling made significant progress in the 2000s, with a focus on enhancing personalization and adaptability in various systems.

A research area called *adaptive hypermedia* [50, 52] emerged at the intersection of hypermedia systems [184] and adaptive user interfaces [199]. Unlike regular hypermedia, adaptive hypermedia tailored the set of hyperlinks offered to users based on a model of their goals, preferences, and knowledge [86]. During the same period, studies also focused on creating user profiles that accurately captured interests based on observations of user behavior on the web (e.g., [62, 118, 121, 207, 257]). With the advent of the *semantic web*, researchers investigated representing and modeling user preferences through *ontologies* [88, 233, 308, 311], which were used to semantically organize and connect user profiles, thereby improving the understanding of user preferences and relationships.

The introduction of the *expert finding* and *expert profiling* tasks [26, 27, 28] in the Enterprise Track at TREC 2005 [80] marked a significant milestone in user modeling research. This event brought the field a lot of attention [103, 210, 251] and is considered a turning point in this domain. During the 2010s, there was a significant evolution toward more advanced user profiling methods, with a greater emphasis on personalization in various digital services, particularly in RSs [3, 198]. Researchers developed sophisti-

cated algorithms to analyze user behavior and preferences, resulting in improved content recommendations [187, 228]. Innovative methods included personality-based user adaptation, in which automated methods were created to recognize personality traits in user behaviors [32, 126] and conversations [225, 353].

Another significant development was the use of *semantic user modeling* techniques, which involved creating computational models to understand user preferences, behaviors, and intentions based on semantic information derived from various data sources. Studies increasingly focused on *context-aware user modeling* to comprehend how user preferences and behaviors were influenced by different contexts [8, 309, 334]. Researchers also explored ways to incorporate social network data into user models, using social network analysis to understand the influence of social connections [329, 393]. The integration of such data and user-generated content helped create more accurate and context-aware user profiles [256].

During the same period, the rise of *big data* led to the exploration of advanced *data mining* methods for user modeling [291, 328]. The application of ML algorithms, such as clustering and classification, on large datasets helped reveal valuable patterns and insights into user behavior [191, 385]. Following this, nearly all significant research in user modeling began to concentrate on implementing *deep neural networks* to model complex user behaviors, aiming to provide more precise predictions and personalized experiences (e.g., [19, 68, 69, 323, 372]).

In recent years, the widespread collection of user data has led to increasing awareness of *privacy* concerns and the development of privacy-preserving techniques [11, 163] such as *federated user modeling* [216, 362]. Moreover, there has been a surge of scientific contributions focused on XAI to enhance the interpretability of ML models, including its application in user modeling [29, 160].

Ethical concerns have gained significant prominence in the past few years in many areas, including user modeling research. Nowadays, there is a commitment to foster transparency, accountability, and fairness in algorithmic decision-making by addressing these challenges [82, 261, 389]. It is essential to ensure that user models maintain representativeness and fairness across different user groups [268]. The evolving landscape includes a growing emphasis on fostering effective human-AI collaboration to enhance the ethical and inclusive dimensions of user modeling [397].

2.4.2 Novel definitions

In the preceding section, we highlighted the abundance of research surrounding user modeling and user profiling, as evidenced by the numerous studies documented in the academic domain. Over the tons of contributions that have been published in these areas, there persists a notable ambiguity and inconsistency in the terminology employed across these fields. To address this issue, our research started with a rigorous and systematic review of the literature, marking the first endeavor of its kind to meticulously deconstruct and analyze the foundational terminology used within the user modeling research area.

Our comprehensive analysis was directed toward clarifying and refining the definitions associated with key concepts, including *user profile* [13, 18, 57, 121, 173, 200, 247, 319],

user model [49, 118, 256, 293, 324], *user profiling* [69, 87, 103, 173, 338, 348, 393], *user modeling* [4, 108, 179, 209, 285], and *user profile modeling* [18]. This initiative aimed to eliminate any prevalent misunderstandings or misinterpretations in their usage, ensuring a precise and uniform application of these terms across the academic community. Our detailed examination revealed that, particularly in publications from the last decade, the terms "user model" and "user profile" (along with "user modeling" and "user profiling") often bear overlapping descriptions. This insight has led us to conclude that these terms, while historically distinct, have evolved to become largely synonymous and can be used interchangeably.

In light of these findings, this dissertation will henceforth employ a single term for these concepts, chosen for its clarity and relevance, without alternating between terminologies. Drawing from our extensive investigation, which can be consulted in depth in our survey [267]. Based on the unique insights gained from our in-depth survey, we introduce two novel, encyclopedic definitions. These definitions are designed not only to standardize usage but also to enrich the domain of user modeling and profiling by providing a clear and authoritative reference for future research:

> *A **user model** (or **user profile**) is a structured representation of an individual user's preferences, needs, behaviors, and demographic details to personalize system interactions. It is derived from direct user feedback or inferred through machine learning and data mining techniques. It supports the predictions of future user intentions and the refinement of systems response to enhance user satisfaction. User models are often instrumental in optimizing the relevance and efficiency of adaptive systems, ensuring that user interactions are aligned with individual needs and preferences.*

> ***User modeling** (or **user profiling**) is the process of acquiring, extracting, and representing user features and personal characteristics to build accurate user models (or user profiles). It encompasses inferring personality traits and behaviors from user-generated data. This dynamic practice includes automatically converting user information into interpretable formats, capturing latent interests, and learning conceptual user representations. Essentially, user modeling constitutes the methodology for building and modifying user models, determining "what" to represent and "how" to effectively represent this information for adaptive and personalized systems.*

2.4.3 The evolving research landscape

In our comprehensive survey [267], along with the thorough taxonomy analysis detailed in the previous section, we have identified several significant paradigm shifts and emerging trends that have reshaped the landscape of user modeling research over the last decade. These developments not only mark a transformative phase in the field but also form the basis of the innovative contributions presented in this manuscript. The discussion that follows explores these changes, demonstrating how user modeling has been

redefined by technological advancements, shifts in user behavior, and new methodological approaches.

Implicit and explicit user modeling In the past, profiling techniques primarily concentrated on analyzing static characteristics. The traditional *explicit user modeling* (also known as *static* or *factual profiling*) required users to directly disclose their information by means of questionnaires, online forms, or explicit ratings and preferences [49, 258, 280, 395]. These approaches became problematic, especially due to privacy concerns, and the process of filling out forms and questionnaires was considered tedious. As a result, the accuracy of this type of modeling decreased over time [173, 174] in favor of the implicit methods, which began to be increasingly adopted. Also referred to as *behavioral* or *adaptive profiling* [103, 173], *implicit user modeling* involves gathering and analyzing dynamic user data in a non-intrusive manner. This may include observing user behaviors and interactions without requiring direct input from the user [131, 140, 224, 279, 373]. Both approaches, explicit and implicit user profiling, have been used together for many years, but modern systems have now shifted their focus to place greater importance on the latter. The practice of utilizing static data for user profiling has evolved into collecting direct information obtained from publicly available data that users have already shared, such as signing up for social media platforms [306, 307]. To refer to this new type of profiling, we coin the term "*pseudo-explicit user profiling*".

User preferences and interests The study of these aspects has evolved alongside the advancements in implicit and explicit modeling. With the rise of digital platforms like *e-commerce* services and *recommender systems*, there has been an abundance of user-generated data, such as social interactions and opinionated text content. This has led to a growing focus on capturing user interests and preferences hidden in their historical behaviors [24, 58, 81, 107, 127, 145, 218, 188, 219, 222, 300, 347]. In this scenario, specific research on *short-* and *long-term preference modeling* started to arise [19, 109, 134, 159, 215, 320, 379].

User behavior modeling The investigation of user behaviors has advanced significantly to incorporate a range of refined modeling techniques and innovative concepts that offer a more profound comprehension of the users in numerous contexts: *Micro and macro behavioral modeling* [132, 355] respectively refer to immediate and large-scale actions taken by the user that reflect their short-term and long-term preferences; *Multi-behavior modeling* [71, 73, 168, 365, 370] integrates diverse forms of user interactions with items rather than depending on a single type; *Sequential behavior modeling* [36, 60, 65, 380] involves considering the temporal sequences of user actions that impact the interests and preferences; *Hierarchical user modeling* [132, 354, 355, 371] is a technique employed in personalized recommender systems, particularly in e-commerce, to model users' real-time interests at varying levels of granularity; *Mobile user modeling* [348, 349, 346] identifies users' interests and behavioral patterns through their activities on mobile devices.

User representation Current research on user modeling tends to focus on specific aspects instead of generalized approaches [325]. Researchers and practitioners have been able to gain a more comprehensive understanding of users in digital environments and effectively meet their needs through the development of concepts such as *universal user representation* and *holistic user modeling*. *Universal user representation* is a concept that creates a generalized profile of a user by encapsulating a broad spectrum of user behaviors and preferences applicable across various domains and applications [130, 179, 381]. *Holistic user modeling* integrates diverse personal data sources to construct a comprehensive representation of the user, providing a complete picture that can be used to personalize experiences across different platforms [123, 239, 240].

Evaluation in user modeling Historically, approaches to assess user modeling methods included a *layered* strategy aimed to separate the evaluation of different aspects to help identify problems in the adaptation process [250]. There are currently two primary methods for assessing the effectiveness of independent user modeling approaches. The first approach involves evaluating the efficiency of the proposed model or method by assessing its ability to accurately predict a user's personal characteristics through a *classification task* [68, 79, 82, 372]. On the other hand, the second approach focuses on generating *simulated data* to minimize the amount of user data gathered while still maintaining the accuracy of profiling and safeguarding the privacy and confidentiality of users' personal data [30].

Graph data structures Similar to other fields, there has been a significant emphasis in user modeling on exploring and implementing graph structures, including knowledge graphs. *Graph structures* are powerful representations of data that capture relationships among data objects and can be efficiently used to represent and analyze user behavior, preferences, or interactions, making them commonplace in real-world applications [70, 214, 343, 348, 374]. *Knowledge graphs*, a specific type of graph structure, have gained attention from academia and industry for their ability to effectively represent complex information. They are employed to accumulate and disseminate knowledge of the real world, which is beneficial for analyzing critical information from people's activities and posts on social media [20, 344, 349, 370].

Deep learning The rise of models based on deep neural networks has played a crucial role in the progress of the user modeling research field. A variety of DL techniques and architectures have been utilized in the field. In particular, *differentiable user models* [162], *attention mechanism* [75, 110, 278, 352], *graph neural networks* [69, 71, 82, 140, 359, 372], *convolutional neural networks* [110, 278, 343, 348], *autoencoders* [5, 109, 216, 348], *recurrent neural networks* [75, 119, 132, 379], *long-short term memory networks* [110, 223, 243, 292, 398], and *transformers* [129, 189, 361, 390, 394].

Beyond-accuracy perspectives The adoption of beyond-accuracy approaches in various domains represents a significant shift similar to the emergence of deep learning

architectures. These methods prioritize values such as *privacy*, *fairness*, and *explainability*. In user modeling, such approaches can make accurate predictions while protecting user privacy [75, 216, 221, 362], addressing biases [1, 82, 268, 261, 302, 391], and promoting transparency [29, 89, 96, 144, 236, 366].

2.4.4 Taxonomy

The comprehensive literature review, on which our survey [267] is based, allowed us to present a detailed overview of the research field of user modeling, a vast and continuously evolving area of study. We traced the significant milestones of the scientific literature, from the introduction of *stereotype user modeling* in 1979 to recent contributions on *beyond-accuracy perspectives*. Based on our analysis, we developed a formal taxonomy that encompasses all the presently active topics in the research area, including emerging trends in the last few years. The proposed taxonomy of the user modeling research field is shown in Figure 2.7, while Figure 2.8 reveals the details of the *Modeling techniques* tree, separately displayed to enhance the visualization perspective.

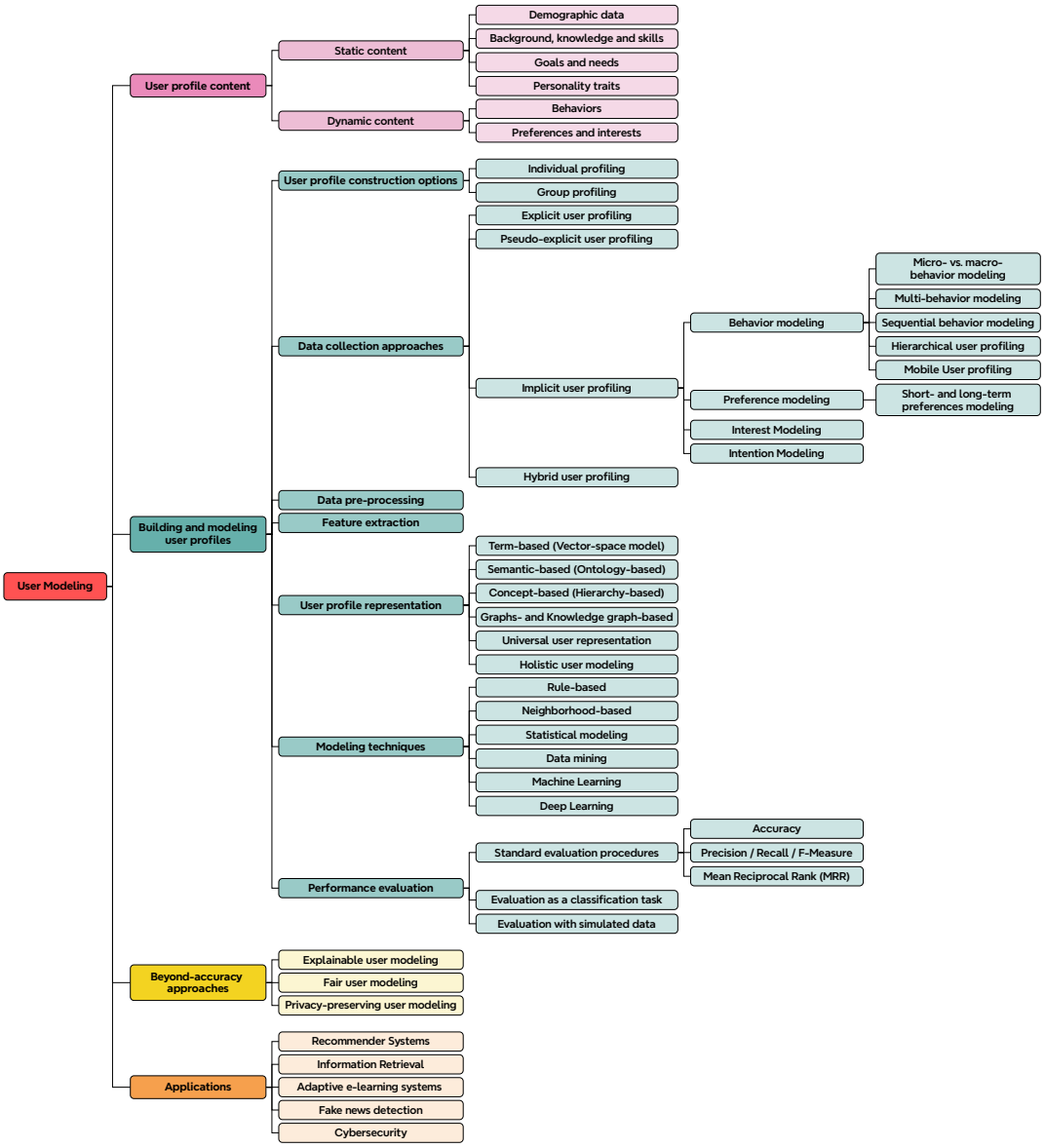


Figure 2.7. Novel taxonomy of the **user modeling** research field, proposed in our comprehensive survey [267]. The *Modeling techniques* tree is displayed in detail in Figure 2.8.

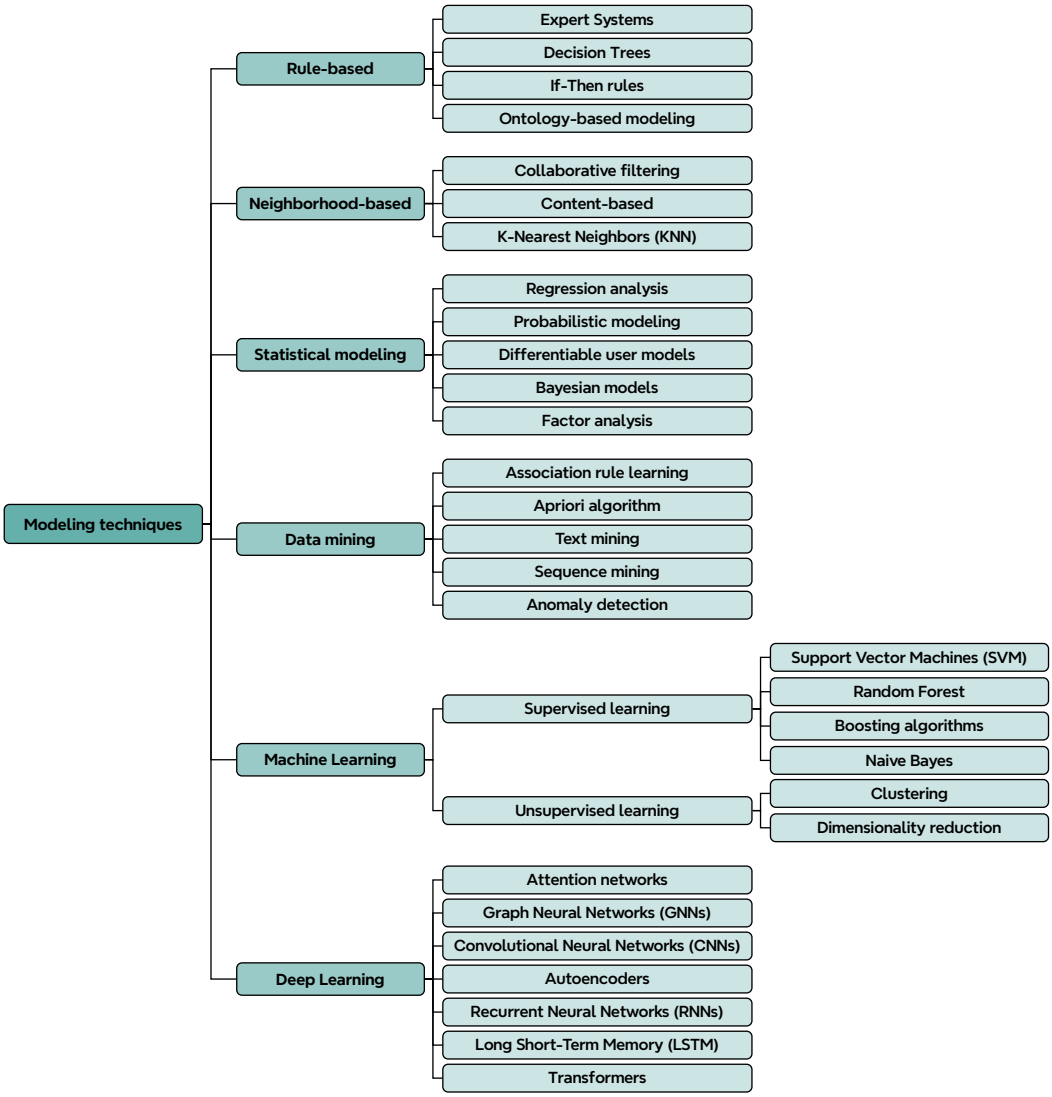


Figure 2.8. Detailed representation of the *Modeling techniques* branch in our novel taxonomy (Figure 2.7).

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Chapter 3

Fairness Analysis in Generic Machine Learning Models

He who would learn to fly one day must first learn to stand and walk and run and climb and dance; one cannot fly into flying.

Friedrich Nietzsche

The work described in this chapter, which represents the initial study of the presented doctoral research, focuses on demonstrating how the application of explainability and fairness techniques can enhance a domain expert's trust and reliance on Artificial Intelligence (AI) systems, in particular, machine learning (ML) models.

To achieve this goal, we propose a Responsible AI framework composed of four main functions: a *dataset & ML model handler*, a *standardized explainability tool*, a *fairness tool*, and a *feedback loop*. The first function allows users to upload a dataset and preprocess it, train different ML models to identify the most effective one for the provided dataset and monitor its performance. The *standardized explainability tool* offers methods for obtaining explanations for each prediction. Our objective is to create a tool that supports the development of a Responsible AI system, regardless of the specific ML model it is based on. The *fairness tool* empowers users to identify biases within the model's behavior and generate an unbiased version of the original model through the adoption of a novel pre-processing bias mitigation algorithm. The *feedback loop* enables a domain expert to evaluate the model's results alongside the related explanations in order to establish a new ground truth for training a more effective ML model.

We have utilized the presented system in the context of the loan approval process, creating a proprietary framework with an intuitive interface and showcasing its effectiveness through field tests and subsequent user studies. We conducted an experimental session to select the best explainability algorithm for the developed framework, following the state-of-the-art *Explanation Goodness Scale*. Additionally, we introduced a new *Trust & Reliance Scale* to assess the system's explainability, and an *A/B test* was conducted to evaluate the fairness feature. Finally, we carried out a *usability test* to assess the user

interface (UI). The results were obtained by issuing the mentioned scales, tests, and usability questionnaires to data scientists and researchers for the explainability algorithms and to bank domain experts and loan officers to evaluate the other functionalities.

The upcoming chapter, whose structure is outlined below, will discuss the components of the implemented systems, focusing on the adopted bias detection approaches and the proposed bias mitigation method. This is in line with the principal topic of the dissertation. The techniques and components that will only be briefly described are thoroughly presented in our journal article. [269].

In Section 3.1, we delve into the underlying reasons and driving factors behind the research, highlighting its significance and relevance. Section 3.2 outlines the research approach, techniques, and procedures used to conduct the study, ensuring a rigorous and systematic investigation. The system design and implementation are presented in Section 3.3. This section provides a detailed overview of the architectural framework of the proposed Responsible AI system. Following this, Section 3.4 presents the practical application of the proposed system in the context of loan approval process, demonstrating its functionality and effectiveness in a real-world scenario. The evaluation in Section 3.5 critically assesses the performance and outcomes of the system through an extensive user study. Finally, Section 3.6 summarizes the key findings and insights derived from the chapter, setting the stage for subsequent discussion.

3.1 Motivation

The financial and banking sectors have always relied on the capacity to forecast the likelihood of specific events happening in the future [43, 148]. Assessing the potential risk of granting a loan at a bank demands extensive knowledge and significant experience from loan and credit officers. This involves analyzing customer information, including personal data, financial status, and credit history, as well as evaluating the specifics of the loan request. As envisioned in recent years by several industrial investigations (e.g., [84]), currently, it is difficult to find a branch or department within a financial institution that does not require predictive analytics. The volume of data needed for this type of analysis in lending, which includes historical information on approved loans, makes it one of the most compelling areas for the application of AI in the banking industry.

In this context, ML models have been employed, for instance, in the prediction of stock prices [316] or, as in the case study described in this chapter, in determining whether to approve a bank customer's loan [301]. Due to the specific application area, the risk linked with the calculated prediction may differ significantly. According to multiple studies [122, 156], even though AI systems are matching or outperforming human performance in numerous fields, their adoption is still met with suspicion, and human expertise is often deemed irreplaceable [166]. Understanding the motivations behind a specific outcome can be more important than the outcome itself in certain situations. It is essential to comprehend why a prediction was made in order to establish confidence in a model's decisions. Trust plays a crucial role in the adoption of machine learning techniques in high-risk applications, leading to the emergence of the fields of *Human-*

Centered AI (HCAI) and *Responsible AI*, as deeply depicted in Section 2.1.

It is becoming increasingly important to consider whether the implementation of AI systems and the decisions made by them should be guided by a set of *ethical principles* to ensure transparency, social equity, and sustainability. In the specific case of automatic predictions in loan approval processes, it is essential to take into consideration several key aspects of European law, such as the *EU AI Act* referenced in Section 2.1.1, which states that individuals evaluated by an automated decision-making system have the right not to be subjected to a solely automated decision, to receive an explanation of the decision, and not to be discriminated against. As highlighted in Perez's report "*Fairness in Machine Learning*"[1], ML practitioners should create models that inherently address potential discrimination and are understandable to users, necessitating high transparency and reproducibility throughout the entire system workflow.

The debate about the necessity of explainability in the AI community is highly contentious. Hinton, for example, considers the constant pursuit of explaining how an AI system operates as a "complete disaster."[2] Our perspective opposes this view, and we consider explainability to be important for two primary reasons: *trust*, as people cannot simply rely on statistical information about model performance to believe that a decision is correct, and *ethics*, as we need to demonstrate that a developed system does not result in any form of discrimination. Therefore, a successful Responsible AI system must be connected to the social sciences [235].

Connected to the notions of explainability and ethics is the concept of *algorithmic fairness* (meticulously illustrated in Section 2.2 and hereinafter in this chapter simply referred to as *fairness*). Understanding how a prediction was generated can reveal discriminatory behavior in machine learning models. This makes it possible to detect and address biases originating from the data provided by humans, which forms the basis of these models. Consequently, the predictions made by these systems may exhibit a preference for the majority group over certain minorities.

The development of explainable UIs is also an essential aspect of creating a reliable and valuable AI system. Currently, this area remains a weak point in Explainable AI (XAI) research, and its assessment is also a critical subject [2]. While it may seem predictable that users interacting with systems providing explanations (as opposed to those without explanations) would be more satisfied, a concrete evaluation is always necessary, particularly in domains where dealing with explainable UIs is not a common practice. This assertion is also backed by evidence in the literature: Millecamp et al. [234] demonstrated that in specific contexts and for particular users, explanations could lead to a lack of confidence in the system; Wang et al. [345] showed that users might prefer a biased model over an unbiased one if proper result explanations are missing.

[1] `https://2021.ai/fairness-in-machine-learning/`. Accessed March 30, 2025.
[2] `https://tinyurl.com/hinton-xai-interview/`. Accessed March 30, 2025.

3.2 Methodology

While automated systems could improve loan approval processes, their application in this field has been limited due to several reasons: (1) loan approval processes are high-risk activities, requiring officers to understand the reasoning behind each ML model prediction. Merely demonstrating that a model performs well as a black box is not sufficient. With skeptical users, the ability to explain *how* it works, *which* data is important, and *when* is crucial; (2) model decisions significantly impact the future of loan applicants, and they must receive explanations for why their application has been rejected; (3) decisions must be unbiased to ensure fair treatment of individuals from different origins, cultures, and backgrounds.

The system introduced in this chapter aims to address the aforementioned challenges by offering a unified solution to develop a comprehensive *trustworthy intelligent system* that leverages the principles of *explainability* and *fairness*. This section provides a thorough examination of the latter concept, emphasizing its significance in this research study. As previously stated, a detailed description of the explainability component is featured in our journal article.

The adoption of fairness criteria and metrics in ML models has garnered significant attention lately due to heightened awareness of the potential risks posed by biased AI systems toward certain groups. In recent years, there has been a proliferation of academic research pertaining to the emerging field of Fair ML (e.g., [33, 63, 74, 77, 155, 231, 232]). The rise in popularity has not only increased the number of scientific publications on fairness but has also prompted the development of several tools designed to monitor a model's behavior and detect any unfair treatment. An example is the *IBM AI Fairness 360* (AIF360) [34], which is among the most significant open-source toolkits for algorithmic fairness. Its aim is to "*inspect, report, and address discrimination and bias in ML models throughout the AI application lifecycle*." It is an adaptable framework that can consolidate most of the metrics and algorithms discussed in this chapter. It also includes a bias explanation feature that provides further insights into the computed metrics. *IBM Watson OpenScale*[3] is another popular tool that offers fairness capabilities by allowing the configuration of a monitor to keep track of the biases present in the model being used. The determination of biases in the model is based on the disparate impact metric. One of OpenScale's primary limitations is that the selection of privileged and unprivileged groups must be done in advance when setting up the fairness monitor. This process can become complex as the number of sensitive attributes increases. Besides, the user may not know which value corresponds to which group.

However, to ensure fairness, it is important that users are informed about any biases and prejudices that may cause AI systems to discriminate against specific individuals or groups. Additionally, AI systems should be designed to be accessible to people of all ages, genders, and capacities. As also discussed in Section 2.2, no standard definitions of "fairness" have been established so far. Yet, in our research domain, it is considered

[3] `https://www.ibm.com/docs/en/cloud-paks/cp-data/4.8.x?topic=services-watson-openscale`. Accessed March 30, 2025.

as the absence of any bias or preference towards an individual or group based on their inherent or acquired characteristics [231]. A recent study on perceptions of fairness in loan allocations [296] found a preference for a specific definition called *calibrated fairness* [217], which aims to select individuals in proportion to their merit. This study showed that when officers have to choose between two loan applicants, they tend to favor splitting the money in proportion to their loan repayment rates rather than an "equal" (i.e., 50/50) split or giving all the money to the candidate with the higher payback rate. This "ratio" decision is permissible under the calibrated fairness definition.

We depicted the potential causes of bias in Section 2.2.3. The observed discrimination in an ML model may result from its training on biased example data. When using historical data to model human behaviors, it is essential to consider that the sample is influenced by the biases introduced by the individuals involved in the decision-making processes. The selection of the appropriate bias mitigation algorithm is particularly influenced by the stage of the ML model pipeline at which the user can intervene (see Section 2.2.6). In general, the earlier the algorithms are implemented, the more adaptable and effective the intervention will be. The selection of the algorithm is also dependent on its own requirements. For example, the *equalized odds post-processing* method, despite being a post-processing technique, needs access to the sensitive feature to calculate the correct label. Some algorithms have constraints in terms of the types of classifiers they can be applied to. Certain algorithms, like *reject option classification*, are deterministic, while others involve a random element, such as the *disparate mistreatment remover*.

In the implementation of the framework described within this chapter, we employed the *disparate impact* metric (Equation (2.3)) as the bias detection technique and developed a pre-processing bias mitigation method based on the *reweighing* algorithm.

3.3 System design and implementation

This section outlines the proprietary framework designed for implementing the aforementioned case study. It then demonstrates the system's main aspects, which involve applying the principles of explainability and fairness to the loan approval process. The framework includes the functionalities depicted in Figure 3.1, which are organized based on their high-level purpose to ensure complete management of the ML model life cycle.

The *dataset & ML model handler* enables users to: import a dataset and save it using a process that includes a consistent preprocessing step, a customized configuration, and a fairness check for initial bias identification; identify the most suitable ML model by simultaneously training multiple models using different algorithms and assessing them using standard metrics; monitor the performance of the models using a range of metrics similar to those used to evaluate a model after training.

The standardized **explainability tool** gives users the capability to receive explanations for every prediction. This enables both loan officers and loan applicants to clearly understand the features that have the most influence on the results, whether positively or negatively.

The **fairness tool** offers a method for conducting a fairness assessment and imple-

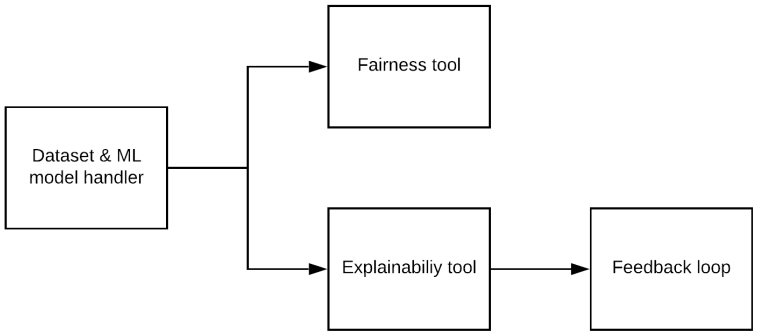


Figure 3.1. High-level components of the Responsible AI framework presented in our study.

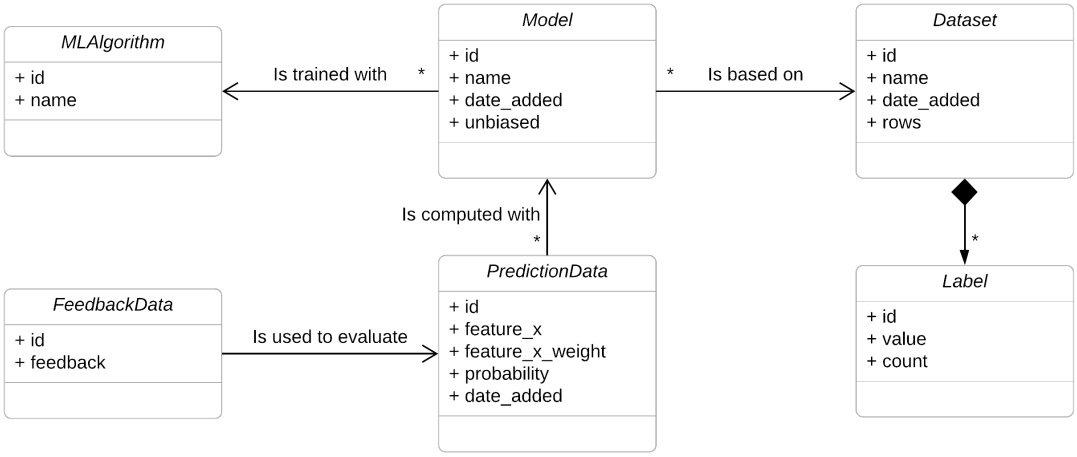


Figure 3.2. Class diagram of the proposed Responsible AI framework.

menting bias mitigation strategies. The utilization of a state-of-the-art fairness criterion enables the identification of biases within the trained model, as discussed in Section 2.2.6. This approach serves to inform users of the presence of a fair system and to offer the option of retraining a new, unbiased version of the model if necessary.

The **feedback loop** facilitates the process by which a subject matter expert, such as a loan officer, provides input on a particular prediction in order to establish a new standard of accuracy and develop an improved machine learning model.

The class diagram depicted in Figure 3.2 drafts the framework's conceptual and schematic representation. Every class is summarized below.

The *Dataset* class provides details about the data used for training the model. It includes the dataset's row count and an identifying name. The `id` attribute is employed to access the dataset content from local storage.

The values that can be assigned to the labels of a dataset are represented by the *Label* class. To distinguish each label, the information about its value and its frequency in the associated dataset is included.

The *Model* class refers to the model that has been trained using a particular dataset. The attributes of each model include an identifier, a descriptive name, and the date of addition to the system. The boolean attribute `unbiased` serves to denote whether a given model has been derived through the utilization of a bias mitigation algorithm applied to a preexisting model.

MLAlgorithm describes the algorithm employed to train the model. Additionally, it is utilized to offer extra information to the user while mitigating bias.

The prediction computed by a model for a given input instance provided by the user is represented by the *PredictionData* class. The probability returned by the model stores the outcome of the prediction. Besides, the entity contains the feature value and the related weight (or score) obtained using a *model-agnostic interpretability algorithm* for each attribute of the instance being predicted.

The information in *FeedbackData* pertains to the feedback given for a specific prediction. It is indicated by a boolean attribute, with a value of *true* indicating that the prediction corresponds with the user's expectation or the actual outcome.

3.3.1 Application workflow

The capabilities of the developed framework are outlined in Figure 3.3. In this section, the key components of each application flow are described. In order to facilitate comprehension of the diagram, it is essential to establish two underlying premises: (1) components that share the same shape, size, and name are considered to be identical; their duplication is solely for the purpose of enhancing visual representation; (2) the black dashed lines depicted in the diagram illustrate the linkage between the data and the particular processes employed.

Loan Approval System User Interface allows users to choose the operations they want to perform using the *Tab menu*. It is designed for use in a web application with a user-friendly layout.

The *Load dataset* function is used to import a dataset and save it in the system. Initially, this is the only available function when the system is started. Before the loaded dataset can be stored effectively, a predefined *Data preprocessing* step is necessary to prepare the data for subsequent processes. The *Dataset setup* module enables users to review and adjust dataset parameters, such as the saving name, column names, and data types. Multiple datasets can be imported and saved in the system concurrently, allowing users to select them as needed.

The *ML model training* functionality allows for the initiation of the training phase following the storage and availability of a dataset for selection. Simultaneously, multiple models are constructed utilizing various ML algorithms. Trained models are shown to users, accompanied by metrics such as *accuracy*, *precision*, *recall*, and *F1-score*, in order to assess their performance and compare them for selecting the best model to be stored in the system.

Users have the option to choose one of the available ML models and *request a prediction* along with its explanation. In our scenario, the predictions represent the likelihood

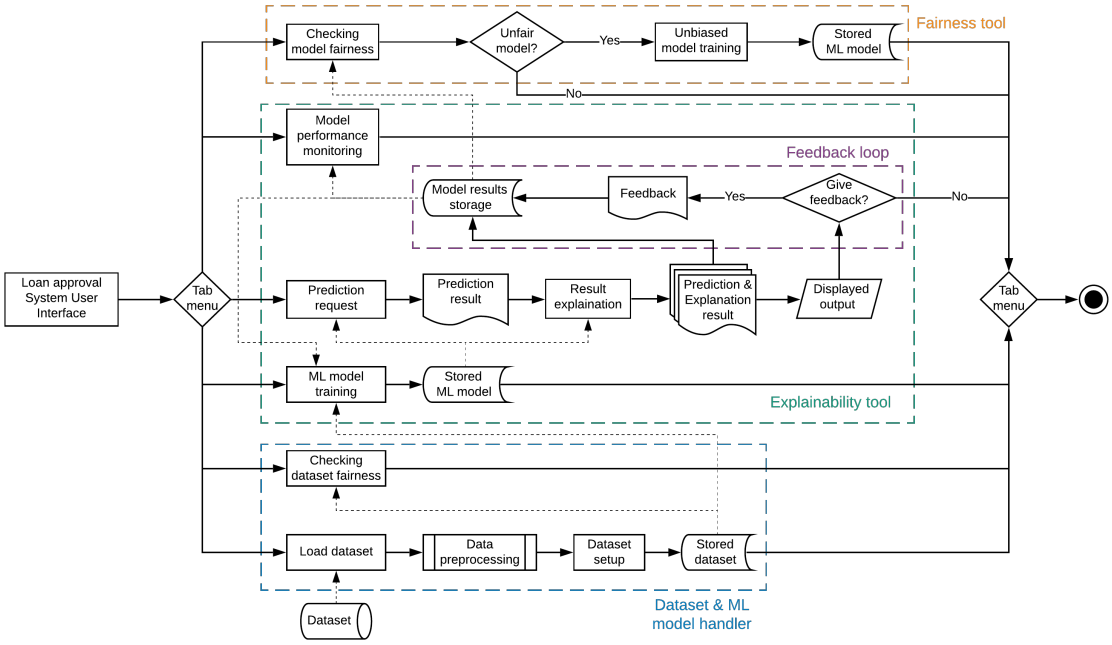


Figure 3.3. Application workflow of the presented Responsible AI framework.

of customers repaying a loan based on their credit histories and specific personal information provided as input. Following the computation of the output, users, who are domain experts at this stage, can provide feedback on the specific outcome, thereby allowing for the monitoring of the model's results in use. The predictions, explanations, and feedback are all stored in a *model result storage*.

The functionality for *monitoring model performance* utilizes user feedback to calculate statistics on the performance of the managed models, using the same metrics as those used in *ML model training*.

The process of examining dataset and model *fairness* and implementing strategies to *mitigate bias* operates in the following manner. Examination of both the initially stored dataset and the employed model is essential to assess the existence of any biases. During the process of dataset examination for diagnostic purposes, the objective is to identify any potential unfair attributes and trace them back to the original distribution of labels. Meanwhile, the insights derived from model predictions are used to assess the impact of sensitive attributes on the behavior of the model. In the event that certain decisions are deemed to be inequitable, it may be appropriate to train and store an *unbiased model*.

3.3.2 Explainability tool

The developed framework is designed to offer multiple methods for acquiring the explanation of a specific prediction. The *explainability tool* consists of two components, the *configuration* module and the *explainer interface*, which are summarized below. The meticulous illustration is available in our corresponding journal article [269].

The *Configuration* class is responsible for carrying out the necessary pre-processing steps to utilize the explainability algorithms and train explainer models. It begins by extracting categorical and numerical features from the dataset for evaluation and then applies the one-hot encoding procedure to the categorical values. Furthermore, this tool readies the instance to be explained by interpretability algorithms through the utilization of the data gathered in the earlier initialization phase.

The common interface called *Explainer interface* standardizes access to various interpretability algorithms, enabling seamless switching between explainers or simultaneous use of different explainers. Each algorithm is initialized using the previously described configuration class, and their explanations are produced. The tool can exploit three different explainability algorithms: *LIME* [286], *SHAP* [220], and *Anchors* [287].

3.3.3 Fairness tool

The system being presented offers a second crucial feature, which is the ability to examine the initial dataset's label distributions and the behavior of the trained models for detecting bias and training an unbiased model.

It can analyze the original dataset and the behavior of models to identify potential unfairness. The dataset is used to identify biases from the original data, mainly for diagnostic reasons. Nevertheless, the model's predictions are employed to assess its performance instead. If a systemic bias is identified, then a version of the model can be retrained without bias and preserved for subsequent use in making predictions. In addition to these distinctions, Algorithm 1 delineates the bias detection procedure for both the initial dataset and trained models.

Algorithm 1 Bias detection procedure

```
procedure ComputePrivilegeClasses(D)
    C = ∅;
    𝔈 = ∅;
    t ← len(D);
    G ← Select * From D GroupBy s;
    for all g ∈ G do
        if (len(C) ≥ t) and (disparate_impact(C, g) < 0.8) then
            𝔈 ← 𝔈 ∪ {C};
            C ← ∅;
        end if
        C ← C ∪ g;
    end for
    𝔈 ← 𝔈 ∪ {C};
    return 𝔈;
end procedure
```

The dataset labeled as input D (also referring to the data structure that stores the model prediction results) is examined for biases by utilizing the *disparate impact* metric

(Equation (2.3)). The process involves grouping dataset rows $d \in D$ based on the value of the sensitive attribute s. These groups are known as the *sensitive groups* $g \in G$. For each sensitive group g, the positive outcome ratio is calculated, and G is divided into one or more *privilege classes* $C \in \mathfrak{C}$, where $C = \{g_i \mid g_i \in G,\ 0 < i \leq |G|\}$, based on two criteria: each privilege class C must represent at least 5% of the entire population of D and the disparate impact between C and any other sensitive group g (or vice versa) must be less than 0.8. The first requirement ensures that each privilege class contains a statistically significant number of instances, while the threshold of 0.8 has been chosen to comply with the *80%-rule* (see Section 2.2.5).

The algorithm's output consists of a collection of privilege classes denoted as $\mathfrak{C}$. The class with the highest rate of positive outcomes is identified as the *privileged class*, while the remaining classes are known as the *unprivileged classes*. If the cardinality of $\mathfrak{C}$ exceeds 2, then the dataset or model under evaluation is deemed to be *biased*.

The system allows for training a fairer version of the model if the results are biased. To accomplish this, Algorithm 1 is utilized to divide the model into two categories: the *privileged class*, representing the sensitive feature values with the highest proportion of positive outcomes, and the *unprivileged class*, encompassing the remaining values. It is important to note that there might be sensitive feature values for which predictions are not currently available. These values will be categorized into the unprivileged class based on the previously mentioned criteria. This approach is chosen to ensure that the system, in cases where it lacks information about how a model evaluates a specific value, does not exacerbate existing unknown biases by assigning it to the unprivileged class.

After determining the allocation of feature values between privileged and unprivileged classes, a bias mitigation procedure based on the *reweighing* algorithm is employed. We selected this algorithm for several reasons: (1) the system can access the dataset used to train the examined model, allowing us to apply a preprocessing strategy like the reweighing algorithm, which is likely to yield better results; (2) the reweighing algorithm makes decisions based on the disparate impact criterion, which is the legal definition used to differentiate between privileged and unprivileged groups. (3) the algorithm produces a set of weights as its output, which is easier to interpret than other techniques. Once the new set of weights is established, the unbiased model can be trained using the identical ML algorithm employed for the original model. Subsequently, it can be saved and accessible for querying.

3.4 Case study

The following section's main focus is to demonstrate how the described system can be applied to the loan approval process context and to give a general view of the UI of the developed framework, without delving into its technical implementation details. The diagram displayed in Figure 3.4 illustrates the accessibility of the functionalities described in the preceding section to various user types within the presented system, facilitated by the developed framework.

Figure 3.5 displays the interface for loading a dataset. For our case study, we received

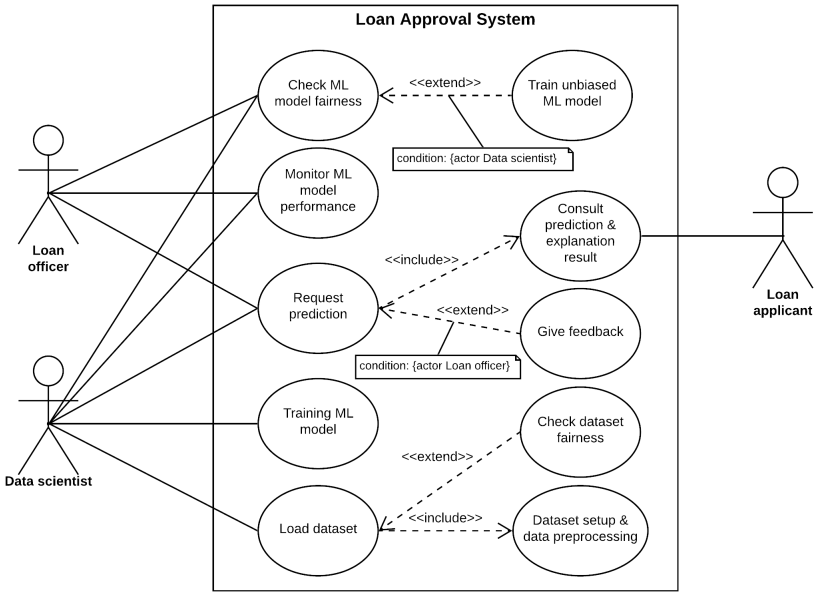


Figure 3.4. Use case diagram of the presented Responsible AI framework.

the data from an Italian banking institution after it underwent a *pseudonymisation*[4] process. The interface showcases the attributes of the loaded dataset and the outcomes of initial bias detection.

Once the dataset is loaded, users can automatically train a new ML model (see Figure 3.6) using different algorithms. In this specific case study, as it is a binary problem, we opted for *Logistic Regression*, *Random Forest*, and *Naive Bayes* algorithms. Since the dataset is unbalanced, the system ranks the trained models by the *F1-score* metric.

Upon soliciting a prediction, individuals have the capability to review the model's output alongside its corresponding explanation in a user-friendly format. Figure 3.7 depicts the presentation of the prediction result and its associated probability in the top-left box, as well as the corresponding explanation in the correct box. In the given example, these components are generated using the *SHAP* algorithm. Afterward, the user can provide feedback on the prediction through the buttons in the bottom-left box.

When the user selects the *Fairness* tab, they are presented with the classification of privilege for the most recently uploaded dataset (see Figure 3.8). The case study under consideration involves the identification of *nationality* as the sensitive attribute. The partitioning displayed in the analysis is achieved through the procedure listed in Algorithm 1. Within the interface, users can access the *Training* tab from the navigation menu on the left-hand side of the screen. This allows them to initiate the training

[4]This processing method of personal data makes it impossible to link the data to a specific individual without using additional information. This additional information is kept separately and protected by technical and organizational measures to prevent personal data from being linked to an identified or identifiable person.

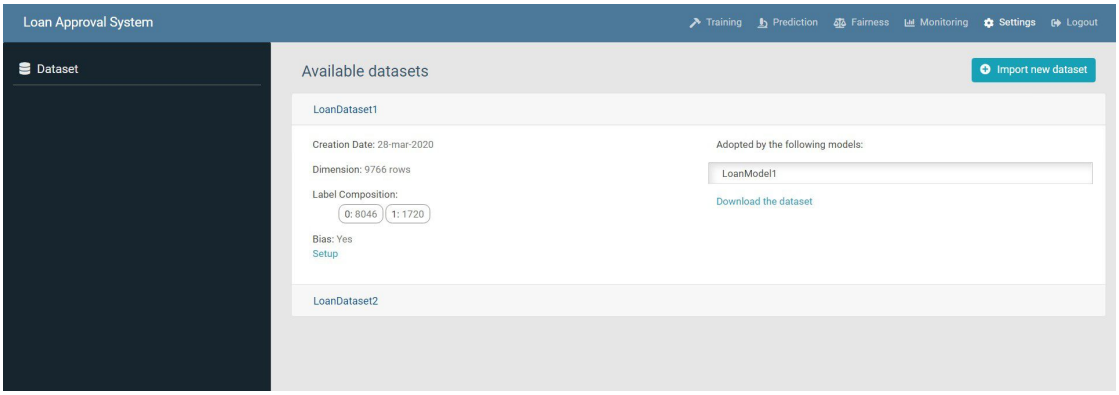


Figure 3.5. Responsible AI framework UI: Load dataset

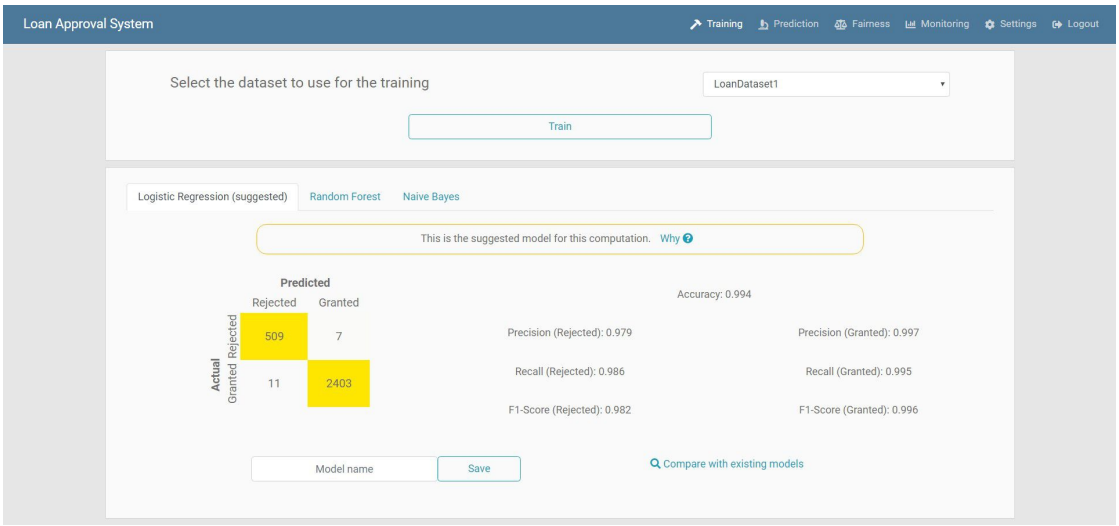


Figure 3.6. Responsible AI framework UI: ML model training

of a new model version utilizing the *reweighing* bias mitigation algorithm. After the completion of training, the model is permanently stored within the system alongside its previous version and can be accessed for prediction within the *Prediction* tab of the main interface.

A comparison between the explanations produced by an unfair and a fair model for the same instance is depicted in Figures 3.9 and 3.10. The left-hand navigation menu in Figure 3.8 allows for manual verification of fairness across a dataset, although this feature is automatically executed by the system upon dataset loading.

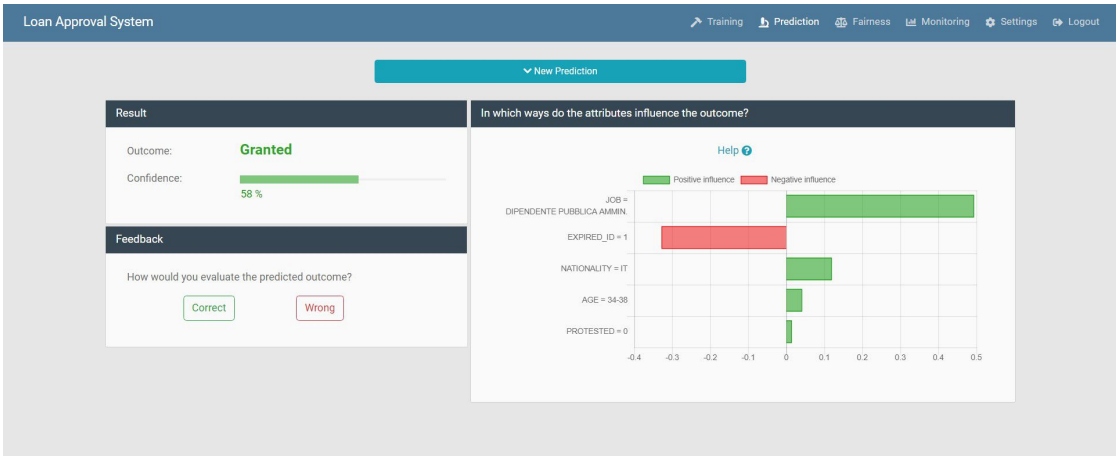


Figure 3.7. Responsible AI framework UI: Displayed prediction and explanation output

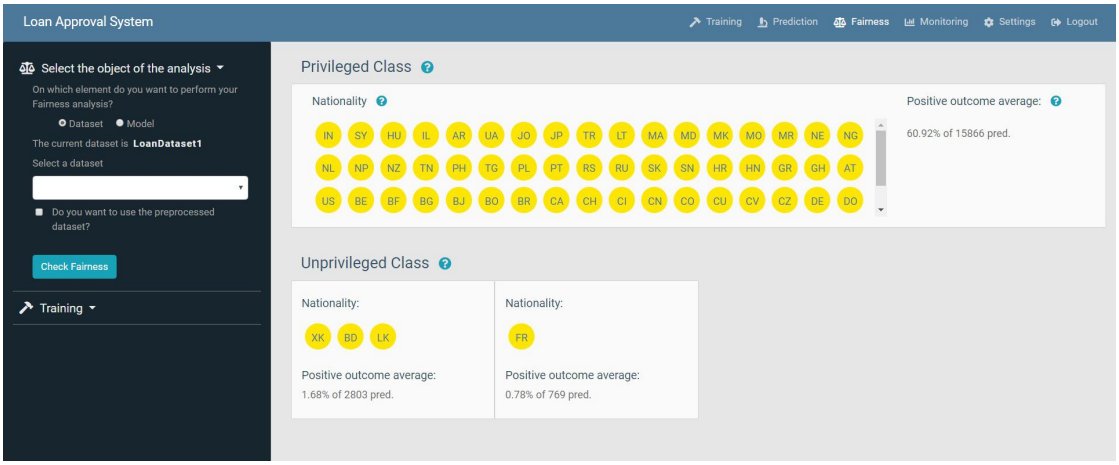


Figure 3.8. Responsible AI framework UI: Bias detection

3.5 Evaluation

Four different types of evaluation have been conducted on the presented Responsible AI framework:

1. Assessment of the most suitable explainability algorithm. This analysis, carried out by a group of data experts and researchers, centered on the algorithms that can be used in the advanced system and were further explained in Section 3.3.2. The related user study involved the utilization of the *Explanation Goodness Checklist* [153]. A group of 54 people who were not previously familiar with our system were selected to use the interface we designed for one month. After that, they were asked to fill out a checklist about their experience. The group of participants was made up of an equal number of data scientists and computer science researchers.

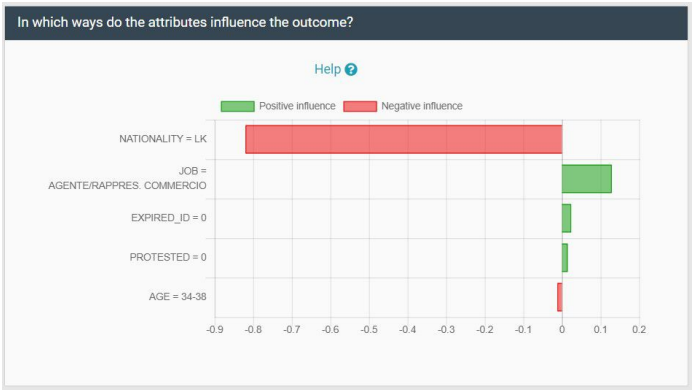


Figure 3.9. Responsible AI framework UI: Unfair model

Figure 3.10. Responsible AI framework UI: Fair model

The majority of participants (90%) reported involvement in the daily development of machine learning models, while a substantial portion (70%) indicated familiarity with XAI techniques. The gender distribution comprises 75% males and 25% females, with an average age of 27.2. The pool has been divided into three homogeneous subgroups based on the pre-determined factors. Our study utilized a *between-subject* evaluation, whereby each of the three subgroups was assigned a different algorithm in order to ensure independence in the evaluation process and mitigate potential prejudices arising from prior exposure to alternative techniques. The three explainability methods have been applied to a trained ML model with the same characteristics. The results, which are illustrated in detail in our journal article [269], showed the *SHAP* algorithm as the most satisfying, complete, and reliable.

2. Evaluation of the *Loan Approval System*, specifically from the perspective of explainability. This assessment is conducted utilizing a newly proposed *Trust & Reliance Scale* (Appendix A), which is based on the *Trust Scale Recommended for XAI* [153]. The specified new scale was used to measure the outcomes with a group

of banking industry professionals and loan officers. Section 3.5.1 presents the peculiarities of this evaluation part.

3. In the third part, presented in Section 3.5.2, the outcomes of the *A/B test* and *targeted interviews* conducted to assess the efficiency of the fairness aspect in the proposed system are showcased.

4. A *usability test* of the UI was conducted to evaluate user satisfaction, and the findings are presented in the final portion of this section, specifically in Section 3.5.3.

3.5.1 Explainability perspective

The proposed novel *Trust & Reliance Scale* is described in Appendix A. The scale we are using to assess the effectiveness of explanations for predictions is primarily based on the *Trust Scale Recommended for XAI* [153] (Q1, Q3, Q4, Q6, Q7). We have modified this scale to create a new one that is better suited for evaluating our system according to our approach. Specifically, we have eliminated questions related to predictability and efficiency and included three new items. One question, derived from Adams' work [7] (Q2), directly asks users if they trust the tool's output. We have also included a question from Hoffman's *Explanation Satisfaction Scale* [153] (Q8) to emphasize the importance of explanations to the evaluator, as well as a new question (i.e., Q5) to prompt users to consider trusting the system's response even if it differs from their own.

This new scale is implemented as a *5-point Likert scale*, based on existing literature, which suggests that the five-point format is less perplexing and helps to lower respondents' frustration, leading to higher response rates and improved response quality [25, 94]. Each user provides a response ranging from *Strongly disagree* to *Strongly agree* for each statement.

The scale is designed for experienced users and was tested after two months of continuous system use by a group of 42 bank domain experts, 33 of whom are currently loan officers. All participants are practitioners from the Italian banking institution that provided the dataset for the system prototype. The average age of the participants is 39.3 years, and their average years of experience in loan approval processes are 9.6.

To establish a baseline and effectively measure the impact of the explanations, we split the chosen testers into two similar subgroups and created two distinct testing environments. In the first environment, the group was not informed about the explanations, and the user interface was adjusted to only show the prediction results with *label* and *confidence*, as depicted in Figure 3.11. Meanwhile, the second group interacted with the actual system prototype and the user interface presented in Section 3.4 (in particular, refer to Figure 3.7).

The evaluation results are displayed in Figures 3.12 and 3.13. Visualizing the Likert scale results using *diverging stacked bar charts* is a graphical display technique that is based on Heiberger and Robbins's research on presenting findings using rating scales [149].

The result charts analysis clearly indicates that providing explanations for predictions has improved the overall assessment of the system. In both test environments, users

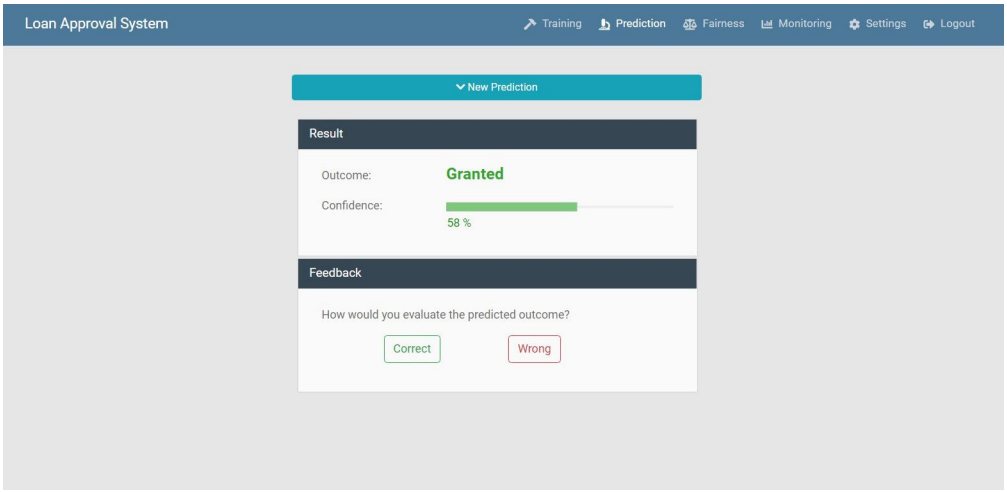


Figure 3.11. Responsible AI framework UI: Predictions *without* explanations

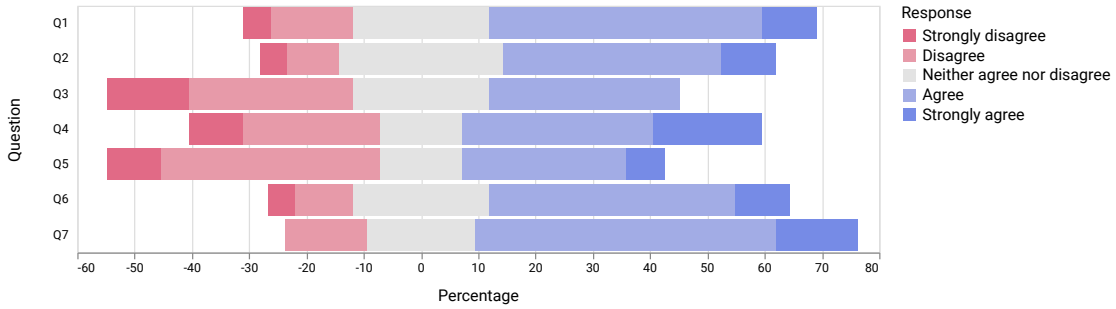


Figure 3.12. System evaluation *without* explanations (baseline)

perceived the system to work well (Q1) and appreciated the use of the automatic system for decision-making (Q7). However, there was a tangible improvement in the system's perceived trustworthiness and reliability (Q2, Q3, Q6). The explicit question about the usefulness of explanations in the second test (Q8) confirmed this perception. Displaying predictions' explanations also led to a decrease in "non-opinion" answers overall and an increase in the number of users who would change their minds based on the system's response (Q5). Surprisingly, in both environments, most users believe that such a system can produce better results than a novice human (Q4). Furthermore, we examined the characteristics of users who disagreed about the reliability (Q3) and confidence (Q5) in the system with displayed explanations. The analysis revealed that the average expertise in loan approval processes is 11.6 years, which is 2.1 years more than the overall average of the participants. This finding highlights that experienced loan officers may not be enthusiastic about integrating new technologies into their daily work.

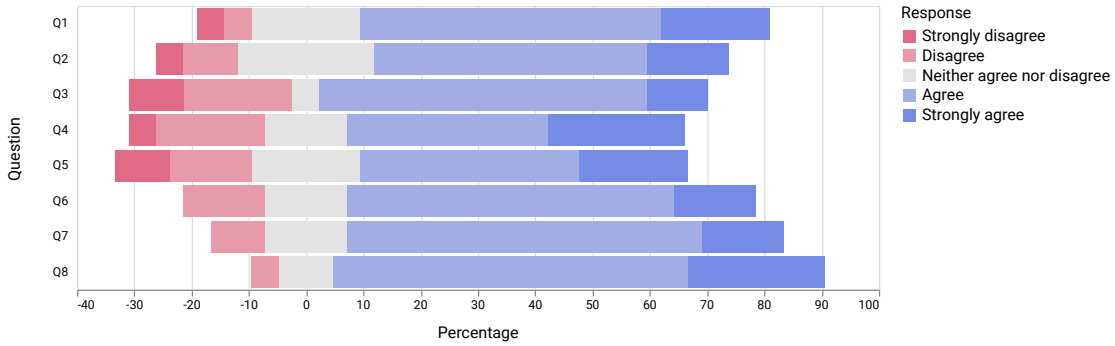


Figure 3.13. System evaluation *with* explanations

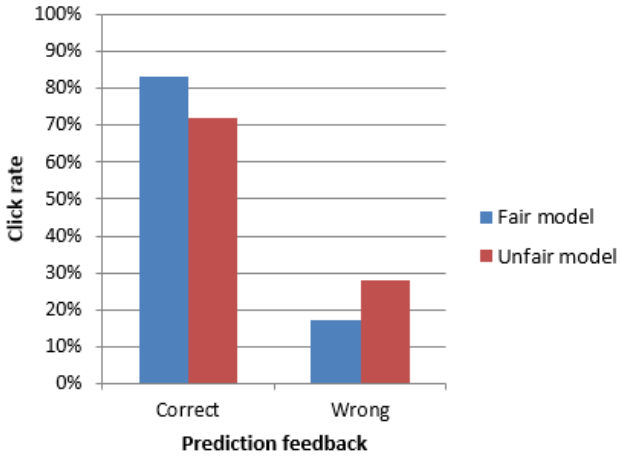


Figure 3.14. System's fairness evaluation

3.5.2 Fairness perspective

To evaluate the framework's fairness aspect effectively, we opted to conduct an *A/B test* as outlined below. Initially, we chose 24 loan officers (with an average age of 34.6 years and an average of 5.5 years of experience in the industry) who were not part of the previous assessment to participate in a testing session to assess the effectiveness of the feedback loop. Their task during the session was to evaluate the accuracy of each displayed prediction and its accompanying explanations by clicking on the dedicated button located at the bottom-left section of the UI depicted in Figure 3.7. The participants were unknowingly split into two similar subgroups to compare how they assessed the accuracy of predictions when using two contrasting models. The initial group engaged with an unfair model, as depicted in Figure 3.9, whereas the second group interacted with a fair model, as shown in Figure 3.10, where the attribute *nationality* was totally absent. The evaluation, which lasted for two hours, involved presenting 50 predictions to each user. The results in terms of click rate on the feedback buttons are shown in Figure 3.14.

The chart illustrates that unfair-model testers received a higher percentage of negative

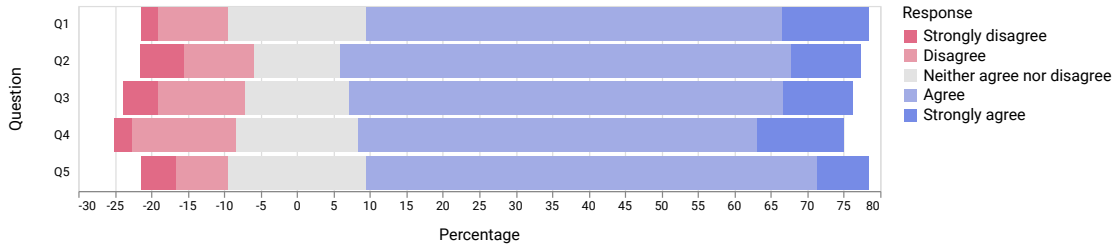


Figure 3.15. UI evaluation results: Dataset & ML model handler

responses. Following this finding, targeted interviews were conducted with some of the loan officers involved in both interactions for the second part of the evaluation. The key findings from the interviews indicate that 92% of the fair-model testers reported that they "*focused on evaluating the accuracy of the prediction based on their expertise*," while 88% of the unfair-model testers indicated that their focus often centered on the weight of the nationality attribute, as they would "*never agree to confirm the rejection or approval of a loan application in which the primary factor is a potentially discriminatory individual characteristic such as the applicant's nationality.*" Even though it was not included in the test, they all agreed that visualizing the explanations of the predictions was crucial for this type of automated system.

3.5.3 User interface

Finally, the developed explainable UI is qualitatively evaluated to measure user satisfaction with the system's usability.

The bank domain experts who participated in the previous system evaluation also attended this experimental session. The questionnaire, which can be found in Appendix B, is based on the usability test proposed in one of our pre-doctoral publications (i.e., [273]) and is structured following a methodology presented by IBM [205] but adapted to a five-point format for the reasons mentioned above. Each participant tested the three functionalities for one month and then evaluated them using the same procedure described in the previous section.

The three primary functionalities of the system, including the *dataset & ML model handler*, *explainability tool*, and *fairness tool*, were tested. The results, shown in Figures 3.15 to 3.17, indicate that users find the UI effective. However, the fairness tool needs improvement to make it easier to find the required information for specific tasks.

3.6 Summary

In this chapter, we introduced a Responsible AI framework focusing on the ethical principles of *explainability* and *fairness* in AI. The system is applied to loan approval processes using a proprietary framework to manage the ML model life cycle. Four functionalities were developed: a *dataset & ML model handler*, a standardized *explainability*

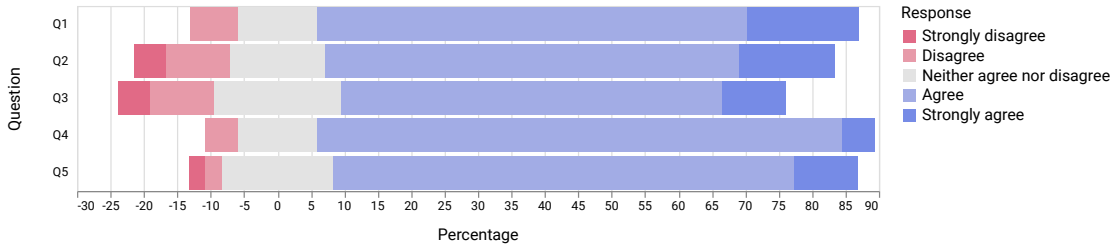


Figure 3.16. UI evaluation results: Explainability Tool

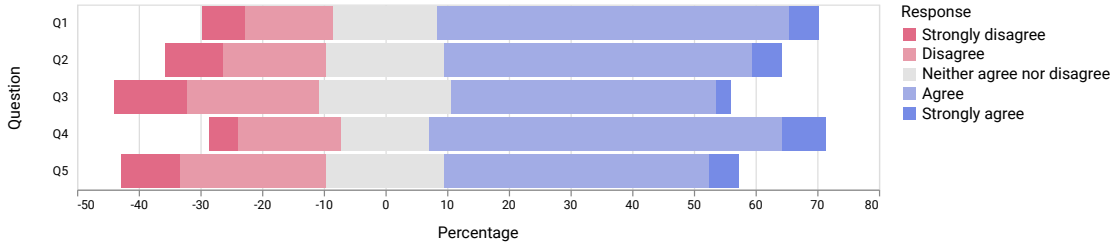


Figure 3.17. UI evaluation results: Fairness Tool

tool, a *fairness tool*, and a *feedback loop*. The system's effectiveness was demonstrated through comprehensive user studies. SHAP was chosen as the preferred explainability algorithm, and the novel proposed *Trust & Reliance Scale* was used to assess increased trust. The fairness tool includes a bias detection and mitigation algorithm based on *reweighing* to address model biases, and it was evaluated through an A/B test. A user-friendly explainable UI was developed and tested for usability.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Chapter 4

Fairness Assessment of Graph Neural Networks in Binary User Modeling Scenarios

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.

Marcel Proust

In this chapter, we will discuss the fairness assessment of state-of-the-art Graph Neural Networks (GNNs) designed for behavioral user modeling tasks in real-world binary scenarios. Additionally, we will address the challenges that exist for fairness assessment in such scenarios, setting the stage for our research in multiclass and multigroup bias detection.

In particular, for the first part, we conduct two user modeling tasks by performing binary classification on two real-world datasets, utilizing the most effective GNNs available in this domain. Subsequently, we evaluate *disparate impact* and *disparate mistreatment* in GNNs tailored for behavioral user profiling, employing four distinct algorithmic fairness metrics. Through an extensive series of experiments, we identify three significant insights into the examined models, linking their different user modeling approaches to the fairness metric scores.

In the second part of the study, we will examine the results of the fairness analysis described earlier to provide considerations about the issues associated with using binary fairness metrics. We will conduct two different types of experiments, adapting the ones we discussed in the first part to present our perspective on the challenges we mentioned. In the first experiment, we will focus on the analysis of the *absolute difference* score usage of the computed fairness metrics. In the second experiment, we will conduct a preliminary study to explore a responsible fairness analysis considering the original *multiclass* distribution of the sensitive attribute that we are investigating.

In Section 4.1, we present the research studies and outline the significance of exam-

ining fairness in GNNs for behavioral user modeling. Following this, Section 4.2 details the specific GNN-based models analyzed, the datasets used in our experiments, and the metrics adopted for assessing fairness. Section 4.3 provides an in-depth description of our study aimed at evaluating the fairness of these models, offering insights into how different modeling paradigms impact bias detection metrics scores. We then address broader issues in Section 4.4, discussing two significant open challenges in current algorithmic fairness practices. Finally, the chapter concludes with Section 4.5, encapsulating our research's key findings and implications.

4.1 Motivation

In the past few years, *user modeling* (Section 2.4) has emerged as a crucial subject in various real-world situations, particularly social networks [211] and e-commerce [360], owing to the vast volume of data offered by web applications and platforms. User modeling (or user profiling) aims to deduce an individual's interests, personality traits, or behaviors from collected data in order to create an effective user representation, known indeed as a *user model* (or user profile), which is commonly utilized by adaptive and personalized systems [103, 267]. In the initial stages of modeling, the focus was placed solely on the examination of static attributes (referred to as *explicit user modeling*), typically relying on data obtained from online forms and surveys. Nevertheless, these methods have demonstrated inefficacy as users exhibit a lack of concern regarding the direct provision of their personal information. The current trend in modern systems, as outlined in Section 2.4.3, emphasizes an implicit approach for modeling users' data through the analysis of individuals' actions and interactions (referred to as *implicit user modeling*). The aforementioned strategy is commonly known as **behavioral user modeling** in literature, and it is a key topic in the presented doctoral research.

An effective method of representing such behaviors is to adopt graphs, in which the connections between users are depicted by edges, and the users are denoted by nodes. **Graph Neural Networks**, as accurately introduced in Section 2.3, have exhibited efficacy in modeling graph data across various domains, including recommender systems [147, 378], natural language processing [375], text mining [321], and user profiling [68, 69, 372].

To offer some background from the literature in this context, the first steps toward user modeling on graph data were taken by Li et al. [208] in 2012, who leveraged a heterogeneous graph based on "following" and "tweeting" interactions to infer users' locations. Rahimi et al. [281] suggested a geolocation model that relies on Graph Convolutional Networks (GCNs) to detect users' location by incorporating text and network data. A Heterogeneous Graph Attention Network (HGAT) was introduced by Chen et al. [69] to learn user representations by considering the graph structure and incorporating an attention mechanism to discern the importance of each node's neighbor.

In 2021, two of the most effective GNN-based architectures for user modeling were introduced, and these are also the primary models examined in this doctoral research and are further elucidated in Section 4.2.1. Chen et al. [68] presented a model based on GCN

that demonstrates the advantages of enhancing node representation before conducting user profiling tasks. Yan et al. [372] suggested a Heterogeneous Graph Network (HGN) to enhance prediction accuracy by considering various types of relations and entities for user profiling, in contrast to previous methods relying on single types. In general, in Section 2.4.3, we stated how current methods assess user profiling models based on their ability to effectively classify a user's personal characteristics, such as gender or age.

While GNNs have demonstrated success in classifying user models, it is important to acknowledge that, like all machine learning (ML) systems trained on historical data, they are susceptible to learning biases inherent in the data and subsequently reflecting these biases in their output. The observed phenomenon can be primarily attributed to the topological characteristics of graph structures and the conventional message-passing mechanism employed by GNNs (see Section 2.3.4). This process can exacerbate discriminatory effects, as nodes sharing the same sensitive attribute are more likely to be interconnected than those with differing attributes.

The concept of **algorithmic fairness** (Section 2.2) has become increasingly important as automated decision-making systems are being more widely used. Over the past few years, a limited number of scientific works addressed the issue of fairness evaluation in GNNs (e.g., [10, 82, 97, 232]). However, at the beginning of our study, none of them specifically investigated the potential for discrimination in the state-of-the-art GNN-based models for user modeling tasks. A significant amount of contributions has been instead released regarding overall approaches for identifying and addressing bias in ML models (e.g., [33, 63, 77, 335, 341]), each focusing uniquely on a specific aspect of what might be considered *fair*.

The existence of hidden unfair practices in these models presents pragmatic risks. In fact, while they do not intentionally create unfairness (i.e., *disparate impact*, see Section 2.2.4) by solely focusing on behavioral data, ML models can still lead to systematic disparities in services if they are more effective for certain demographic groups. For instance, if a system consistently provides less accurate predictions for *gender* within a specific *age* group, that group will always receive inadequate service, such as ads targeted toward the opposite gender. Therefore, identifying and addressing unfairness in behavioral user modeling is essential in this field.

Although a widely accepted definition of fairness has not been established, most of the measures focusing on *classification parity*[1] commonly aim to identify and rectify bias and inequity in *binary* problems [72]. The widespread adoption of this practice can be attributed to two primary motivations, as identified by Caton and Haas [63]: (1) many applications involving ML models are inherently binary (e.g., hiring processes, loan granting procedures, spam detection); (2) mathematically, it is more suitable to quantify fairness on a binary dependent variable. While these two justifications are technically valid, our study aims to delve into and analyze the potential implications of applying such binary metrics in user modeling, particularly in real-world scenarios, from an ethical perspective. Our arguments align with a similar criticism put forth by Barocas et al. [33]:

[1]It means that predictive performance scores, such as true positive, true negative, false positive, and false negative rates, should be the same across groups identified by the chosen sensitive attributes.

"*Most proposed fairness interventions start by assuming such a (binary) categorization. But when building real systems, enforcing rigid categories of people can be ethically questionable.*".

Moreover, an additional challenge arose from the research on bias detection literature. When assessing the performance of a model in producing equitable outcomes, it is frequently analyzed through the measurement of the *absolute difference* between the scores of the two sensitive groups being studied. However, this approach entails potential risks from both a systemic and user standpoint, as it makes it incredibly difficult to determine disadvantaged groups within any possible combination of model, dataset, and fairness metrics. Therefore, the implementation of concrete interventions to address these issues in a real-world scenario becomes not feasible.

4.2 Methodology

This section illustrates the state-of-the-art GNN-based models adopted in the studies presented in the chapter, along with the real-world datasets and fairness metrics employed in the experimental phases.

4.2.1 Analized models

The evaluation conducted in this study leverages two recently published GNN-based models that exemplify the latest advancements in user modeling tasks, namely CatGCN and RHGN.

CatGCN [68] is a model based on a graph convolutional network (GCN) that is designed to conduct graph learning using categorical node features. Instead of utilizing the original node representation, the model incorporates two additional types of interaction into the learning process. The first type is a local interaction that involves multiplication and is carried out on every pair of node features. The second type is an addition-based interaction, where the model leverages an artificial feature graph. By introducing these interaction types prior to graph convolution, the model aims to enhance the effectiveness of user modeling. The architecture of this model is shown in Figure 4.1.

RHGN [372] stands for Relation-aware Heterogeneous Graph Network, which is created to represent various relations on a graph containing different types of entities. The main components of this model include a transformer-like multi-relation attention, which is used to understand the importance of nodes and determine the significance of meta-relations on the graph, and a heterogeneous graph propagation network, which is used to gather information from multiple sources. This architecture, shown in Figure 4.2, performs better than several GNN-based models on tasks related to user modeling.

4.2.2 Datasets

Two public real-world datasets, specifically obtained from the popular e-commerce platforms Alibaba and JD, were selected for the user modeling studies illustrated in this chapter.

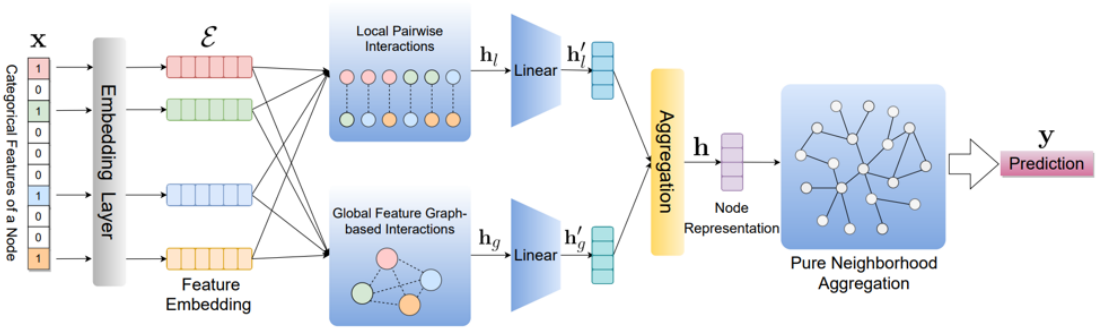


Figure 4.1. Architecture of CatGCN [68].

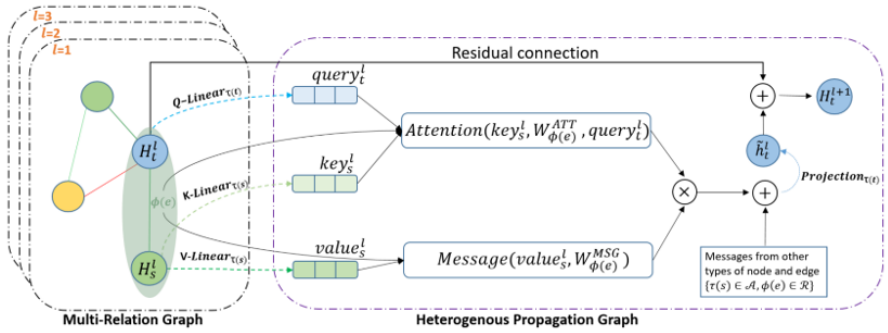


Figure 4.2. Architecture of RHGN [372].

Alibaba dataset[2] contains click-through rate data for ads on the Taobao platform, and it was provided by the Alibaba Group's Tianchi Lab in 2018. Both CatGCN [68] and RHGN [372] models were originally evaluated using this dataset. The heterogeneous graph used as input for the models consists of two types of nodes: users and items (i.e., products). User nodes contain attributes such as gender, age, consumption grade, student status, and region of living. Item nodes have a single attribute, which is the category to which the product belongs. The relationship between user and item nodes is based on clicks, and the edges are not weighted. In accordance with CatGCN's experimental setup, we used the types of products as categorical features associated with the users for the same model. Therefore, only items clicked by at least two users were considered to establish the *co-click* relationship, which was used as the model's local interaction. To ensure consistency between the evaluated GNNs, the same filtering procedure was applied to RHGN before constructing the heterogeneous graph.

JD dataset[3] comprises 100 000 users randomly selected from JD.com, which is one of the largest e-commerce platforms globally. The data was gathered by Chen et al. [69] and includes user profiles, information about items (i.e., products), as well as click and

[2] `https://tianchi.aliyun.com/dataset/dataDetail?dataId=56`. Accessed March 30, 2025.

[3] `https://github.com/guyulongcs/IJCAI2019_HGAT`. Accessed March 30, 2025.

Table 4.1. Alibaba and JD dataset characteristics.

Dataset	Users	Items	Edges	Features
Alibaba	166 958	64 553	427 464	2 820
JD	38 322	49 634	315 970	2 056

order logs spanning from February 2018 to February 2019. This dataset was utilized in the original RHGN paper [372] for experimental purposes. Gender and age form the user profile data, while category information, brand, and price constitute the product data. These details are used to generate the user and item nodes for the input graph. Due to the dataset's considerable size, and since our work does not specifically focus on evaluating models' performance in user modeling tasks, we selected a sample representing 15% of the items. Additionally, we decided to consider only one type of relationship, namely *click*, as the graph edges. This approach ensures that our experimental setups are comparable. A *co-click* relationship is employed as the local interaction for CatGCN to replicate the process of the Alibaba dataset.

Table 4.1 displays the characteristics of the two datasets described.

4.2.3 Adopted metrics

The fairness metrics employed in the presented studies follow the notation described in Table 2.1. The primary focus of this chapter is to evaluate the fairness of the GNNs introduced in the previous section in relation to *disparate impact* and *disparate mistreatment*.

According to the definition provided in Section 2.2.4, the concept of disparate impact arises when a model unfairly discriminates against certain groups, even if the model does not directly use the sensitive attribute to make predictions but instead relies on other related attributes. This is the case with the analyzed GNNs, where user models are built by combining information from neighboring nodes, and the sensitive attribute is not explicitly considered during classification. The idea of disparate impact is useful when no clear link exists between the predicted label and the sensitive attribute in the training data. In other words, it is difficult to determine the fairness of a decision for a group member based on historical data. In our study, we follow a common practice applied in various studies on fairness in decision-making systems (e.g., [82, 382]) and evaluate the disparate impact of the analyzed models using **statistical parity** (Equation (2.2)) and **equal opportunity** (Equation (2.4)) metrics. To expand the analysis of disparate impact from previous studies, we also utilize the **overall accuracy equality** (Equation (2.6)) metric to assess the relative accuracy among different groups.

In our case studies, we focus on a situation where it is challenging to determine the accuracy of a prediction involving sensitive attribute values. In our research, we propose that a comprehensive evaluation of fairness should consistently consider the aspect of disparate mistreatment (see Section 2.2.4). This notion examines the *misclassification rates* for user groups with varying sensitive attribute values instead of focusing on cor-

rected predictions. In addition, disparate mistreatment is important in scenarios where misclassification costs depend on the group affected by the error. In particular, we employ the **treatment equality** (Equation (2.8)) metric to assess this fairness perspective.

4.3 Fairness analysis

We perform thorough empirical research to examine the following research questions in order to conduct the fairness evaluation of the aforementioned GNN-based models in the binary user modeling scenario:

RQ1 How does the variation in input types of the examined GNNs and the construction of user models influence fairness?

RQ2 How *fair* are the user models created by the state-of-the-art GNNs that were analyzed?

RQ3 To what extent is it necessary to consider disparate mistreatment in order to fully assess the presence of unfairness?

Below, we outline the experimental settings set up to address each research question and the parameters selected to train the models before assessing fairness. To conclude the section, we present the evaluation results and discuss the derived insights.

4.3.1 Experimental settings

In order to measure the disparate impact and disparate mistreatment of the models being analyzed, we operationalize the metrics described in Section 4.2.3 by defining the following scores, according to similar contributions in the field [35, 82]:

$$\Delta_{SP} = |P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1)| \tag{4.1}$$

$$\Delta_{EO} = |P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid y = 1, s = 1)| \tag{4.2}$$

$$\begin{aligned}\Delta_{OAE} = |&P(\hat{y} = 0 \mid y = 0, s = 0) + P(\hat{y} = 1 \mid y = 1, s = 0) - \\ &- P(\hat{y} = 0 \mid y = 0, s = 1) - P(\hat{y} = 1 \mid y = 1, s = 1)|\end{aligned} \tag{4.3}$$

$$\Delta_{TE} = \left| \frac{P(\hat{y} = 1 \mid y = 0, s = 0)}{P(\hat{y} = 0 \mid y = 1, s = 0)} - \frac{P(\hat{y} = 1 \mid y = 0, s = 1)}{P(\hat{y} = 0 \mid y = 1, s = 1)} \right| \tag{4.4}$$

We investigated **RQ1** and **RQ2** to assess disparate impact by performing a user modeling task for CatGCN and RHGN models and calculating the related fairness score in terms of Δ_{SP}, Δ_{EO} and Δ_{OAE}, defined, respectively, by Equations (4.1) to (4.3).

For our experiments, we specifically focused on considering the users' *gender* as the label for the user modeling task, and their *age* as the attribute to be evaluated for fairness in both datasets. Given that the evaluation takes place in a binary scenario, we needed to split the sensitive feature into two groups (generating the *bin-age* attribute). The age

Table 4.2. Distribution of label and sensitive attribute values of ALIBABA and JD datasets for fairness assessment in binary scenario.

Dataset	**Label**	**Count (Percentage)**	
		Class 1	*Class 0*
ALIBABA	gender	42 192 (25.3%)	124 766 (74.7%)
JD	gender	13 735 (35.8%)	24 587 (64.2%)
Dataset	**Sens. Attr.**	**Count (Percentage)**	
		Class 1	*Class 0*
ALIBABA	bin-age	71 583 (42.9%)	95 375 (57.1%)
JD	bin-age	25 717 (67.1%)	12 605 (32.9%)

range for each class in the ALIBABA dataset is not defined, and it is only distinguished by a label. The split is created to establish a distinct separation between the two groups. In the JD dataset, labels for the age attribute are provided, and the feature is binarized to split users under and over 35 years old. The distribution of target class and sensitive attribute values within the datasets is presented in Table 4.2.

In particular, to address **RQ1**, we evaluated the discrimination level of the two models by comparing the scores of the three mentioned metrics. This allowed us to assess the impact of different user modeling paradigms on fairness scores. The classifier fairness increases as these scores decrease.

For **RQ2**, we compared the scores obtained in terms of Δ_{SP} (Equation (4.1)) and Δ_{EO} (Equation (4.2)) with the results of FAIRGNN (in the original publication [82]), a recent GNN architecture developed to generate fair outcomes for node classification.

In order to answer **RQ3**, we broaden our fairness analysis to include Δ_{TE} (Equation (4.4)). Our goal is to assess how much the examined models discriminate against users in terms of disparate mistreatment as opposed to disparate impact.

For the user profiling tasks, the hyper-parameters of the models were specified in the following manner. The *learning rate* for CATGCN is explored within the range {0.01, 0.1}, the L_2 regularization coefficient and the *dropout ratio* are searched in {1e-5, 1e-4} and {0.1, 0.3, 0.5, 0.7}, respectively, while the aggregation parameter α value is obtained from {0.1, 0.3, 0.5, 0.7, 0.9}. RHGN model requires more computational time for each individual experiment, resulting in a narrower range of hyper-parameter options compared to other models. The *learning rate* and the L_2 regularization coefficient are directly set to 0.01 and 0.001, respectively; the hidden dimension of the two layers of the entity-level aggregation network gets value from {32, 64}, while the number of heads in multi-head attention is searched in {1, 2}. All other parameters are configured based on the original papers.

Following the grid search, fairness experiments are conducted 10 times, and the test set is used to evaluate the probabilities. The experiments are carried out using an Nvidia Quadro RTX 8000 48GB GPU, and the source code is publicly available[4].

[4] `https://github.com/erasmopurif/do_gnns_build_fair_models`.

Table 4.3. Experimental results of the binary user modeling task. The best result for each dataset and metric is reported in **bold**. The ↑ symbol means that higher scores are better.

Dataset	Label	Model	Performance		
			Accuracy (↑)	*F1-score* (↑)	*AUC ROC* (↑)
Alibaba	gender	CatGCN	0.787 ±0.017	**0.714** ±0.006	**0.714** ±0.008
		RHGN	**0.812** ±0.005	0.704 ±0.017	0.681 ±0.016
JD	gender	CatGCN	0.721 ±0.007	**0.706** ±0.006	**0.712** ±0.006
		RHGN	**0.735** ±0.005	0.696 ±0.007	0.658 ±0.008

Table 4.4. Experimental results of the fairness assessment in the binary scenario in terms of Δ_{SP} and Δ_{EO}. The best result for each dataset and metric is reported in **bold**. The ↓ symbol means that lower scores are better.

Dataset	Sens. Attr.	Model	Fairness	
			Δ_{SP} (↓)	Δ_{EO} (↓)
Alibaba	bin-age	CatGCN	0.046 ±0.019	0.147 ±0.080
		RHGN	**0.018** ±0.013	**0.133** ±0.086
JD	bin-age	CatGCN	0.033 ±0.013	0.050 ±0.017
		RHGN	**0.009** ±0.007	**0.041** ±0.017

Table 4.5. Experimental results of the fairness assessment in the binary scenario in terms of Δ_{OAE} and Δ_{TE}. The best result for each dataset and metric is reported in **bold**. The ↓ symbol means that lower scores are better.

Dataset	Sens. Attr.	Model	Fairness	
			Δ_{OAE} (↓)	Δ_{TE} (↓)
Alibaba	bin-age	CatGCN	0.175 ±0.109	0.068 ±0.021
		RHGN	**0.148** ±0.101	**0.017** ±0.013
JD	bin-age	CatGCN	0.062 ±0.020	0.150 ±0.066
		RHGN	**0.054** ±0.017	**0.019** ±0.015

Table 4.6. Variations in fairness scores between CATGCN and RHGN. Differences in averages are considered. The best result for each dataset and metric is reported in **bold**. The ↓ symbol means that lower scores are better.

Dataset	Variations in fairness scores			
	Δ_{SP} (↓)	Δ_{EO} (↓)	Δ_{OAE} (↓)	Δ_{TE} (↓)
ALIBABA	0.028	0.014	0.027	**0.051**
JD	0.024	0.009	0.008	**0.131**

4.3.2 Evaluation results

For every dataset and model, we initially present the performance results of the user modeling task evaluation in terms of *accuracy*, *F1-score*, and *AUC ROC* in Table 4.3. Subsequently, we provide the fairness assessment results in Tables 4.4 and 4.5.

Based on performance, RHGN outperforms its counterpart in both datasets. However, considering the F1-score and the AUC ROC, CATGCN appears to be more effective in these two aspects. Therefore, CATGCN is less influenced by false positives and false negatives. Conversely, RHGN generates more true positives and true negatives.

When examining fairness values, it was found that RHGN leads to less discrimination in its outcomes compared to CATGCN. This result holds true for all the metrics that were considered.

RQ1 finding *The effectiveness of RHGN in depicting users through multiple interaction modeling yields fairer results compared to a model like CATGCN that solely depends on binary connections between users and items. Moreover, CATGCN exacerbates bias by modeling users' local interactions (e.g., the co-click relationship).*

When we compare the fairness scores with the results of the baseline model, i.e., FAIRGNN, we can observe different situations: if we consider that the values should be close to 0 for a model to be considered fair for a specific metric, we find that only RHGN is effective with respect to Δ_{SP} in both experimental settings. However, in all other cases, neither of the two analyzed models can be considered fair.

RQ2 finding *Despite RHGN proving to be a more equitable model than CATGCN, it is equally important to implement a debiasing process to ensure that the user models generated by both GNNs are fair.*

Table 4.6 shows the extended fairness evaluation by presenting the variations in the metric scores between CATGCN and RHGN. The outcomes indicate that the most significant variance is observed for Δ_{TE}. This result emphasizes the importance of considering both disparate impact and disparate mistreatment metrics to gain a comprehensive understanding of the fairness landscape.

Table 4.7. Experimental results of the fairness assessment in the binary scenario in terms of Δ^*_{SP} and Δ^*_{EO} (i.e., Δ_{SP} and Δ_{EO} without absolute value). Negative results reflect the situation in which the Δ_{SP} or Δ_{EO} value is higher for $s = 1$, meaning that this could be the advantaged group.

Dataset	Model	Δ^*_{SP}	Δ^*_{EO}
Alibaba	CatGCN	-0.045 ±0.021	0.139 ±0.074
	RHGN	0.019 ±0.012	-0.133 ±0.086
JD	CatGCN	0.033 ±0.013	-0.052 ±0.016
	RHGN	0.009 ±0.007	-0.042 ±0.017

RQ3 finding *In conditions where the accuracy of a decision regarding the desired label with respect to the sensitive attributes is unclear, or where there is a significant cost for incorrect predictions, a thorough fairness evaluation should always consider disparate mistreatment assessment, as disparate impact findings may be misleading in these particular cases.*

4.4 Challenges in binary scenarios

In the algorithmic fairness literature, many researchers disagree regarding the practice of binarizing the target class and using absolute value scores to compute fairness, as discussed in Section 4.1. To illustrate the limitations of these approaches, we conducted an exploratory evaluation consisting of two different experiments that are based on the assessment described in Section 4.3.

The first experiment focused on the *absolute difference* of the metrics commonly employed in fairness analyses. The assumed approach was to eliminate the absolute value from the models' fairness assessment while running the same experiments reported in Table 4.4. Specifically, the metrics adopted are the following:

$$\Delta^*_{SP} = P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1) \tag{4.5}$$

$$\Delta^*_{EO} = P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid y = 1, s = 1) \tag{4.6}$$

The results shown in Table 4.7 illustrate the calculated values of Δ^*_{SP} (Equation (4.5)) and Δ^*_{EO} (Equation (4.6)), demonstrating a clear pattern of alternating positive and negative scores. This observed pattern indicates that unfairness, irrespective of the exact numerical value, could be biased against either one of the sensitive groups depending on the model and dataset combination.

To analyze fairness in binary situations, our study involved running experiments to assess how binarization practices can impact the perception of fairness evaluation. The following discussion will only concentrate on the results related to a specific combination of model and dataset as a preliminary examination. In the next chapter, we will show how the implications derived from these experiments can be extended and analyzed more thoroughly in different contexts.

Table 4.8. Experimental results of the preliminary study in binary and multiclass sensitive attribute groups for RHGN model and ALIBABA dataset. The *statistical parity* scores (*SP* in table) refer to the computation of a single side of Equation (2.2) for each sensitive group in the binary and multiclass settings.

Binary group	*SP*	Multiclass group	*SP*
s_A	0.887 ±0.015	s_0	0.81 ±0.02
		s_1	0.91 ±0.02
		s_2	0.91 ±0.01
		s_3	0.92 ±0.01
s_B	0.797 ±0.055	s_4	0.89 ±0.01
		s_5	0.72 ±0.03
		s_6	0.78 ±0.07

For the specific experiment, we focused on the RHGN model and ALIBABA dataset described, respectively, in Sections 4.2.1 and 4.2.2. We adopted the original binary classification task but with a different setting for the sensitive attribute (i.e., *age*): we analyzed its original multiclass distribution (i.e., seven groups, referred to as s_0-s_6) and computed each individual *statistical parity* (SP) probability, which means considering a single side of Equation (4.1) for every sensitive group. Additionally, as per the original study in Section 4.3, we transformed the attribute into binary groups and once again calculated the probabilities for the binary groups. The generated binary sensitive groups are structured as follows: $s_A = \{s_0, s_1, s_2, s_3\}$, $s_B = \{s_4, s_5, s_6\}$.

The results displayed in Table 4.8 indicate that converting data into binary form can result in an inaccurate evaluation of particular subgroups. In this experiment, subgroup s_0 should be viewed as disadvantaged when assessed in a detailed multiclass analysis but would be considered advantaged when included in the binary group s_A. Conversely, s_4 would be seen as disadvantaged in the binary group while individually advantaged.

4.4.1 Ethical considerations

The results presented in the previous section have several ethical implications, leading us to argue the following positions about the challenges we discussed in this chapter and have paved the way for a comprehensive study of fairness in multi-valued scenarios.

Ethical implication 1 *Relying exclusively on the absolute difference score in fairness analysis can be risky. This approach prevents from clearly identifying disadvantaged groups for each combination of model, dataset, and fairness metrics, which in turn hinders the ability to implement tailored interventions to address these issues in real-world scenarios.*

Ethical implication 2 *Many current studies on evaluating the fairness of automated systems involve converting sensitive attributes, which are inherently multiclass, into binary attributes to comply with standard fairness metrics definitions. From our perspective, assessing fairness by examining the distribution of sensitive groups is crucial for*

two main reasons. First, if the system is less effective for certain specific individual groups, they will receive less effective or wrong services, such as targeted advertisements or recommendations. Second, reducing the various classes and groups into a binary representation can result in an inaccurate assessment of model fairness, potentially distorting the original data conditions.

4.5 Summary

This chapter examined recent GNN-based behavioral user modeling architectures. Unlike previous studies that primarily focused on the predictive performance of these models, we focused on potential disparities in classification outcomes across different demographic groups, highlighting unfairness issues. Our analysis, conducted on two state-of-the-art models and real-world datasets using four fairness metrics, revealed that directly modeling raw user interactions can disproportionately misclassify a specific demographic group compared to its counterpart.

Additionally, we identified and discussed two significant open challenges in the context of algorithmic fairness. First, we critiqued the common practice of conducting fairness assessments exclusively in binary classification scenarios, emphasizing the need for a multiclass approach that recognizes the complexity of disadvantaged groups. Second, we questioned the reliance on the absolute difference of metric scores to evaluate model fairness, advocating for a more nuanced understanding of these metrics. Through a case study leveraging the same GNN-based models of the initial analysis, we argued for these methodological improvements and discussed the ethical implications arising from our experimental findings.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Chapter 5

Multiclass and Multigroup Fairness Assessment

Fairness does not mean everyone gets the same. Fairness means everyone gets what they need.

Rick Riordan

In this chapter, we aim to contribute significant advancements to the domain of algorithmic fairness by proposing innovative metrics tailored for a responsible evaluation of user modeling systems within intricate, multi-valued contexts. Building upon the foundation laid by our preliminary analysis in the previous research study (see Chapter 4), we thoroughly investigate both the ethical dimensions and the often controversial aspects associated with traditional binary fairness assessments.

In response to these challenges, we broaden the scope of classification fairness metrics to include scenarios where both the target classes and the sensitive attributes are multiclass, marking our work as one of the early comprehensive efforts in this area. We undertake a rigorous evaluation using four real-world datasets, providing a thorough assessment of fairness across both straightforward binary settings and more complex multiclass or multigroup configurations. Through our research, we expose the inherent risks and potential misrepresentations caused by the oversimplification of attributes into binary categories, which can erroneously reveal an appearance of fairness. Our findings emphasize the crucial importance of implementing more comprehensive and context-aware methods to accurately identify and mitigate unfair practices in user modeling, demonstrating the valuable insights gained from this thorough analytical exploration.

Section 5.1 sets the foundation by discussing the motivation behind the need for multiclass fairness metrics and providing the context and justification for the advancements proposed in our study. Section 5.2 details our methodology, including an explanation of the baseline binary fairness metrics on which our proposed metrics are based. Additionally, this section describes the Graph Neural Network (GNN) models and datasets utilized in our evaluations. In Section 5.3, we introduce our proposed multigroup and multiclass fairness metrics. This core section elaborates on the extensions we made to

traditional fairness metrics to better address the complexities found in these scenarios. Section 5.4 focuses on the experimental assessment of our new fairness metrics, providing empirical evidence of their effectiveness in identifying disparities in multiclass and multigroup settings. The chapter concludes with Section 5.5, which summarizes the major findings and contributions of our research.

5.1 Motivation

In light of the widespread use of automated decision-making systems across various domains, there is a growing recognition that the design and implementation of these models and their outcomes should align with a set of ethical standards (e.g., those prescribed by the *EU AI Act*[1]). This trend has led to a greater emphasis on investigation into topics such as transparency [29, 350], privacy [163, 276], sustainability [242, 314], and social equity [135]. As already discussed in Chapter 4, **algorithmic fairness** (Section 2.2) has been the focus of considerable attention in academic research and industry projects, primarily driven by the increased awareness of the potential harms that unfair artificial intelligence (AI) systems may present to specific social groups. On the one hand, several research studies have been conducted to explore potential sources of bias in automated systems (see Section 2.2.3). The two main categories for these sources are usually *biased data* and *algorithms* that can be influenced by the biases present in the training datasets. The existence of discrimination has practical and legal implications for organizations that rely on automated systems to make significant decisions [63]. Thus, it is crucial to quantitatively measure bias and fairness in machine learning (ML) models in a *responsible* way, referring to the principles of Responsible AI (see Section 2.1.2), which focus on accountability, reliability, and transparency in the development of such systems, and are correlated to similar notions expressed in Human-Centered AI (Section 2.1).

Typically, when we mention "bias", we are referring to a situation where an ML model shows a systematic (i.e., repeated over time) preference for one class over another. This means that the model has a significantly lower error rate for one category compared to another. More specifically, in our study, we deal with *group fairness*, which focuses on the results and effects of advantaged and disadvantaged groups. Group fairness, in its broadest sense, divides a population into groups based on protected characteristics and seeks to ensure fairness among these groups. Nevertheless, following our previous study illustrated in Chapter 4, regardless of the specific notion of fairness adopted, the existing literature clearly shows a gap between the strategies and methods for binary and multiclass scenarios.

There have been only a few contributions addressing this issue in recent years. Blakeney et al. [41] introduced two metrics, namely *Combined Error Variance* (CEV) and *Symmetric Distance Error* (SDE), to compute the biases of each sensitive group while comparing two different models. CEV assesses how likely a deep neural network is to reduce performance on one class in favor of others, while SDE calculates the disparities among the classes to be chosen more or less often based on the abundance of their

[1] `https://artificialintelligenceact.eu/the-act/`. Accessed March 30, 2025.

training examples. Putzel and Lee [277] addressed the issue of altering the results of a black-box classifier by extending the post-processing method suggested by Hardt et al. [141] in order to generate fair predictions for the examined model. Denis et al. [90] expanded the existing definition of *demographic parity* [111] to apply to multi-class classification scenarios, addressing both exact and approximate fairness cases. Additionally, they presented optimal solutions for the classifier in both situations. Alghamdi et al. [14] aim to develop unbiased probabilistic classifiers for multi-class classification tasks. Their proposed method involves transforming a biased pre-trained classifier into a set of models that satisfy specific fairness criteria for target groups. The resulting transformed model is obtained by adjusting the pre-trained classifier's outputs using a multiplicative factor. Furthermore, the authors have introduced an iterative algorithm that can be parallelized to compute the transformed classifier and ensure both sample complexity and convergence guarantees.

The main limitation identified in previous research, which we aim to address, is the lack of a thorough investigation into the effects of using binary fairness metrics in actual, real-life situations. In most cases, the primary goal of the proposed procedures and methods is to tackle the mathematical aspect of binary categorization to detect and mitigate bias.

5.2 Methodology

In this section, we provide the preliminary methodology adopted in the study presented in the chapter. The employed fairness metrics and GNN-based models will be shortly described below as they are the same utilized for the previous investigation, so they were already described in detail in Chapter 4. The datasets will be illustrated in a specific section.

Standard fairness metrics These metrics constitute the foundation for our extended multiclass and multigroup measures. As for the previous study, we consider four algorithmic fairness metrics belonging to *disparate impact* and *disparate mistreatment* categories, described in Section 2.2.4. The specific metrics are **statistical parity** (SP, Equation (2.2)), **equal opportunity** (EO, Equation (2.4)), **overall accuracy equality** (OAE, Equation (2.6)), and **treatment equality** (TE, Equation (2.8)).

Adopted GNN models The basis of our fairness evaluation relies on two modern GNN-based models, which currently stand among the most impactful developments in user modeling. These are **CatGCN** [68] and **RHGN** [372], both described in Section 4.2.1.

5.2.1 Datasets

The evaluation of the study presented in this chapter involved the adoption of four real-world datasets, i.e., Alibaba, JD, Pokec, and NBA. The description of the first

two datasets has already been provided in Section 4.2.2; here, we only discuss the specific attribute selection and processing for the related experimental setting.

For **ALIBABA**, we chose the product types as the categorical features associated with the users for the same model, following the experimental setup of CATGCN. Therefore, we considered only the items clicked by at least two users to establish the *co-click* relationship, which was used as the model's local interaction. To ensure consistency across the analyzed GNNs, we applied the same filtering process to RHGN before creating the heterogeneous graph. We selected the user's *consumption grade* (referred to as *buy*) as the target class for the user modeling task and the user's *age* as the sensitive attribute. For the binary scenario, we created the *bin-buy* variable from the original 3-level *buy* attribute by combining the *mid* ($y = 1$) and *high* ($y = 2$) levels. Additionally, we generated the *bin-age* variable from the 7-level *age* attribute by merging labels as follows: $s_A = \{s_0, s_1, s_2, s_3\}$ and $s_B = \{s_4, s_5, s_6\}$. In the ALIBABA dataset, the age range for each category is not indicated and is solely identified by a label. Two different binarizations have been utilized to establish a distinct division between the two groups.

JD leverages a *co-click* relationship as CATGCN's local interaction, as for ALIBABA. In order to ensure consistency across various experiments, we create a variable called "expense level" and employ it as the target class for profiling tasks. We utilize the pre-existing *purchase* connection between user and item nodes, as well as the number of purchased items (i.e., *count*) and the individual prices (i.e., *price*), to calculate the total expense of a user. After eliminating duplicate values, the expense list was segmented into four quartiles to determine the thresholds for establishing a variable with four levels. The variable *bin-exp* was created by separating the *low* level ($y = 0$) and combining the remaining ones, in accordance with the method used in the ALIBABA dataset. The sensitive attribute is the *age* variable with 5 levels, and we binarized it (*bin-age*) by categorizing users as under or over 35 years old. The resulting binary sensitive attribute is formed by the following groups: $s_A = \{s_0, s_1\}$ and $s_B = \{s_2, s_3, s_4\}$.

POKEC stands as the most widely used social network in Slovakia, closely resembling Facebook and the former Twitter, X. This dataset[2] has already been used in other relevant works (e.g., [82, 212, 232]). The dataset consists of anonymized information from the complete social network in 2012 and was made available by Takac and Zabovsky [322]. The input graph's nodes are homogeneous and stand for platform users, each with distinct characteristics such as gender, age, hobbies, interests, and profession. The connections between users are depicted as edges in a *follow* relationship, with no assigned weights. Concerning the user modeling task, we use the *working field* as the class to be predicted. The attribute's categories are only distinguished by names and do not have comprehensive explanations. Due to this, the binarization procedure, which creates the *bin-work-field* variable, is carried out by separating the most common category ($y = 0$) to create an

[2]Original dataset: `https://snap.stanford.edu/data/soc-pokec.html`.
Version adopted in our study: `https://github.com/EnyanDai/FairGNN/tree/main/dataset/pokec`. Accessed March 30, 2025.

Table 5.1. Characteristics of the used datasets.

Dataset	Users	Items	Edges	Features
Alibaba	166 958	64 553	427 464	2 820
JD	38 322	49 634	315 970	2 056
Pokec	13 504	-	882 765	70
NBA	403	-	16 570	178

evident distinction between the two groups, ensuring that the groups are clearly defined. The *age* is the sensitive attribute used in this case. In this dataset, every node contains a specific age of users. To create a set of levels that are meaningful and almost balanced, we adopted the following ranges to form five groups: under 18, 18-23, 24-28, 28-35, and over 35. As this is a social network, for the binary scenario, we chose to categorize users as either under or over 18 years old when creating the *bin-age* attribute. In particular, the generated groups are: $s_A = \{s_0\}$ and $s_B = \{s_1, s_2, s_3, s_4\}$.

The **NBA** dataset[3] utilized in our study is the extension of a Kaggle dataset[4] composed of information about ca. 400 NBA basketball players already employed in other contributions (e.g., [82, 212]). The 2016-2017 season's performance data for players, along with additional details such as nationality, age, and salary, form the characteristics of the homogeneous input graph nodes. The connections between players on a social network (i.e., Twitter, accessed through the official crawling API) are defined as unweighted graph edges. The user modeling task uses the three-level *salary* attribute from the dataset as the target class. It is transformed into a binary scenario by isolating the top-level class ($y = 2$) to create the *bin-salary* attribute. We used the *age* attribute as the sensitive attribute. We initially grouped individual values into three categories based on meaningful criteria for basketball players (i.e., under 25, 25-30, and over 30). Subsequently, we combined the two highest groups to form the *bin-age* variable using the following division: $s_A = \{s_0\}$ and $s_B = \{s_1, s_2\}$.

Information about the four datasets is presented in Table 5.1, with *features* denoting the size of CatGCN's input categorical feature array dimension. Tables 5.2 and 5.3 show the distribution of the target classes and sensitive attribute groups within the datasets in both the original and binarized scenarios.

In our research, we converted the target classes and sensitive attributes into binary form in a way to conduct a thorough analysis of class distributions. The applied method allowed us to investigate both balanced and unbalanced distributions of positive and negative classes. By using this binarization approach, we guaranteed that our investigation could effectively assess the impact of different class distributions on the results, improving the reliability and generalizability of our findings.

[3] `https://github.com/EnyanDai/FairGNN/tree/main/dataset/NBA`. Accessed March 30, 2025.
[4] `https://www.kaggle.com/noahgift/social-power-nba`. Accessed March 30, 2025.

Table 5.2. Distribution of the *original* target classes and sensitive attribute groups of the adopted datasets.

Dataset	Label	% Class/Group						
		0	*1*	*2*	*3*	*4*	*5*	*6*
ALIBABA	buy	32.48%	60.30%	7.22%	-	-	-	-
	age	21.74%	1.61%	17.56%	23.72%	30.83%	4.53%	0.01%
JD	expense	40.99%	15.68%	23.97%	19.36%	-	-	-
	age	23.59%	7.53%	50.17%	16.95%	1.76%	-	-
POKEC	work-field	47.67%	21.10%	13.12%	12.41%	5.70%	-	-
	age	38.85%	30.10%	13.64%	9.98%	7.43%	-	-
NBA	salary	22.33%	38.21%	39.45%	-	-	-	-
	age	39.95%	37.37%	22.58%	-	-	-	-

Table 5.3. Distribution of the *binarized* target classes and sensitive attribute groups of the adopted datasets.

Dataset	Label	% Class/Group	
		0	*1*
ALIBABA	bin-buy	32.48%	67.52%
	bin-age	64.63%	35.37%
JD	bin-exp	40.99%	59.01%
	bin-age	67.12%	32.88%
POKEC	bin-work-field	47.67%	52.33%
	bin-age	38.85%	61.15%
NBA	bin-salary	60.55%	39.45%
	bin-age	60.05%	39.95%

5.3 Multiclass and Multigroup Fairness Metrics

This section outlines the driving factors and the progression that resulted in the definition of **multigroup fairness metrics**, leading to the broader **multiclass and multigroup metrics**.

In Section 4.1, we explained that one of the main reasons for the widespread use of binary fairness metrics is that many AI applications with ethical implications are inherently binary, such as deciding whether to hire or not to hire. However, this rationale doesn't hold true when sensitive attributes are taken into account because it is widely acknowledged that almost no human characteristics can truly be categorized as binary, including gender and especially age.

For the **multigroup fairness metrics**, as for the standard definition of binary metrics (see Section 2.2.5), we consider $y \in \{0, 1\}$ as the binary target label and $\hat{y} \in \{0, 1\}$

as the model prediction. Let N represent the count of sensitive attribute groups s. The equations below are defined to ensure that the resulting score is uniform across groups, thereby meeting the prescribed fairness criteria.

Multigroup statistical parity

$$P(\hat{y} = 1 \mid s = n), \forall n \in \{0, ..., N-1\} \tag{5.1}$$

Multigroup equal opportunity

$$P(\hat{y} = 1 \mid y = 1, s = n), \forall n \in \{0, ..., N-1\} \tag{5.2}$$

Multigroup overall accuracy equality

$$P(\hat{y} = 0 \mid y = 0, s = n) + P(\hat{y} = 1 \mid y = 1, s = n), \forall n \in \{0, ..., N-1\} \tag{5.3}$$

Multigroup treatment equality

$$\frac{P(\hat{y} = 1 \mid y = 0, s = n)}{P(\hat{y} = 0 \mid y = 1, s = n)}, \forall n \in \{0, ..., N-1\} \tag{5.4}$$

The second reason supporting the definition of binary fairness metrics provided by Caton and Haas [63] is linked to the simplicity of mathematically quantifying a binary variable as opposed to a variable with multiple values. Building upon the aforementioned definitions for multiple groups, we suggest a further expansion of **multiclass and multigroup fairness metrics** without introducing any additional mathematical complexity. This aims to demonstrate that a straightforward generalization can result in a more comprehensive and profound analysis of fairness.

Let M and N denote the number of classes y, $\hat{y}$ and sensitive groups s, correspondingly. All the metrics shown below should have the same value for each class and group.

Multiclass and multigroup statistical parity

$$P(\hat{y} = m \mid s = n), \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{5.5}$$

Multiclass and multigroup equal opportunity

$$P(\hat{y} = m \mid y = m, s = n), \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{5.6}$$

Multiclass and multigroup overall accuracy equality

$$\sum_{m=0}^{M-1} P(\hat{y} = m \mid y = m, s = n), \forall n \in \{0, ..., N-1\} \tag{5.7}$$

Multiclass and multigroup treatment equality

$$\frac{P(\hat{y} = m \mid y \neq m, s = n)}{P(\hat{y} \neq m \mid y = m, s = n)}, \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{5.8}$$

It is important to note that, in the case of multiclass and multigroup scenarios, the definition of *equal opportunity* in Equation (5.6) would also be applicable to the expansion of the *equalized odds* metric [141] in that same setting.

5.4 Experimental Fairness Assessment

In this section, we illustrate the empirical investigation carried out to evaluate the impact of the proposed multiclass and multigroup fairness metrics compared to typical binary measurements. We aim to address the following research questions:

RQ1 To what extent can multigroup fairness metrics impact a model's fairness evaluation with respect to the related binary metrics?

RQ2 To what extent can multiclass and multigroup fairness metrics improve bias detection and future mitigation in real-world cases?

Below, we outline the experiments conducted to address these research questions and the related parameters. The experimental findings serve as the conclusion for this section.

5.4.1 Experimental setting

To ensure that the GNN-based models are able to effectively handle user modeling tasks using the specified dataset, target classes, and sensitive attributes, we implement a process for selecting hyperparameters as outlined below. For CATGCN, the *learning rate* is tuned among {0.001, 0.01, 0.1}, the L_2 regularization coefficient and the *dropout ratio* are searched in {1e-5, 1e-4} and {0.1, 0.3, 0.5, 0.7}, respectively, and the aggregation parameter α is explored within {0.1, 0.3, 0.5, 0.7, 0.9}. For RHGN, the *learning rate* and the L_2 regularization coefficient are tuned among {0.01, 0.1} and {1e-5, 1e-4}, respectively; the hidden dimension of the two layers of the entity-level aggregation network is searched in {32, 64}, while the number of heads in multi-head attention is explored within {1, 2}. All other parameters are set to default values from the original papers. After performing the grid search, we conducted the experiments 40 times for each fairness metric. The specified operations were performed using an Nvidia Quadro RTX 8000 48GB GPU, and the source code is publicly available[5].

5.4.2 Experimental results

This section analyzes the outcomes and findings of the practical studies conducted for each research question. Prior to examining the fairness evaluation, Table 5.4 presents

[5] `https://github.com/erasmopurif/toward-responsible-fairness-analysis/`.

Table 5.4. Experiment results of the user modeling tasks for each combination of dataset, model, and setting (binary or multiclass). The ↑ symbol means that higher scores are better.

Dataset	Model	Performance (binary)		Performance (multiclass)	
		Accuracy (↑)	*F1-score* (↑)	*Accuracy* (↑)	*F1-score* (↑)
Alibaba	CatGCN	0.776 ±0.021	0.718 ±0.005	0.535 ±0.031	0.501 ±0.012
	RHGN	0.803 ±0.006	0.711 ±0.016	0.618 ±0.002	0.587 ±0.018
JD	CatGCN	0.732 ±0.008	0.706 ±0.006	0.502 ±0.002	0.498 ±0.013
	RHGN	0.738 ±0.004	0.702 ±0.007	0.575 ±0.010	0.525 ±0.017
Pokec	CatGCN	0.808 ±0.002	0.797 ±0.002	0.445 ±0.004	0.398 ±0.006
	RHGN	0.799 ±0.022	0.779 ±0.013	0.455 ±0.004	0.404 ±0.003
NBA	CatGCN	0.743 ±0.074	0.709 ±0.052	0.593 ±0.067	0.541 ±0.072
	RHGN	0.768 ±0.043	0.721 ±0.071	0.581 ±0.051	0.527 ±0.035

the results of the user modeling task, displaying the performance scores in terms of *accuracy* and *F1-score* for each combination of dataset, model, and setting (i.e., binary or multiclass). The purpose of showing this table is to enhance comprehension of the effectiveness of the selected models and to emphasize the significance of the presented fairness metrics in their various forms.

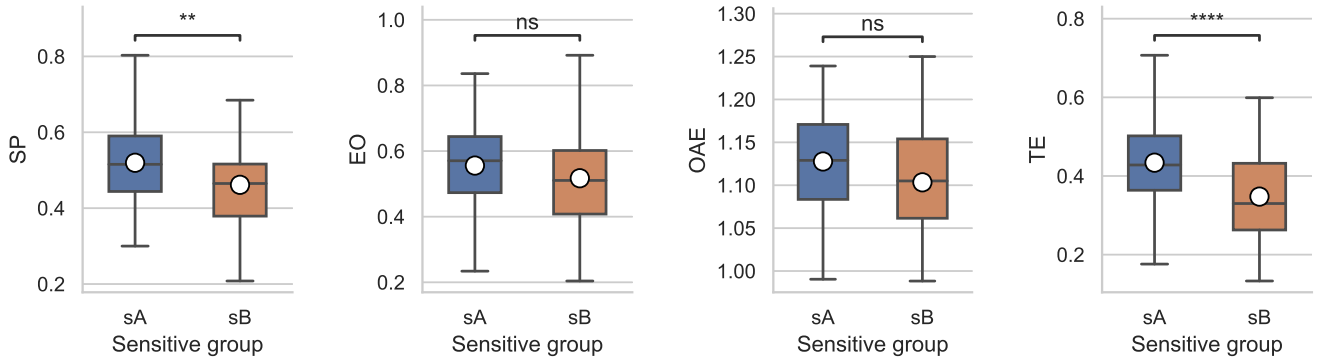
Comparing multigroup and binary fairness evaluation (RQ1)

This study examines the benefits of using multigroup metrics instead of binary metrics to accurately assess fairness. The focus is on a binary target class, and we compare the analysis of binary and multigroup sensitive attributes. For both GNN models, CatGCN and RHGN, we initially run the user modeling task (i.e., classification of *bin-buy* class for Alibaba, *bin-exp* for JD, *bin-work-field* for Pokec, and *bin-salary* for NBA), then we computed the scores of the binary fairness metrics, defined by Equations (2.2), (2.4), (2.6) and (2.8), and multigroup fairness metrics, defined by Equations (5.1) to (5.4).

To enhance the reliability of our results and verify that the differences observed are not due to chance, we used a *Mann-Whitney-Wilcoxon test*[6] [227, 358] for each pair of groups. We conducted the statistical test with 1000 iterations to ensure consistency and reproducibility of our findings. Further, we employed the *Bonferroni correction*[7] [146], a conservative statistical method designed to address the issue of multiple comparisons.

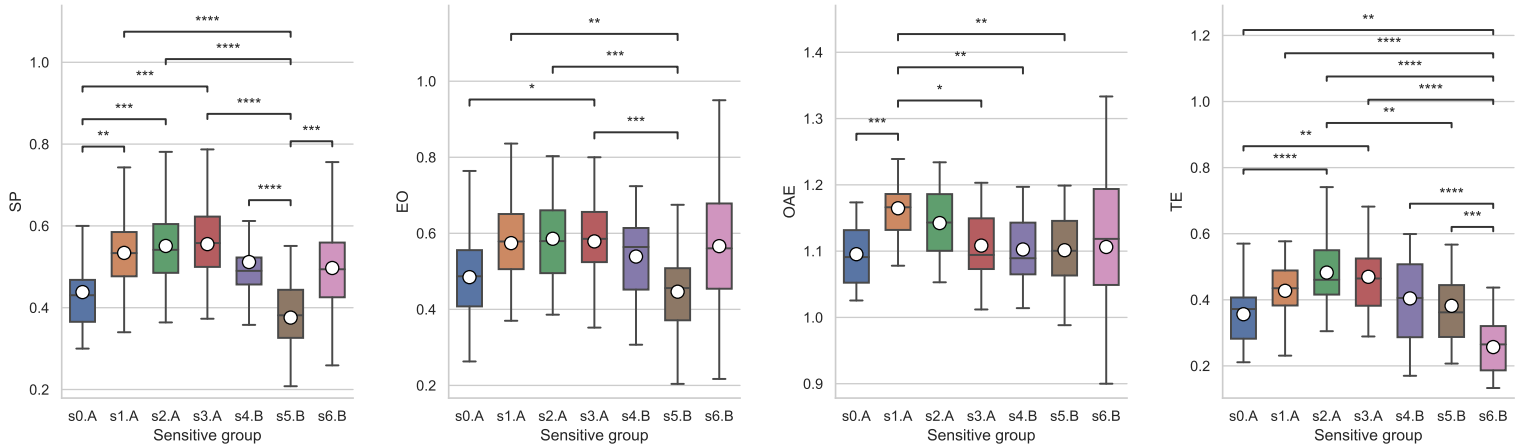
[6]The Mann-Whitney-Wilcoxon test is utilized to evaluate if two independent samples originate from the same distribution and is a non-parametric test. It is different from the t-test as it does not necessitate the assumption of normal distribution, thus making it a more adaptable and dependable option for our data's distribution traits.

[7]The Bonferroni correction is used to modify the significance thresholds to address the higher likelihood of encountering significant outcomes by chance when conducting multiple tests at the same time. To achieve this, the desired significance level is divided by the number of comparisons made, helping to reduce the possibility of false positives and ensuring that the reported differences are genuinely statistically significant rather than simply a product of random variation or the large number of tests conducted.



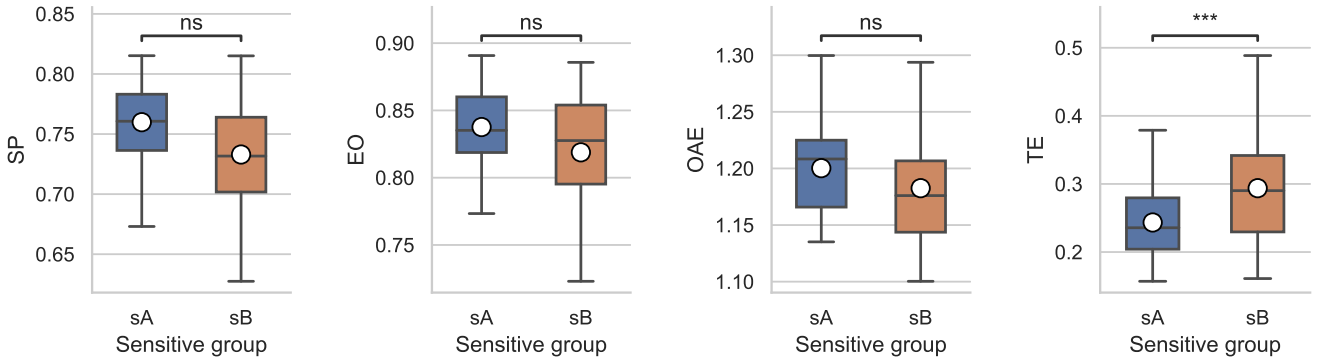
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.1. Fairness assessment of **CatGCN** on **Alibaba** dataset in binary class (positive output) and binary group scenario.



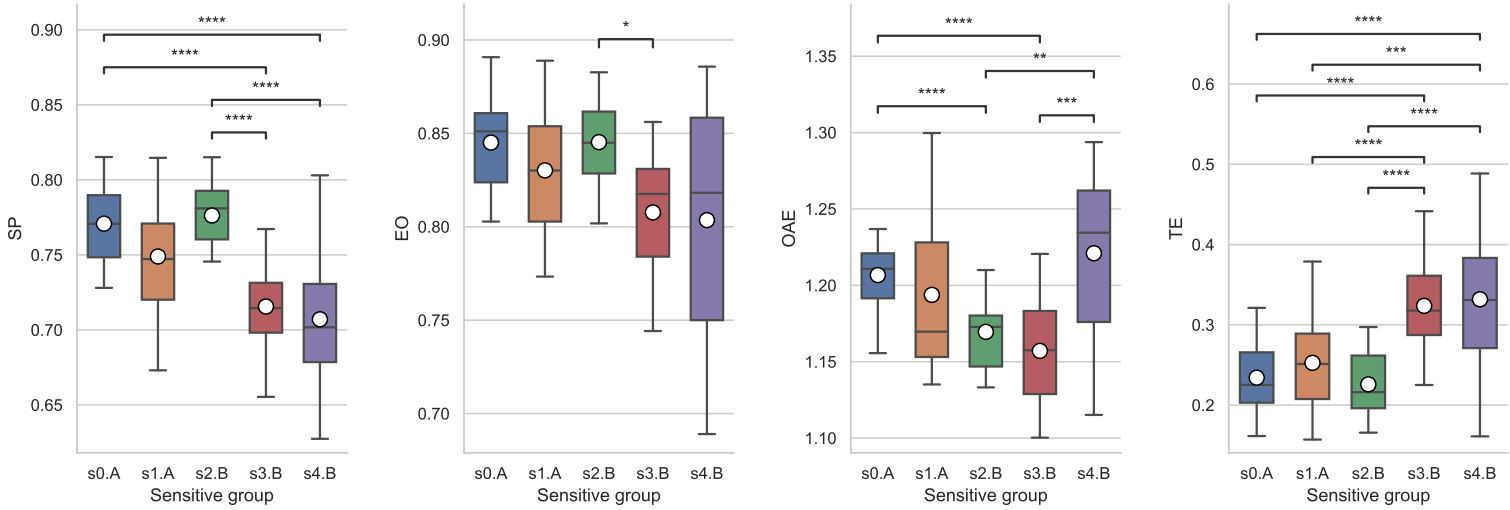
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.2. Fairness assessment of **CatGCN** on **Alibaba** dataset in the binary class (pos. output) and multigroup scenario.



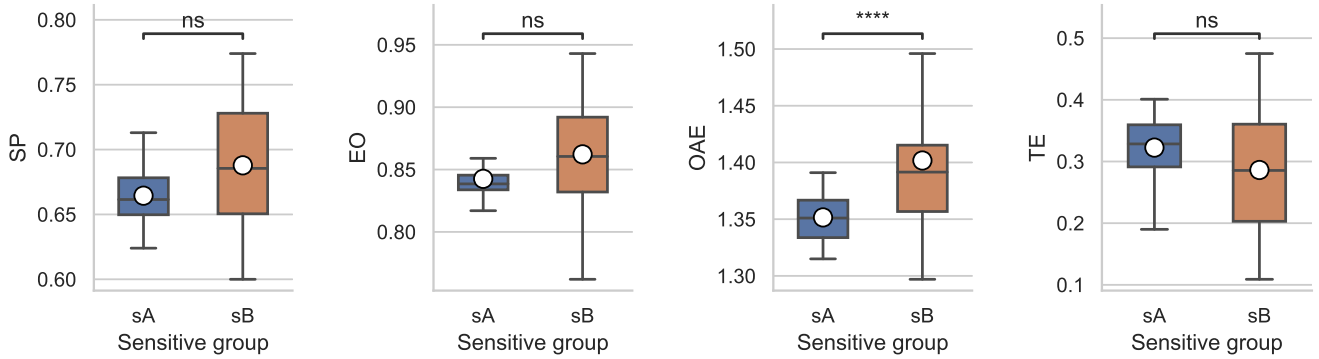
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.3. Fairness assessment of **CatGCN** on **JD** dataset in the binary class (positive output) and binary group scenario.



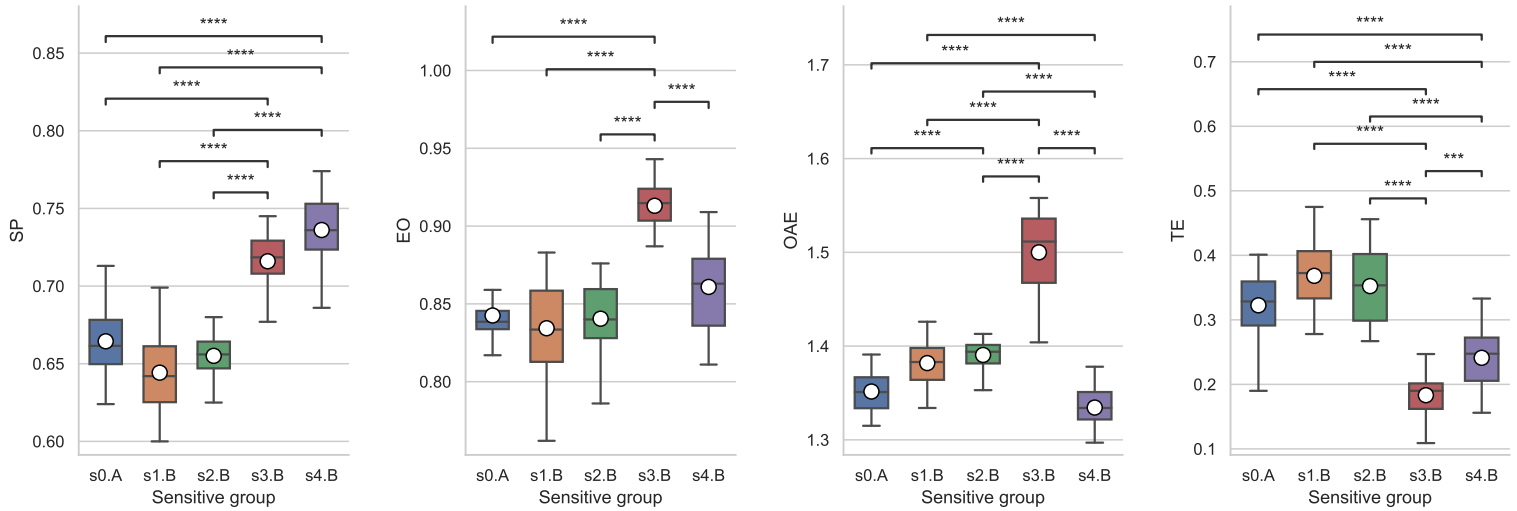
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.4. Fairness assessment of **CatGCN** on **JD** dataset in the binary class (positive output) and multigroup scenario.



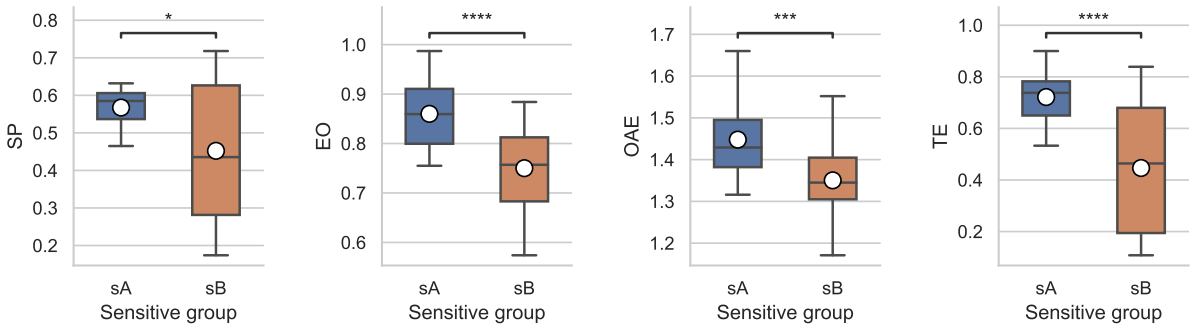
(a) SP (b) EO (c) OAE (d) TE

Figure 5.5. Fairness assessment of **CatGCN** on **Pokec** dataset in the binary class (pos. output) and binary group scenario.



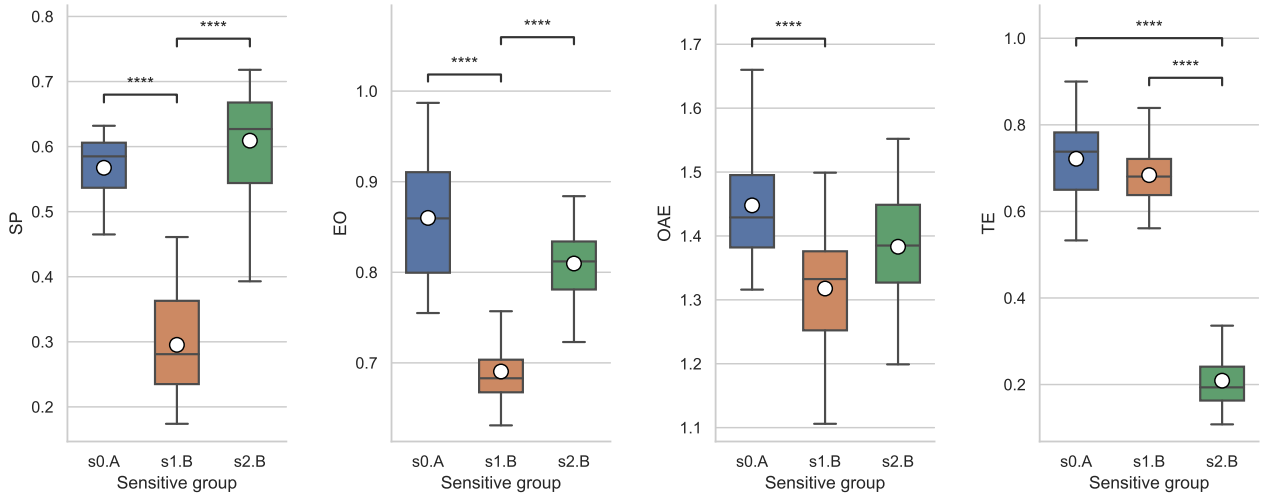
(a) SP (b) EO (c) OAE (d) TE

Figure 5.6. Fairness assessment of **CatGCN** on **Pokec** dataset in the binary class (positive output) and multigroup scenario.



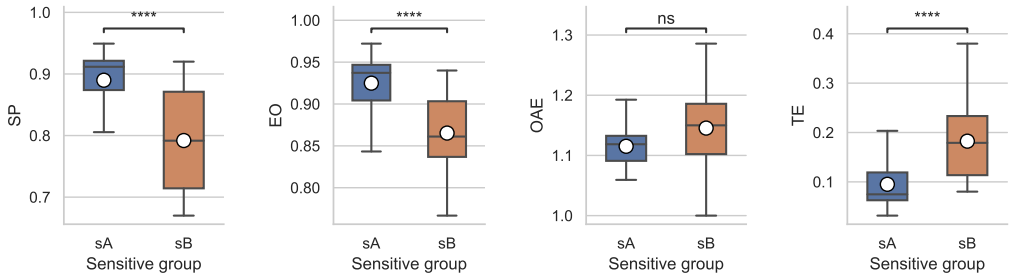
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.7. Fairness assessment of **CatGCN** on **NBA** dataset in the binary class (positive output) and binary group scenario.



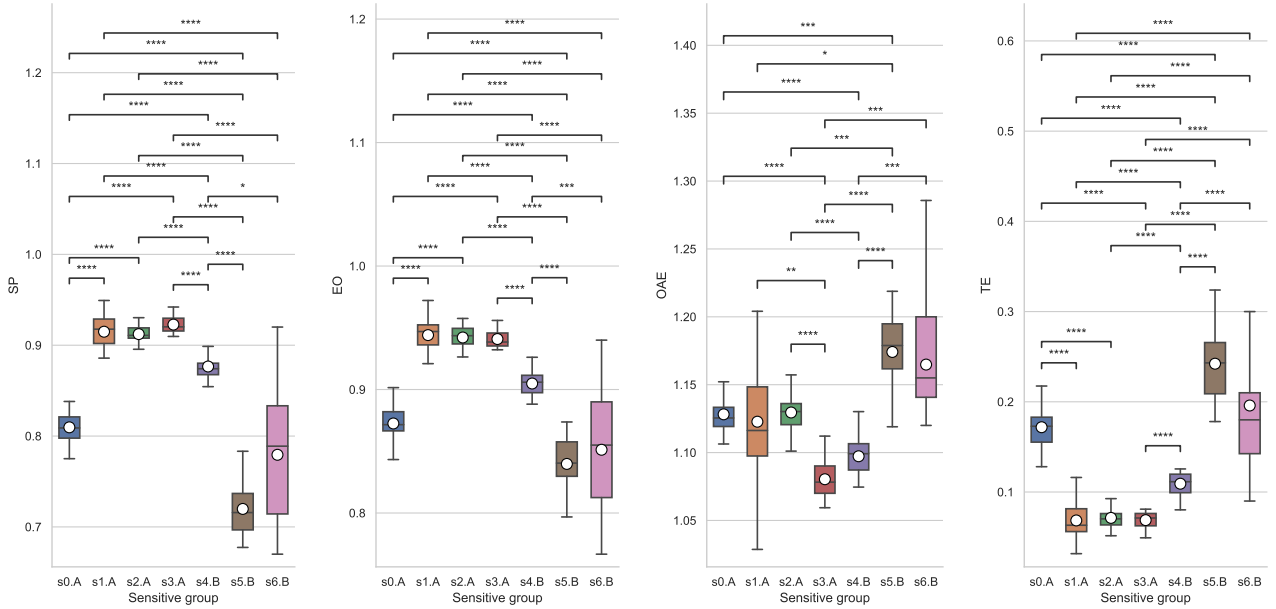
(a) SP **(b)** EO **(c)** OAE **(d)** TE

Figure 5.8. Fairness assessment of **CatGCN** on **NBA** dataset in the binary class (positive output) and multigroup scenario.



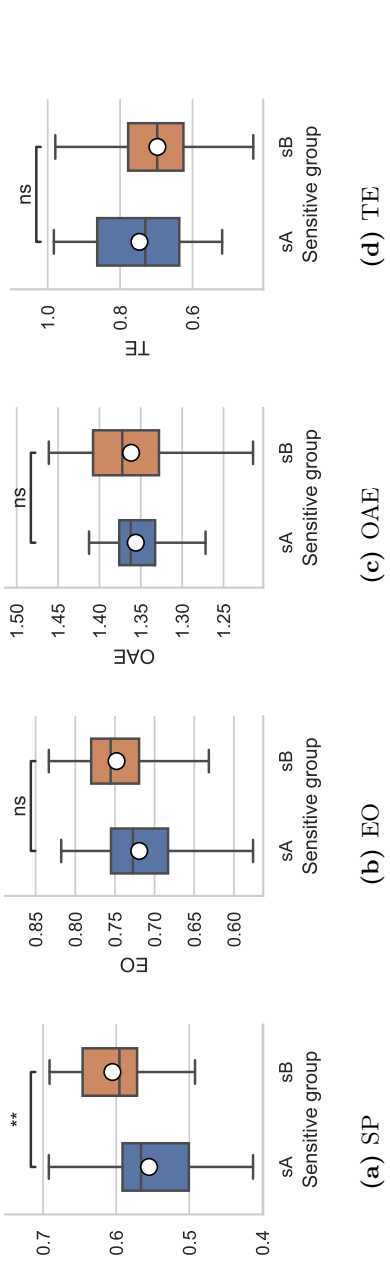
(a) SP (b) EO (c) OAE (d) TE

Figure 5.9. Fairness assessment of **RHGN** on **Alibaba** dataset in the binary class (positive output) and binary group scenario.



(a) SP (b) EO (c) OAE (d) TE

Figure 5.10. Fairness assessment of **RHGN** on **Alibaba** dataset in the binary class (positive output) and multigroup scenario.

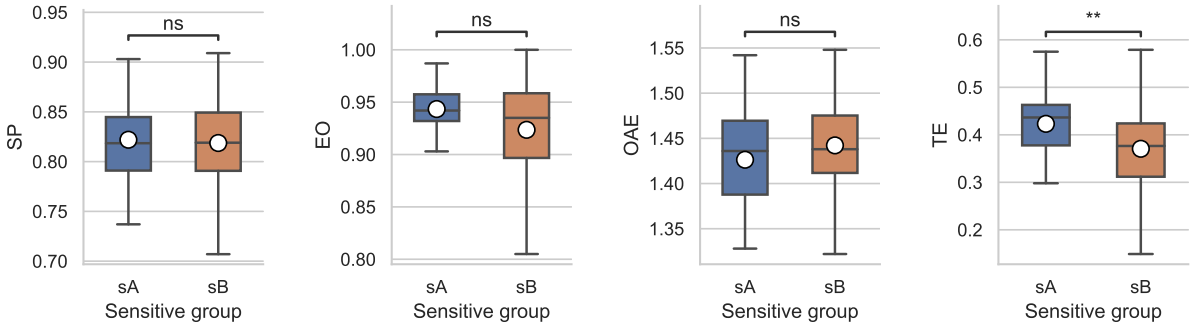


(a) SP (b) EO (c) OAE (d) TE

Figure 5.11. Fairness assessment of **RHGN** on **JD** dataset in the binary class (positive output) and binary group scenario.

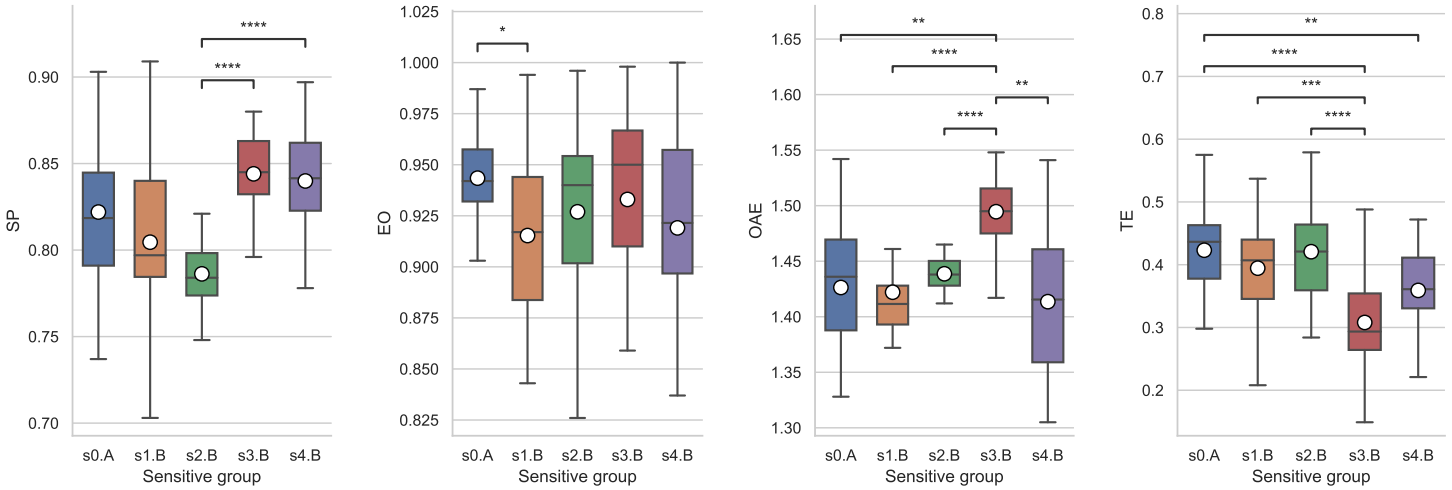
(a) SP (b) EO (c) OAE (d) TE

Figure 5.12. Fairness assessment of **RHGN** on **JD** dataset in the binary class (positive output) and multigroup scenario.



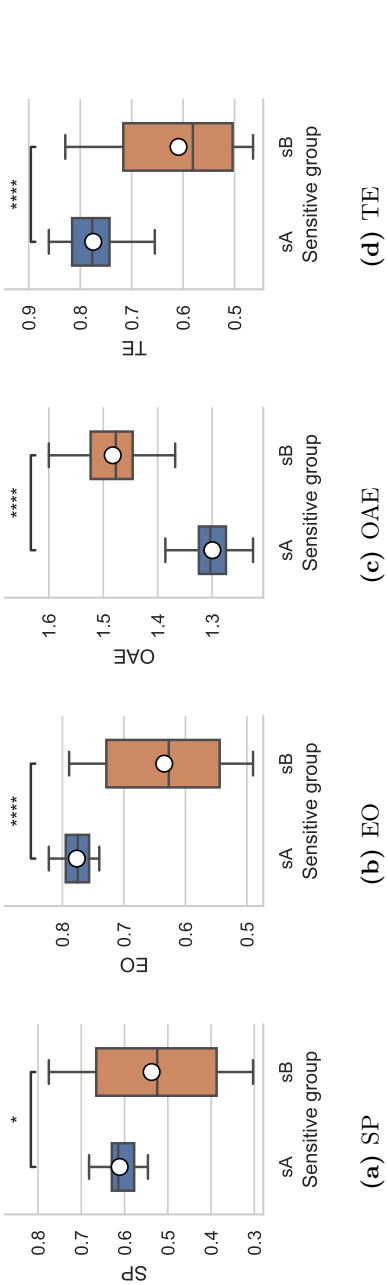
(a) SP (b) EO (c) OAE (d) TE

Figure 5.13. Fairness assessment of **RHGN** on **Pokec** dataset in the binary class (positive output) and binary group scenario.



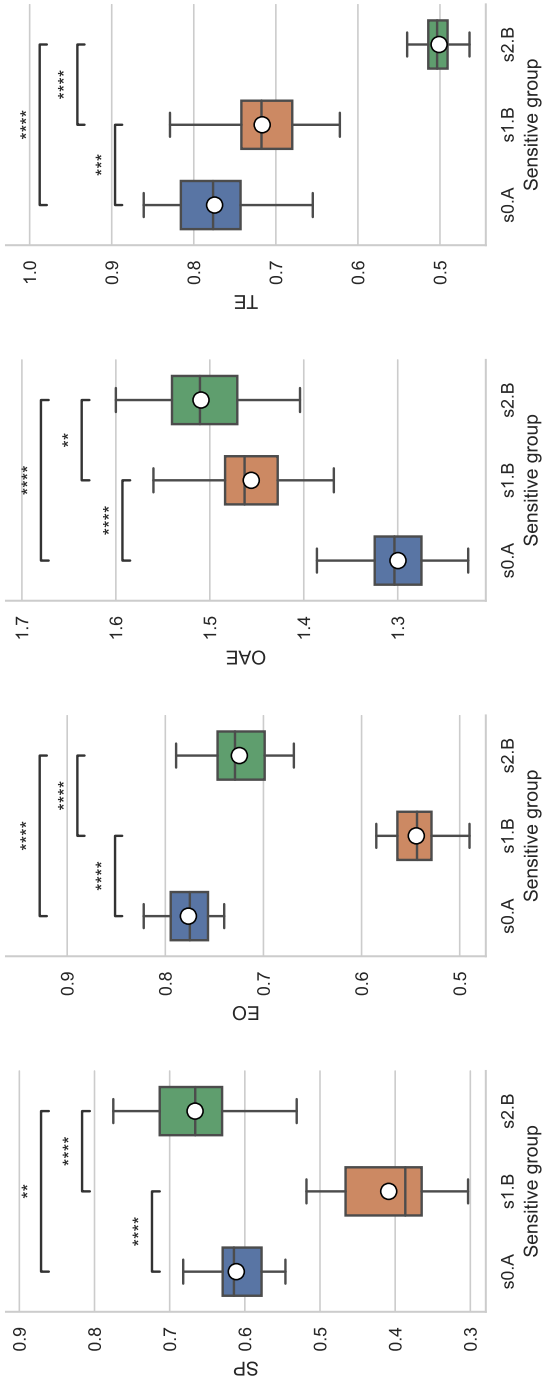
(a) SP (b) EO (c) OAE (d) TE

Figure 5.14. Fairness assessment of **RHGN** on **Pokec** dataset in the binary class (positive output) and multigroup scenario.



(a) SP (b) EO (c) OAE (d) TE

Figure 5.15. Fairness assessment of **RHGN** on **NBA** dataset in the binary class (positive output) and binary group scenario.



(a) SP (b) EO (c) OAE (d) TE

Figure 5.16. Fairness assessment of **RHGN** on **NBA** dataset in the binary class (positive output) and multigroup scenario.

Table 5.5. Qualitative analysis of the comparative results between *binary* and *multigroup* scenarios leading to the considerations for **RQ1**. The *multigroup* column includes the differences from the binary case.

Dataset	**Metric**	**Model** / *Setting* (Ref. Figs.)					
		CatGCN			**RHGN**		
		binary	*multigroup*		*binary*	*multigroup*	
Alibaba	SP	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$, $s_{6.B}$ adv.	(5.1a-5.2a)	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$ adv.	(5.9a-5.10a)
	EO	fair	$s_{0.A}$, $s_{5.B}$ dis.	(5.1b-5.2b)	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$ adv.	(5.9b-5.10b)
	OAE	fair	$s_{1.A}$, $s_{2.A}$ adv.	(5.1c-5.2c)	fair	$s_{3.A}$, $s_{4.B}$ dis., $s_{5.B}$, $s_{6.B}$ adv.	(5.9c-5.10c)
	TE	s_A adv.	$s_{0.A}$ dis.	(5.1d-5.2d)	s_B adv.	$s_{4.B}$ dis.	(5.9d-5.10d)
JD	SP	fair	$s_{3.B}$, $s_{4.B}$ dis.	(5.3a-5.4a)	s_B adv.	$s_{3.B}$ dis.	(5.11a-5.12a)
	EO	fair	$s_{3.B}$ dis.	(5.3b-5.4b)	fair	$s_{0.A}$ dis.	(5.11b-5.12b)
	OAE	fair	$s_{2.B}$, $s_{3.B}$ dis.	(5.3c-5.4c)	fair	$s_{2.B}$ adv.	(5.11c-5.12c)
	TE	s_B adv.	$s_{2.B}$ dis.	(5.3d-5.4d)	fair	$s_{0.A}$ adv., $s_{1.A}$, $s_{2.B}$ dis.	(5.11d-5.12d)
Pokec	SP	fair	$s_{3.B}$, $s_{4.B}$ adv.	(5.5a-5.6a)	fair	$s_{2.B}$ dis.	(5.13a-5.14a)
	EO	fair	$s_{3.B}$ adv.	(5.5b-5.6b)	fair	*no diff.*	(5.13b-5.14b)
	OAE	s_B adv.	$s_{1.B}$, $s_{2.B}$, $s_{4.B}$ dis.	(5.5c-5.6c)	fair	$s_{3.B}$ adv.	(5.13c-5.14c)
	TE	fair	$s_{3.B}$, $s_{4.B}$ dis.	(5.5d-5.6d)	s_A adv.	$s_{1.B}$, $s_{2.B}$ adv.	(5.13d-5.14d)
NBA	SP	s_A adv.	$s_{2.B}$ adv.	(5.7a-5.8a)	s_A adv.	$s_{2.B}$ adv.	(5.15a-5.16a)
	EO	s_A adv.	$s_{2.B}$ adv.	(5.7b-5.8b)	s_A adv.	$s_{2.B}$ adv.	(5.15a-5.16a)
	OAE	s_A adv.	*no diff.*	(5.7c-5.8c)	s_B adv.	$s_{1.B}$ dis.	(5.15a-5.16a)
	TE	s_A adv.	$s_{1.B}$ adv.	(5.7d-5.8d)	s_A adv.	*no diff.*	(5.15a-5.16a)

In Table 5.5, we show a comprehensive qualitative analysis of the comparative results between binary and multigroup scenarios. Specifically, we present the differences from the related binary case for each combination of dataset, metric, model, and setting in the multigroup scenario.

The results of the experiments for every possible combination of the model, dataset, and scenario mentioned previously can be found in Figures 5.1 to 5.16. In the displayed charts, each pair of box plots is annotated with a symbol reflecting the statistical significance of the difference between the two compared groups based on the p-value, which measures the strength of the evidence against the null hypothesis. In particular:

- A non-significant difference, denoted by `[ns]` or simply no annotation, suggests that the evidence is not substantial enough to dismiss the null hypothesis for the difference between the groups. This implies that the observed difference could be attributed to random chance rather than systematic bias.

- Symbols from `[*]` to `[****]` represent increasing levels of statistical significance, linked to decreasing p-values. When there is a statistically significant difference,

Table 5.6. Description of the cases derived from the assessment of the comparative results between *binary* and *multigroup* scenarios.

#	**Binary scenario**	**Multigroup scenario**
1	Binarized group advantaged	Related fine-grained original groups (all or some) disadvantaged
2	Binarized group advantaged	Opposite (i.e., belonging to the other binarized group) fine-grained original groups (all or some) advantaged
3	Fair result	Some fine-grained groups particularly disadvantaged (or advantaged to the detriment of others)

it means the chance of the observed data happening under the null hypothesis is low. This distinction suggests genuine, consistent differences between the groups, and in our study, it indicates the existence of unfairness. The specific significance levels in question are:

`[*]` Significant difference with a p-value less than 0.05 but greater than 0.01.

`[**]` More significant difference with a p-value less than 0.01 but greater than 0.001.

`[***]` Highly significant difference with a p-value less than 0.001 but greater than 0.0001.

`[****]` Extremely significant difference with a p-value less than 0.0001.

Based on the analysis of the experiment results in this scenario, we have identified three pivotal cases, which are detailed in Table 5.6.

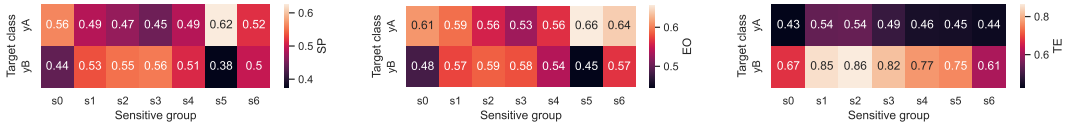
To illustrate the real-world applications of these results, we can analyze the tests conducted on the JD dataset using the CatGCN model. In these, the expense level is used as the classification target, and the sensitive attribute is age. These experiments correspond to Figures 5.3a and 5.4a and Figures 5.3d and 5.4d. In the first scenario, when working under *statistical parity* constraints, the binary age groups are fair, and no intervention is necessary. However, in a detailed analysis that considers multigroup, it becomes apparent that two specific age subgroups are at a disadvantage, and it may be appropriate to plan an intervention to address these inequities. When considering the second situation and looking at the scores for *treatment equality*, the findings indicate that one age group has an advantage in a binary comparison. However, upon conducting a more detailed examination across fine-grained groups, it becomes apparent that within that particular binary age group, there exists a specific subgroup that experiences a disadvantage. Should a bias mitigation process be applied in the binary scenario, it would exacerbate the discrimination toward this particular age subgroup.

Observation 1 *Using multigroup metrics for a detailed fairness analysis helps uncover real discrimination against sensitive groups that may not be apparent when using a binary evaluation. This is important in cases where biases are not easily identified or when an underprivileged group is mistakenly considered to be in a favorable position.*

Assessing multiclass and multigroup fairness metrics in real-world contexts (RQ2)

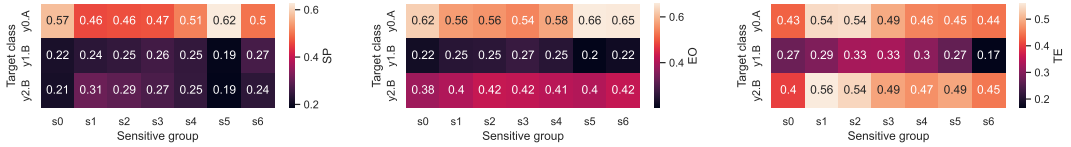
Most of the standard (binary) fairness metrics depend on selecting a *positive* class to calculate them. As discussed in this chapter, when transforming an originally multiclass target variable into binary, the choice of the positive class is almost arbitrary because neither class can be distinctly considered positive. In this study, our aim is to explore why it is preferable to use multiclass and multigroup fairness evaluation to understand all potential biases introduced by the models. We conduct the same profiling analysis outlined in the preceding section and calculate the metrics specified in Equations (5.5) to (5.8).

In Figures 5.17 to 5.32, we show the results related to **RQ2** for each combination of model, dataset, and metrics.



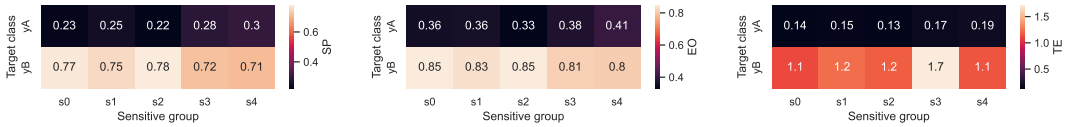
(a) SP **(b)** EO **(c)** TE

Figure 5.17. Fairness assessment of **CatGCN** on **Alibaba** dataset in the binary class (both outputs) and multigroup scenario.



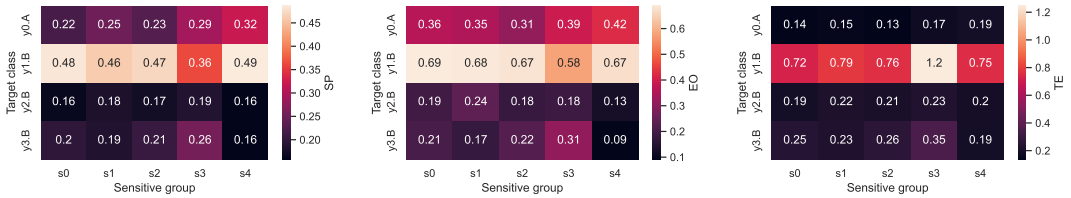
(a) SP **(b)** EO **(c)** TE

Figure 5.18. Fairness assessment of **CatGCN** on **Alibaba** dataset in the multiclass and multigroup scenario.



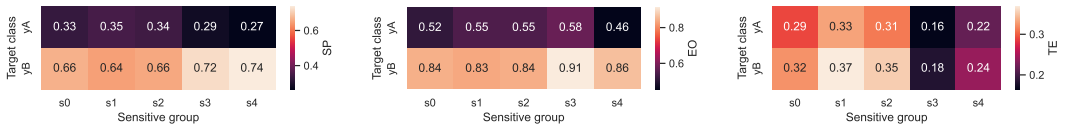
(a) SP **(b)** EO **(c)** TE

Figure 5.19. Fairness assessment of **CatGCN** on **JD** dataset in the binary class (both outputs) and multigroup scenario.



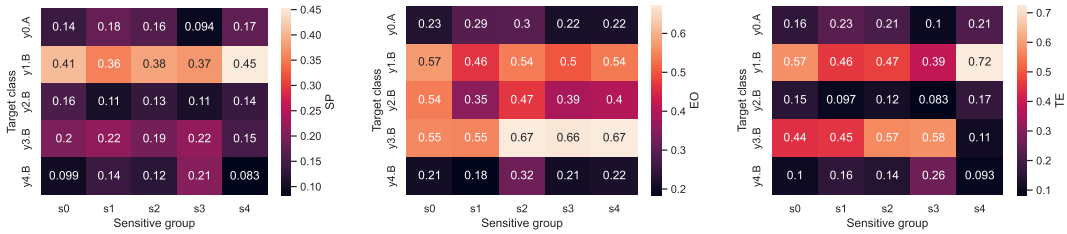
(a) SP **(b)** EO **(c)** TE

Figure 5.20. Fairness assessment of **CatGCN** on **JD** dataset in the multiclass and multigroup scenario.



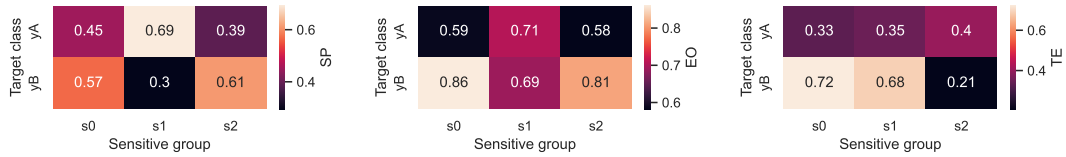
(a) SP **(b)** EO **(c)** TE

Figure 5.21. Fairness assessment of **CATGCN** on **POKEC** dataset in the binary class (both outputs) and multigroup scenario.



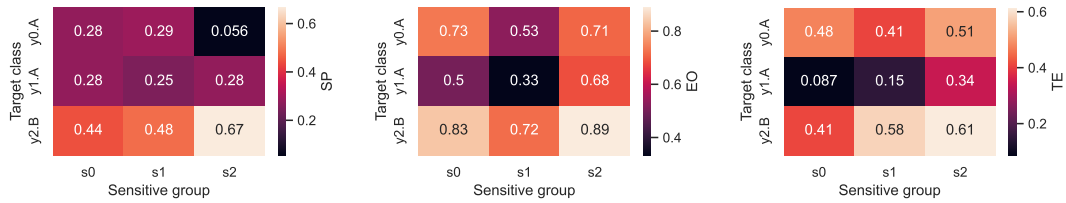
(a) SP **(b)** EO **(c)** TE

Figure 5.22. Fairness assessment of **CATGCN** on **POKEC** dataset in the multiclass and multigroup scenario.



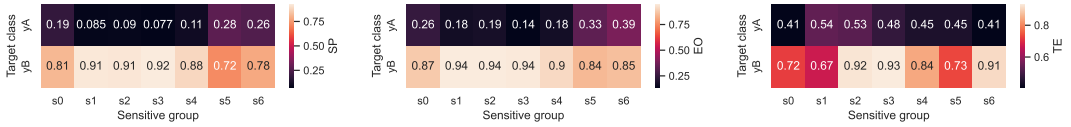
(a) SP **(b)** EO **(c)** TE

Figure 5.23. Fairness assessment of **CATGCN** on **NBA** dataset in the binary class (both outputs) and multigroup scenario.



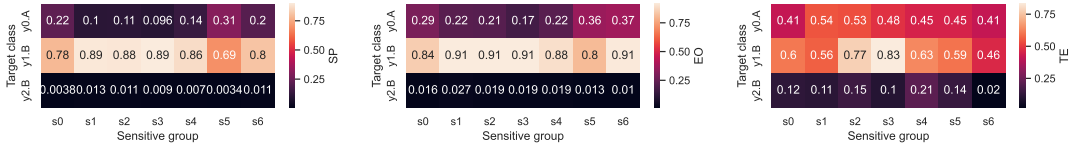
(a) SP **(b)** EO **(c)** TE

Figure 5.24. Fairness assessment of **CATGCN** on **NBA** dataset in the multiclass and multigroup scenario.



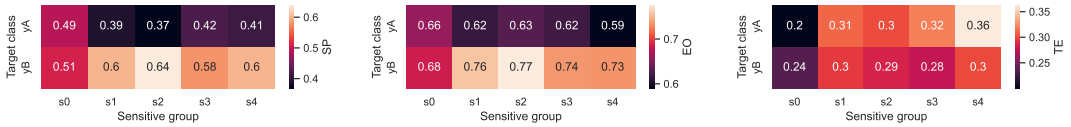
(a) SP **(b)** EO **(c)** TE

Figure 5.25. Fairness assessment of **RHGN** on **Alibaba** dataset in the binary class (both outputs) and multigroup scenario.



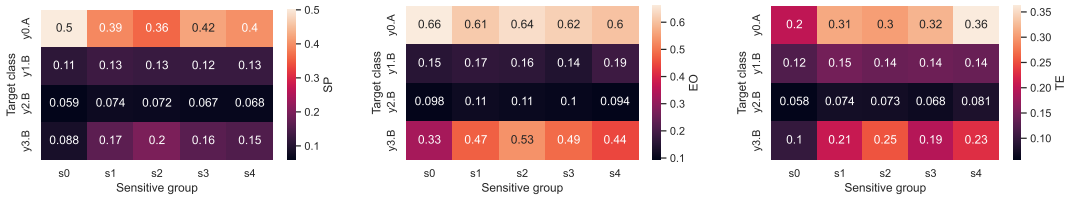
(a) SP **(b)** EO **(c)** TE

Figure 5.26. Fairness assessment of **RHGN** on **Alibaba** dataset in the multiclass and multigroup scenario.



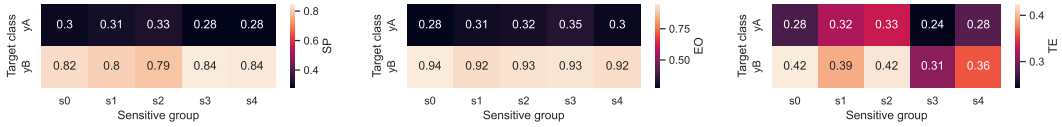
(a) SP **(b)** EO **(c)** TE

Figure 5.27. Fairness assessment of **RHGN** on **JD** dataset in the binary class (both outputs) and multigroup scenario.



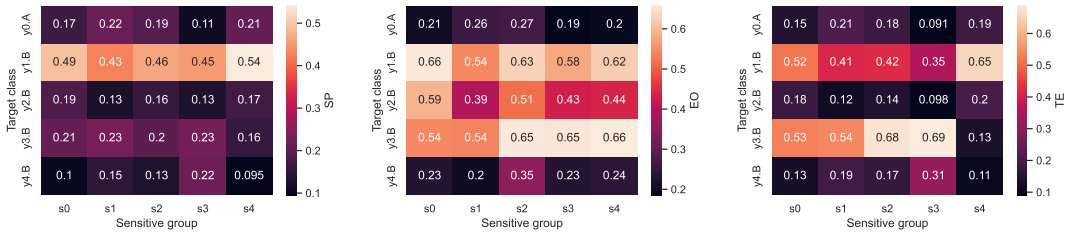
(a) SP **(b)** EO **(c)** TE

Figure 5.28. Fairness assessment of **RHGN** on **JD** dataset in the multiclass and multigroup scenario.



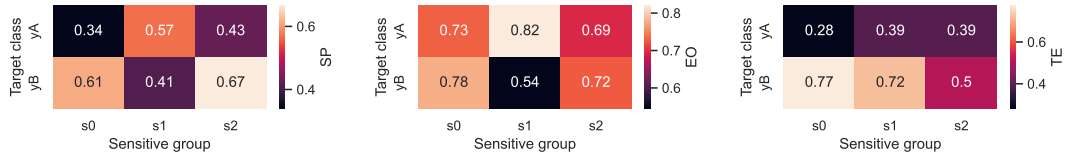
(a) SP **(b)** EO **(c)** TE

Figure 5.29. Fairness assessment of **RHGN** on **POKEC** dataset in the binary class (both outputs) and multigroup scenario.



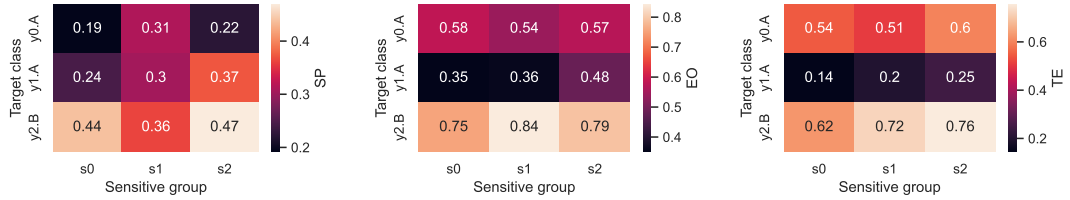
(a) SP **(b)** EO **(c)** TE

Figure 5.30. Fairness assessment of **RHGN** on **POKEC** dataset in the multiclass and multigroup scenario.



(a) SP **(b)** EO **(c)** TE

Figure 5.31. Fairness assessment of **RHGN** on **NBA** dataset in the binary class (both outputs) and multigroup scenario.



(a) SP **(b)** EO **(c)** TE

Figure 5.32. Fairness assessment of **RHGN** on **NBA** dataset in the multiclass and multigroup scenario.

Table 5.7. Qualitative analysis of the comparative results between *multigroup* and *multiclass* scenarios leading to the considerations for **RQ2**. The symbols refer to the derived cases described in Table 5.8.

Dataset	Metric	Model			
		CatGCN	*Ref. Figs.*	**RHGN**	*Ref. Figs.*
Alibaba	SP	†, ◇	(5.17a, 5.18a)	✻	(5.25a, 5.26a)
	EO	◇	(5.17b, 5.18b)	✻	(5.25b, 5.26b)
	TE	✻	(5.17c, 5.18c)	✻	(5.25c, 5.26c)
JD	SP	✻	(5.19a, 5.20a)	†, ◇	(5.27a, 5.28a)
	EO	✻	(5.19b, 5.20b)	†, ◇	(5.27b, 5.28b)
	TE	✻	(5.19c, 5.20c)	◇	(5.27c, 5.28c)
Pokec	SP	✻, ⩓	(5.21a, 5.22a)	✻	(5.29a, 5.30a)
	EO	✻, ⩓	(5.21b, 5.22b)	✻	(5.29b, 5.30b)
	TE	⊙	(5.21c, 5.22c)	✻, ⩓	(5.29c, 5.30c)
NBA	SP	†	(5.23a, 5.24a)	☆	(5.31a, 5.32a)
	EO	◇	(5.23b, 5.24b)	†, ◇	(5.31b, 5.32b)
	TE	†	(5.23c, 5.24c)	-	(5.31c, 5.32c)

Regarding the metrics, the evaluation does not include the multiclass and multigroup *OAE* (Equation (5.7)) because their results would have been the same as in the previous experiments due to their definition, which adds up the probabilities of the target classes. For clarity, only the average values are displayed in the result charts without additional statistics. The method of binarizing the single classes in the previous evaluation is explained in Section 5.2.1.

Equivalently to the binary-multigroup scenario (Section 5.4.2), we presented in Table 5.7 an effective method to visually display the results of the qualitative analysis comparing multigroup and multiclass scenarios. For every combination of dataset, metric, and model, the table shows symbols representing the findings associated with each case derived from the analysis of the experiment results in this context, as described in Table 5.8.

To illustrate the real-world implications of these results, we can analyze the tests carried out using the CatGCN model on the Alibaba dataset. Specifically, the experiments focused on classifying consumption grade as the target class and age as the sensitive attribute, as depicted in Figures 5.17a and 5.18a.

When we estimate the fairness scores using *statistical parity*, we encounter two distinct scenarios. In this assessment, the sensitive attribute is always treated as multi-valued. For a specific age group, we observe that the mid and high-consumption levels are considered privileged in the binary evaluation but disadvantaged in the multiclass assessment. Conversely, for another age group, the binary results display fairness, but a more detailed analysis reveals that the mid and high-consumption levels are once again disadvantaged when assessed individually.

Table 5.8. Description of the cases derived from the assessment of the comparative results between *multigroup* and *multiclass* scenarios.

#	Symbol	Multigroup scenario	Multigroup and multiclass scenario
1	†	Binarized class advantaged	All related fine-grained classes significantly disadvantaged
2	◇	Fair result	All fine-grained classes, belonging to the same binarized class, disadvantaged
3	∗	Binarized class advantaged	Only one or a few of the related fine-grained classes significantly disadvantaged
4	⊙	Fair result	Only one or a few of the related fine-grained classes, belonging to the same binarized class, disadvantaged
5	☆	Unfair result	Fair result
6	⩓	Binarized class disadvantaged	Even greater unfairness against that class (and related fine-grained classes)

Observation 2 *In situations where there is no explicit "positive" category, using multiclass and multigroup fairness metrics to assess both binary and multiclass scenarios reveals the true extent of model discrimination. This approach offers a comprehensive understanding of which groups are being discriminated against in all categories and the disparity in fairness between them. This knowledge is crucial for devising an effective bias mitigation strategy.*

5.5 Summary

In this chapter, we introduced a new method for assessing algorithmic fairness, focusing on real-world applications. We expanded common *binary fairness metrics* (i.e., statistical parity, equal opportunity, overall accuracy equality, and treatment equality) into *multigroup and multiclass metrics*, which is the main contribution of the presented study. By analyzing user modeling tasks using advanced GNN-based models, specifically CATGCN and RHGN, across four datasets, we explored the finer implications of fairness in multiple subgroups. Our findings reveal that multigroup metrics are crucial for uncovering hidden biases and unfair treatment in seemingly equitable situations, enhancing the detection of discrimination against minority groups. Furthermore, our work examined the effectiveness of multiparty and multiplex metrics in contexts lacking a clear beneficial outcome, offering a detailed assessment of biases across different groups and classes. This comprehensive approach supports identifying and addressing the diversity gaps, paving the way for strategies to mitigate biases in complex algorithmic systems. We also highlighted the importance of detailed studies on the interplay between model-dataset combinations and fairness outcomes, underscoring the need for focused research in this area.

Chapter 6

Bias Mitigation for Graph Neural Networks in Binary User Modeling Scenarios

The important thing in science is
not so much to obtain new facts as to
discover new ways of thinking about them.

William Lawrence Bragg

This chapter presents the results of our research on bias mitigation in binary behavioral user modeling scenarios, which follows up the insights derived from the fairness analysis described in Chapter 4, and it is also based on the studies discussed in the same chapter (see specifically Section 4.1), which disclose how unfairness in Graph Neural Networks (GNNs) could be associated with graph topology and the message-passing mechanism (see Section 2.3.4) used to train these models.

In particular, we propose an innovative approach to bias mitigation in GNNs called **FAME**, which stands for Fairness-Aware MEssages, and it is designed to promote fair outcomes by directly modifying the message-passing mechanism, thereby reducing the propagation and amplification of biases during the GNN training process. Unlike existing methods that either balance neighborhood aggregation or apply debiasing techniques post-aggregation, our approach incorporates a bias correction parameter directly into the standard message-passing procedure. This parameter adjusts the influence of sensitive attributes on neighboring nodes, ensuring equitable representations for graph data. We propose two variants of our approach to cater to different types of GNN architectures: FAME, designed for models based on Graph Convolutional Networks (GCNs, see Section 2.3.5), and A-FAME (Attention-FAME), for models based on Graph Attention Networks (GATs, see Section 2.3.5). The effectiveness of our proposed methods is proved by conducting a comprehensive experiment set on three datasets, evaluating the performance under two algorithmic fairness conditions, and comparing against six state-of-the-art baselines, including vanilla GNNs, GNN-agnostic methods, and other

GNN-specific bias mitigation techniques.

6.1 Motivation

In our modern interconnected society, graph data is everywhere, serving as a foundational structure in many natural and artificial systems. Social networks, communication systems, molecules, and transportation networks are just a few examples of such systems constituted by a complex web of relationships [332]. Understanding and analyzing these connections is crucial for making sense of intricate structures and extracting valuable insights. In this context, we have already discussed in the presented manuscript how GNNs (Section 2.3) have recently gained prominence for their ability to successfully process and model graph-structured data and how they have been effectively applied in various fields, especially user modeling (as discussed in Section 4.1).

A core and fundamental operation in GNNs is the **message-passing** mechanism (Section 2.3.4), where nodes exchange and aggregate information from their neighbors in an iterative process [120], which enables GNNs to capture patterns and dependencies within the graph, producing effective node representations for the related downstream applications.

As already outlined in Chapters 4 and 5, despite the acknowledged success in classification tasks, GNNs are susceptible to acquiring biases from the historical data they are trained on and subsequently exhibiting them in their predictions. Due to the particular graph topology, where nodes with similar characteristics and sensitive attributes are likely to be linked to each other [282], multiple research has shown that the learning process of the GNN tends to worsen the spread of bias, making message passing a critical element in reinforcing the propagation of discrimination within these models.

Research studies on **algorithmic fairness** (Section 2.2) have recently become increasingly important in developing techniques to identify and address biases in machine learning (ML) models (see Section 2.2.6). In recent years, many academic contributions have been made to the field of GNNs. These contributions can be broadly categorized as either *GNN-agnostic*, meaning they are universally applicable to any GNN architecture, or *GNN-specific*, indicating that they are tailored to a particular GNN model, such as convolution-based or attention-based, requiring distinct versions to accommodate different architecture types. Within the first category, Merchant and Castillo [232] introduced two approaches, i.e., PFR-AX, which lessens the separability between nodes in protected groups and those in unprotected groups, and PostProcess, which uses a black-box policy to modify the model predictions in order to minimize differences in error rates across various demographic groups. EDITS, proposed by Dong et al. [98], is a method to adversarially alter graph data with a bias-penalization objective function that exploits the Wasserstein distance [336]. Agarwal et al. [10] presented NIFTY, an algorithm to improve the training objective of a GNN by employing layer-wise weight normalization to concurrently reduce bias and improve robustness. Regarding GNN-specific approaches, Dai et al. [82] developed FairGNN, a model that aims to produce fair predictions by leveraging an adversarial debiasing method that addresses the lack of sensitive attribute.

In the same category, Lin et al. [212] recently proposed BeMap, a technique to balance the neighborhood aggregation procedure to mitigate the impact of biased representations, while Jiang et al. [167] presented FMP (Fair Message Passing), a model that explicitly incorporates the use of sensitive attributes in forward propagation for classifying nodes, employing cross-entropy loss without the need for data pre-processing.

The studies produced, such as those previously mentioned, emphasized that strategies for mitigating bias in ML typically concentrate on *pre-processing* and *post-processing* methods (see Section 2.2.6). The former involve modifying the input data to eliminate biases before training, while the latter adjust a model's predictions to achieve fairness. Nevertheless, these approaches, while not tailored to specific models, are frequently less impactful than *in-processing* methods (see Section 2.2.6) that deal with biases that arise throughout the training phase.

Furthermore, differently from what happens for standard ML models [137, 384], breaking the links between sensitive attributes and other attributes is not sufficient when working with graph data and GNNs [232]. Being in close proximity to other nodes within the same protected group might indirectly imply membership, and this influence can impact node representations. Hence, reducing bias may involve training the system to lessen the importance of connections in adjacency information. In this specific situation, addressing unfairness during the model's fundamental training operations by intervening in the message-passing process offers a promising approach to diminishing bias propagation. This intervention has the potential to result in fairer outcomes. To our knowledge, only a small number of contributions have suggested solutions in this area so far (i.e., BeMap [212] and FMP [167]).

6.2 Methodology

In this section, we will begin by explaining the message-passing mechanism and the GNN layers that form the foundation of our innovative techniques. Next, we will outline the fairness metrics and scores used in the experimental setups, along with their mathematical definitions. Lastly, we will present the datasets used in the evaluation phase.

6.2.1 Message-Passing Algorithm

Recalling Section 2.3.2, we commonly define a *graph* as $G = (V, E)$, where V and E denote, respectively, the set of nodes and the set of edges that compose the graph; there is a specific edge $e = (u, v) \in E$ only if a connection between two nodes $u \in V$ and $v \in V$ exists. The *adjacency matrix* is represents as $\mathbf{A} = [a_{uv}] \in \{0, 1\}^{|V|\times|V|}$, where each element a_{uv} indicates the presence or absence of an edge between nodes u and v. *Node features* are defined as $\mathbf{X} \in \mathbb{R}^{|V|\times d}$, where each row denotes a vector $\mathbf{x}_u \in \mathbb{R}^d, \forall u \in V$, with dimension d.

In the **message-passing** algorithm (see Section 2.3.4), every node u exchanges messages with its neighbors within a process made of several iterations, called *layers*. The node's features are subsequently updated by merging the aggregated messages with the

node's prior state to produce a *hidden embedding* $\mathbf{h}_u$. Let k be the current layer, we describe the basic GNN message-passing function [139] with the following equation:

$$\mathbf{h}_u^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{m}_u^{(k)} + \mathbf{b}^{(k)}\right) \tag{6.1}$$

where σ denotes a non-linear activation function (e.g., ReLU [9]), $\mathbf{W}$ is the *weight matrix*, $\mathbf{b}$ is the *bias term vector*[1], and $\mathbf{m}_u$ defines the result of the *message aggregation* associated with node u.

In the presented study, we focused on two kinds of GNN layer types (see Section 2.3.5), which basically vary in how they combine the messages in the process described.

For the **GCN layer**, the aggregation is denoted as:

$$\mathbf{m}_u^{(k)} = \sum_{v \in \mathcal{N}_u \cup \{u\}} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_v|}} \mathbf{h}_v^{(k-1)} \tag{6.2}$$

where $\mathcal{N}_u$ and $|\mathcal{N}_u| = deg(u)$ represent the set of *neighbors* and the *degree* of node u, respectively.

For the **GAT layer**, the aggregation is defined as:

$$\mathbf{m}_u^{(k)} = \sum_{v \in \mathcal{N}_u \cup \{u\}} \alpha_{uv}^{(k)} \mathbf{h}_v^{(k-1)} \tag{6.3}$$

where α_{uv} represents the coefficients calculated by the *attention mechanism*, which determine the impact of each neighbor's characteristics.

6.2.2 Fairness metrics

The fairness metrics adopted in this study are based on the notation presented in Table 2.1. As a common practice in several studies on algorithmic fairness in ML systems (e.g., [82, 382]), in our evaluation, we compute **statistical parity** (Equation (2.2)) and **equal opportunity** (Equation (2.4)) metrics scores. Similar to the procedure applied in our previous studies (see Chapters 4 and 5), we operationalized Equation (2.2) and Equation (2.4) by defining the notions of *disparity* (Δ_{SP}, Equation (4.1)) and *inequality* (Δ_{EO}, Equation (4.2)), also according to existing contributions [82, 232, 261]. For the sake of readiness, we report below the mentioned equations that are used in this chapter's evaluation:

$$\Delta_{SP} = \mid P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1) \mid$$
$$\Delta_{EO} = \mid P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid y = 1, s = 1) \mid$$

Table 6.1. Datasets characteristics

Dataset	Nodes	Edges	Label	Sens. Attr.
German	1K	21K	good-customer	gender
Credit	30K	1.42M	no-default	age
Pokec-z	67K	617K	work-field	region

6.2.3 Datasets

To conduct our evaluation, we considered three publicly accessible datasets from various fields, as indicated in Table 6.1. We converted the prediction labels and the sensitive attributes into the necessary format to address the particular binary classification task.

German[2] [154] is composed of information about clients of a credit bank in Germany that are connected by their account similarity. In the study presented within this chapter, users are classified as *good* or *bad clients* depending on their *gender* for fairness evaluation.

Credit[3] [377] includes data about credit card users in Taiwan that are connected by their purchasing patterns. We exploit the users' *default* in the following month as the target label of the classification task and their *age* as the sensitive attribute for fairness assessment.

Pokec-z[4] [322] contains data from the most popular social network in Slovakia and has already been adopted in our research described in the previous chapter (see Section 5.2.1). In this dataset, the users are related by friendship and classified by their job's *work field*. For the fairness analysis, we leverage the users' *region* as the sensitive attribute.

6.3 Fairness-Aware Messages

This section illustrates in detail the proposed in-processing bias mitigation methods specifically designed for GNN-based models, named **FAME** (Fairness-Aware Messages) and its variant **A-FAME** (Attention-based Fairness-Aware Messages, or simply Attention-FAME).

[1]This "bias term" does not relate to fairness. In neural networks, it serves as an extra parameter, enabling the model to independently adjust its output based on the input. This helps the network better suit the data by shifting the activation function.

[2]`http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data`.
Accessed March 30, 2025.

[3]`https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients`. Accessed March 30, 2025.

[4]Original dataset: `https://snap.stanford.edu/data/soc-pokec.html`.
Version adopted in our study: `https://github.com/EnyanDai/FairGNN/tree/main/dataset/pokec`. Accessed March 30, 2025.

6.3.1 FAME Layer

Implementing an *in-processing* technique, particularly by intervening in the *message-passing* procedure, presents a favorable approach to reducing the spread of bias and potentially promoting fairer results, as previously outlined in Section 6.1.

This chapter introduces **FAME** as a novel approach to mitigate biases during GNN training. It achieves this by modifying the messages with a *correction term*, which reflects the differences in sensitive attributes between connected nodes at each aggregation step of the message-passing process. The goal of this approach is to decrease the impact of nodes with the same sensitive attribute values.

The proposed method is *GNN-specific*, meaning that for each specific GNN layer, a different variant is needed. The standard version of the FAME layer is designed for compatibility with GCN-based models. Taking the basic GNN message-passing function described in Equation (6.1), applying to it the aggregation depicted by Equation (6.2), and a ReLU [9] activation function, we define the FAME layer function as:

$$\mathbf{h}_u^{(k)} = \mathbf{W}^{(k)} \sum_{v \in \mathcal{N}_u \cup \{u\}} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_v|}} \, \delta_{uv}^{(k)} \, \mathbf{h}_v^{(k-1)} \tag{6.4}$$

where δ_{uv} denotes the **fairness correction term**, which is described as:

$$\delta_{uv}^{(k)} = 1 + \mathbf{b}^{(k)} \Delta_s^{(k)} \tag{6.5}$$

with $\mathbf{b}$ representing the *bias term* of Equation (6.1) that scales Δ_s, i.e., the difference between the sensitive attribute values of the nodes.

6.3.2 A-FAME Layer

A-FAME is the FAME version specifically designed for compatibility with GAT-based models. In particular, this variant applies the *correction term* within the attention mechanism α_{uv} described in Equation (6.3), and adopts the softmax [47] activation function. A-FAME is defined by the following equation:

$$\alpha_{uv}^{(k)} = \text{softmax}\left(\mathbf{e}_{uv}^{(k)} + \delta_{uv}^{(k)}\right) \tag{6.6}$$

where $\mathbf{e}_{uv}$ calculates a pair-wise un-normalized attention score between two neighbors. It leverages an *additive attention*, following the definition provided by Veličković et al. [333]. This variant's **fairness correction term** δ_{uv} is defined as:

$$\delta_{uv}^{(k)} = \mathbf{b}^{(k)} \Delta_s^{(k)} \tag{6.7}$$

where $\mathbf{b}$ and Δ_s represent the same components described in Equation (6.5).

6.4 Evaluation

The specific goal of the study presented in this chapter is to address the research questions posed below:

RQ1 How can we enhance the fairness outcomes by directly modifying the message-passing process?

RQ2 How does altering message passing affect fairness compared to state-of-the-art bias mitigation approaches?

We investigate **RQ1** and **RQ2** by evaluating the performance of the two proposed methods, FAME (Section 6.3.1) and A-FAME (Section 6.3.2), on the three datasets introduced in Section 6.2.3 through conducting a user modeling task for each dataset (which can be translated in a node classification task, as described in Section 2.4.3) and calculating fairness measurements related to *disparity* (Δ_{SP}, Equation (4.1)) and *inequality* (Δ_{EO}, Equation (4.2)).

For the purpose of evaluating the effectiveness of our mitigation techniques, the outcomes for **RQ1** are compared with those of the original GCN and GAT models, while the fairness results obtained to address **RQ2** are compared with six state-of-the-art baselines, including vanilla GNNs, GNN-agnostic methods, and GNN-specific bias mitigation techniques.

6.4.1 Baselines

We evaluated our experimental results (discussed in the next section) by comparing our innovative methods against three types of approaches:

- *Standard*, including Vanilla (i.e., the basic GCN and GAT models, see Section 2.3.5, without interventions) and Unaware (i.e., removing the sensitive attribute from the dataset, see Section 2.2.6);
- *GNN-agnostic*, meaning techniques applicable to every GNN model, which include Edits [98] and the three variants of PFR [232];
- *GNN-specific*, where each developed method should be tailored for a specific kind of GNN model. For this category, we exploited BeMap [212] and FairGNN [82].

6.4.2 Experimental setting

For the user modeling task (i.e., the node classification task), the hyperparameters of the models are chosen as outlined below. Regarding the *GNN-agnostic* and *GNN-specific* baselines, all these models' hyperparameters are configured following their default settings, as specified in the original papers. Concerning the *Standard* approaches and our novel methods proposed in this chapter, the hyperparameters are configured through a grid search, as described below. For both GCN and GAT models, the *learning rate*

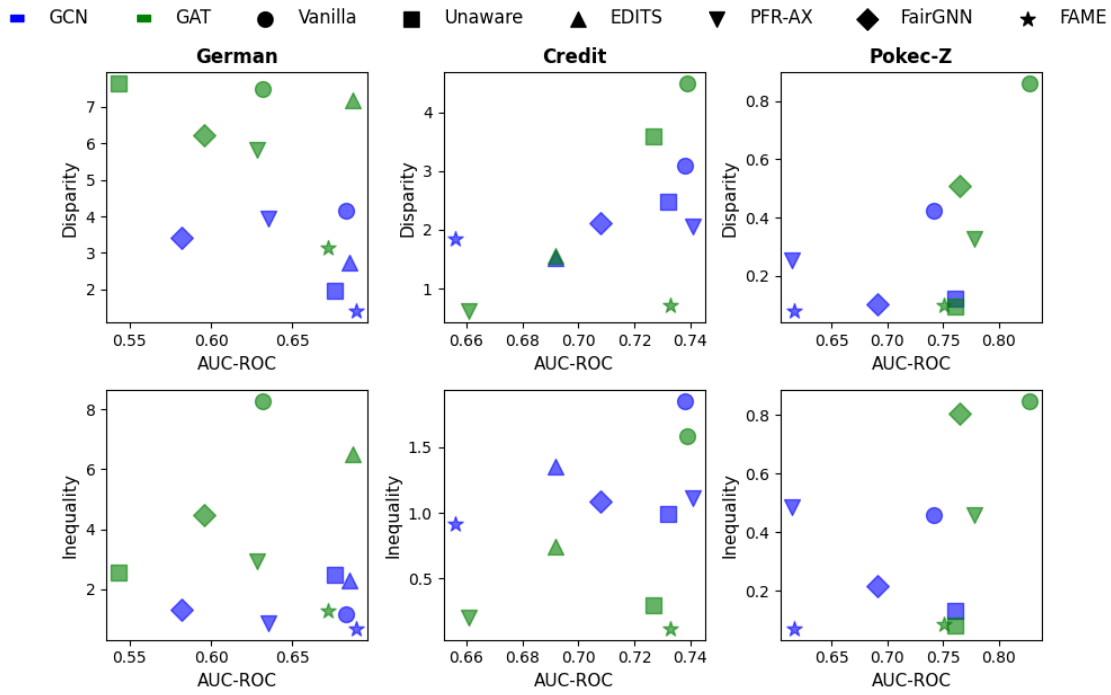


Figure 6.1. Visual experimental results. In all plots, the ideal position is located at the bottom right.

is tuned taking values in {1e-3, 1e-4}, the *weight decay* in {1e-4, 1e-5}, and the *hidden dimensionality size* in {32, 64, 128}. After the grid search, each experiment is executed five times, and the average is considered. All the experiments are performed on an Nvidia Quadro RTX 8000 48GB GPU, except for FAIRGNN, where we utilized a CPU due to compatibility issues with code packages faced during the evaluation phase.

The source code developed during this study, including FAME and A-FAME implementation, as well as the experiments, is publicly available[5].

6.4.3 Experimental results

The results of the conducted experiments are illustrated in Figure 6.1 and in Table 6.2. In the charts, we displayed two approaches for each category in order to provide a clear visualization of the scores. On the other hand, the table presents the complete scores acquired through the carried out experiments.

We start by presenting the performance scores for the classification task using AUC-ROC (shown in decimal format) for every dataset and model combination, followed by the two fairness metrics scores (displayed as percentages). Dashed lines in the table denote *out-of-memory*, and the intuition behind these issues is the following: regarding the experiments conducted with EDITS on POKEC-Z, we experienced a well-known problem

[5] `https://link.erasmopurif.com/FAME`

Table 6.2. Experimental results. Performance scores (AUC-ROC) are reported in decimals, while fairness scores (Disparity and Inequality) are reported in percentages. (↑) indicates that higher values are better, as opposed to (↓), where lower values are better. The specific values represent the average of five runs. Bold indicates the best results, while underlined the second best.

Dataset	Model	Metric	*Standard*		*GNN-agnostic*				*GNN-specific*		
			Vanilla	Unaware	EDITS	PFR-A	PFR-X	PFR-AX	FairGNN	BeMap	FAME
German	GCN	AUC-ROC (↑)	0.683	0.676	0.685	0.677	0.682	0.635	0.582	0.622	**0.689**
		Disparity (↓)	4.141	1.966	2.726	2.276	3.060	3.933	3.401	3.292	**1.402**
		Inequality (↓)	1.186	2.492	2.285	0.736	2.728	0.854	1.334	2.527	**0.697**
	GAT	AUC-ROC (↑)	0.632	0.543	**0.687**	0.626	0.677	0.628	0.596	0.672	0.672
		Disparity (↓)	7.487	7.625	7.177	4.416	3.416	5.831	6.223	4.743	**3.147**
		Inequality (↓)	8.281	2.569	6.525	3.749	3.416	2.942	4.482	1.580	**1.280**
Credit	GCN	AUC-ROC (↑)	0.738	0.732	0.692	0.738	0.740	**0.741**	0.708	0.701	0.656
		Disparity (↓)	3.083	2.473	1.509	**1.320**	2.089	2.058	2.112	4.842	1.853
		Inequality (↓)	1.851	0.996	1.351	1.210	**0.499**	1.115	1.089	1.970	0.912
	GAT	AUC-ROC (↑)	**0.739**	0.727	0.692	0.736	0.743	0.661	-	0.699	0.733
		Disparity (↓)	4.484	3.600	1.561	2.615	1.250	**0.611**	-	3.290	0.712
		Inequality (↓)	1.580	0.298	0.744	0.679	1.005	0.207	-	0.981	**0.121**
Pokec-z	GCN	AUC-ROC (↑)	0.742	**0.761**	-	0.660	0.616	0.615	0.691	0.728	0.617
		Disparity (↓)	0.425	0.120	-	0.235	0.216	0.254	0.101	0.370	**0.077**
		Inequality (↓)	0.458	0.134	-	0.142	0.078	0.486	0.217	0.660	**0.072**
	GAT	AUC-ROC (↑)	**0.828**	0.761	-	0.742	0.742	0.778	0.765	0.737	0.751
		Disparity (↓)	0.860	**0.092**	-	0.680	0.425	0.327	0.510	0.840	0.097
		Inequality (↓)	0.848	0.083	-	0.596	**0.070**	0.458	0.803	0.540	0.085

in the literature about the execution of these experiments, already documented by other researchers (e.g., Merchant and Castillo [232]); concerning FairGNN in the GAT version on Credit, the problem could be attributed to the device adopted for these particular experiments, i.e., the CPU, due to technical issues, as already mentioned in Section 6.4.1.

Even if node classification is not the main focus of the presented study, our methods for reducing bias yield comparable results in all tests. The most significant decrease in AUC-ROC is 9.3%, in the least favorable scenario, compared to the vanilla model.

When we examine the fairness outcomes of our approaches (**RQ1**) and compare them with the standard models, we consistently achieve superior results, reducing bias by an average of 69.8% in terms of *disparity* (Δ_{SP}) and 73.9% in terms of *inequality* (Δ_{EO}).

Observation 1 *Modifying the message-passing procedure by incorporating a correction term that relies on the disparities in sensitive attributes between linked nodes during each stage of the aggregation process effectively limits the influence of nodes with equal sensitive attribute values and mitigates biases.*

When comparing the fairness scores with the selected baselines (**RQ2**), we specifically notice an overall better performance in relation to the other *GNN-specific* approaches (with improvements of 47.7% Δ_{SP} and 53.6% Δ_{EO} on average). Overall, the proposed FAME variants showed the best score in 58% of experiments and the second best score in 25% of cases.

Observation 2 *In the experimental situation provided, our methods for mitigating in-processing bias show strong results when compared to the state-of-the-art bias mitigation strategies, particularly in comparison to GNN-specific approaches, thus establishing the efficacy of this technique.*

6.5 Summary

In this chapter, we presented a novel in-processing bias mitigation method called FAME (short form for Fairness-Aware Messages) specifically tailored for GNNs. It consists of two variants: the standard one, with the same name as the general approach, is designed for GCN-based models, while the second one, called A-FAME (short form for Attention-Based Fairness-Aware Messages or simply Attention-FAME), is designed to be compatible with GAT-based models. These methods adjust the message-passing process by incorporating a correction term based on the differences in sensitive attributes between connected nodes during the aggregation phase. We conducted a comprehensive evaluation by performing user modeling tasks (executed as node classification tasks) on three real-world datasets in two different fairness conditions (i.e., assessing disparity and inequality). Comparing the outcomes of our proposed approaches with six state-of-the-art baselines, the obtained experimental results demonstrated that FAME (and its variant A-FAME) can outperform several existing bias mitigation strategies, setting a valid starting point for deepening research investigation in this promising direction.

Chapter 7

Frameworks for Standardized Fairness Analysis

Vision without execution is just hallucination.

Henry Ford

In this chapter, we will discuss the design and implementation of two novel tools with the objective of standardizing fairness analysis of Graph Neural Networks (GNNs), opening with the motivations leading to these specific research studies in Section 7.1.

The **FairUP** framework, whose components are described in Section 7.2, aims to assess algorithmic fairness on GNN-based models for user modeling tasks and is founded on our first-of-its-kind fairness analysis behavioral user profiling models presented in Chapter 4. In particular, we developed an extensible and unified tool designed to train advanced GNN-based user modeling algorithms on a variety of graph datasets. This framework is equipped to identify biases during the pre-processing and post-processing stages and to address potential unfairness in the original datasets. To allow users to interact with and explore the framework's capabilities, we created an intuitive and accessible user interface (UI). Furthermore, we implemented a prototype version of the framework, which functions using predefined real-world datasets.

The **GNNFairViz** is presented in Section 7.3. It introduces a visual analytics framework that analyzes GNN fairness from a data-centric viewpoint, offering users insights into bias in their models. This tool is GNN agnostic, meaning that it supports various GNN architectures; it also offers interactive visualizations to inspect model bias, enables flexible node selection, and supports fairness diagnostics from the data bias perspective. After describing the developed tool, which integrates a human-in-the-loop approach for investigating fairness issues in GNNs, we illustrate a usage scenario to demonstrate its usability and effectiveness.

7.1 Motivation

GNNs (Section 2.3) have shown ample potential for various applications but face significant challenges related to fairness, especially in human-related decision contexts. As thoroughly discussed in the previous Chapters 4 and 5, these issues are particularly complex due to the unique structure of graph data, which can amplify biases inherent in the data used to train these models. Existing bias detection and mitigation approaches vary widely, and frameworks for effectively assessing and addressing these biases only exist for generic machine learning (ML) models [248]. Examples are *AI Fairness 360* [34], *LiFT* [330], *Fairlearn* [38], and *Fairkit-learn* [170].

To address the critical need for fair and unbiased GNN models, it is essential to develop specific standardized frameworks for computing performance and evaluate fairness for these particular models. Currently, the absence of such methodologies makes it difficult to compare results across different studies and applications. This inconsistency not only hampers the development of universally accepted benchmarks for fairness in GNNs but also impedes the identification of best practices for mitigating bias [197]. A standardized approach would ensure that fairness metrics are applied uniformly across different models and datasets, providing a reliable basis for estimating and improving GNN fairness.

Furthermore, the inherent complexity of fairness issues in GNNs, as in any other ML model, necessitates advanced visual analytics tools. The application of traditional static visualization methods to fairness scores visualization, such as bar charts [196] and scatter plots [158], although useful, fall short of providing the depth of analysis required to fully understand and address biases in GNNs. Recent advancements in interactive visual analysis tools have shown promise in enhancing the thoroughness of fairness analysis [31, 213]. However, most of these tools are designed for general ML models and do not specifically cater to the unique challenges posed by GNNs. The *What-If Tool* [356], for instance, allows users to adjust classifier thresholds and view the impacts on fairness metrics across different subgroups, providing a dynamic way to explore and understand model biases. *DiscriLens* [351] and *FairSight* [12] facilitate a deep understanding of fairness in ML models but are tailored specifically for Euclidean data, whereas GNNs operate on graph data structures, i.e., non-Euclidean data. In graph-based model scenarios, existing tools like *FairRankVis* [367] and *BiaScope* [290] introduce innovative methodologies for exploring and diagnosing algorithmic fairness. Yet, these static tools are limited in their capacity to investigate the intricate relationships between node attributes and graph structures, which frequently underlie the fairness concerns within these models. Hence, it is imperative to establish a comprehensive visual analytics framework capable of formalizing the calculation of fairness and performance metrics, providing interactive visualizations, and enabling thorough examination of model biases. Such a framework would provide GNN practitioners with a method to assess and visually represent the equity and efficacy of their models using standardized procedures. Additionally, it would enable them to comprehend the interaction between node characteristics and graph configuration in influencing model discrimination, as well as obtain practical insights through interactive

visualizations that support bias analysis. This tool would support the identification of specific nodes or subgraphs that disproportionately affect fairness metrics, allowing for targeted interventions. Finally, the ability to visualize fairness metrics in an interactive manner would enhance the transparency and trustworthiness of GNN models. Stakeholders, including developers, researchers, and end-users, can better understand how and why certain biases arise, leading to more informed decision-making processes.

7.2 FairUP

In this section, we describe in detail the FairUP framework, including the different components it is composed of, which allow end users to:

- evaluate the fairness of the input dataset using a specific metric, i.e., *disparate impact* (Equation (2.3));
- reduce potential discrimination found in the dataset by adopting different pre-processing bias mitigation approaches (Section 2.2.6), i.e., *sampling* [172], *reweighting* [172] and *disparate impact remover* [111];
- standardize the input for each GNN model available within the tool by employing a graph structure in *Neo4j*[1] or *NetworkX*[2] format;
- train multiple GNN models, i.e., CatGCN (Section 4.2.1), RHGN (Section 4.2.1), and FairGNN [82], with the possibility to manually set the hyperparameters for each of them;
- assess post-hoc fairness by leveraging four standard metrics in binary scenarios, i.e., *statistical parity* (Equation (2.2)), *equal opportunity* (Equation (2.4)), *overall accuracy equality* (Equation (2.6)), and *treatment equality* (Equation (2.8)).

The logical architecture of the proposed FairUP framework is displayed in Figure 7.1.

7.2.1 Pre-processing component

The pre-processing component can be considered the most important part of the framework due to its main goal to properly prepare the user input data for utilization across all available GNN models. The pre-processing component consists of three modules: the optional *pre-processing fairness evaluation* and *debiasing*, followed by *input standardization*.

[1] `https://neo4j.com/`. Accessed March 30, 2025.
[2] `https://networkx.org/`. Accessed March 30, 2025.

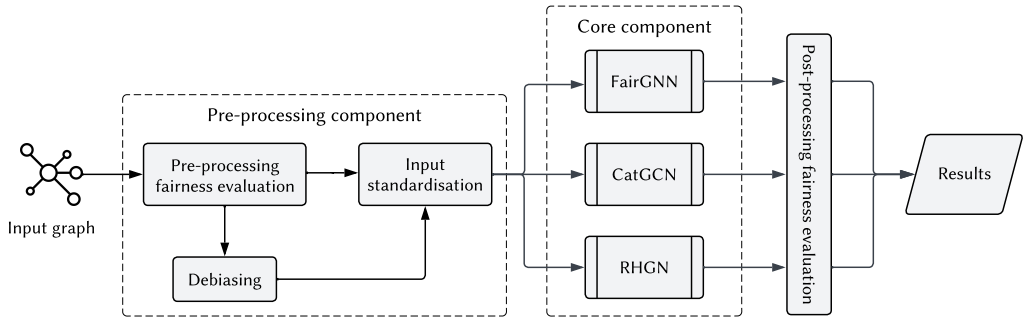


Figure 7.1. Logical architecture of the FAIRUP framework.

Pre-processing fairness evaluation As already discussed, FAIRUP takes as input any graph data from the users in either Neo4j or NetworkX format to help create a unified process. After correctly decoding the input data, the user can assess the fairness of the entire dataset by calculating the score of the *disparate impact* metric (Equation (2.3)). This measure characterizes hidden, often unintentional bias when procedures or systems seem to treat individuals equally based on one of their sensitive characteristics (e.g., age, race, or gender). It also identifies cases where the model discriminates unfairly against specific groups, even if it uses proxy attributes instead of the sensitive attribute directly to make predictions.

Based on the literature and the legal definitions of algorithmic fairness metrics (see Sections 2.2.4 and 2.2.5), a violation of disparate impact arises when the favorable outcome for the disadvantaged group is less than 80% in comparison to the advantaged group.

In the first version of the framework, we consider user modeling tasks in a binary scenario. Thus, the pre-processing component is also responsible for the binarization, if needed, of the selected target class and sensitive attribute for classification and fairness assessment, respectively.

Debiasing After assessing the dataset fairness, if biases are found, the user has the option to utilize a pre-processing debiasing method. The platform provides support for three techniques:

- *Sampling* [172] is a method that aims to re-sample the dataset in order to mitigate or eliminate discrimination. Once the dataset is divided into four groups (denominated as *deprived community with positive class labels*, *deprived community with negative class labels*, *favoured community with positive class labels* and *favoured community with negative class labels*), it computes the expected sizes for each class label and sensitive attribute if the dataset were non-discriminatory. Lastly, a sampling algorithm is applied, either *uniform* or *preferential*.

- The *reweighting* [172] approach aims to address bias in the dataset by assigning varying weights to the dataset entries. Specifically, it assigns higher weights to those entries with an unfavorable sensitive attribute compared to those with a favorable attribute. This is achieved by computing the *expected* and *observed* probability for a specific sensitive attribute label and class label. If the expected probability exceeds the observed probability, bias towards the opposite class label is evident. To counter this, lower weights are allocated to the favored entries.

- *Disparate impact remover* [111], as the name suggests, has been specifically developed to remove disparate impact bias from a dataset. To achieve this, it modifies the sensitive attribute features in order to decrease the correlation between those features and the prediction class and to maintain balance for all prediction classes in the dataset. Disparate impact remover also ensures that the group ranking of the different prediction classes is preserved while editing the sensitive attribute features.

Input standardization Given that each GNN model necessitates a specific input structure, this module is designed to transform the original dataset to conform to the requirements of the selected GNN models. The framework's extensibility ensures that as a GNN model is introduced, the corresponding input standardization procedures can be seamlessly implemented. This means that whenever a new GNN is integrated into the system, an appropriate data preprocessing routine must be developed to meet the novel model's input specifications.

7.2.2 Core component

This component represents the central part of the proposed framework. In its initial version, we incorporated three state-of-the-art GNN-based models that have proven to be highly effective in user modeling tasks, i.e., CatGCN (Section 4.2.1), RHGN (Section 4.2.1) and FairGNN [82]. These models were selected because of their excellent ability to capture and analyze user data, which establishes a solid foundation for the FairUP framework's capabilities. A user can choose any combination and number of the included GNN models for training and evaluation.

7.2.3 Post-processing fairness evaluation

Once the selected models are trained, the framework can evaluate the fairness of their predictions using four standard metrics: *statistical parity* (Equation (2.2)), *equal opportunity* (Equation (2.4)), *overall accuracy equality* (Equation (2.6)), and *treatment equality* (Equation (2.8)). These metrics enable a comprehensive assessment of the models' fairness in order to check whether the predictions are equitable and unbiased across different user groups.

The post-processing fairness evaluation component is built upon the same settings of the assessment presented in Chapter 4, and the employed metrics follow the notation

described in Table 2.1. As in the aforementioned chapter, we operationalize the fairness metrics to use as defined by Equations (4.1) to (4.4).

7.2.4 User interface

In this section, we will illustrate the functionalities of the FAIRUP user interface (UI), developed to allow the users to interact with the framework. The UI is implemented using *Streamlit*[3], an open-source Python framework that enables data scientists and ML practitioners to create dynamic data applications and user-friendly UIs with minimal code.

All resources produced during the implementation of FAIRUP are publicly available, including the source code[4], a web application[5], and a demonstration video[6].

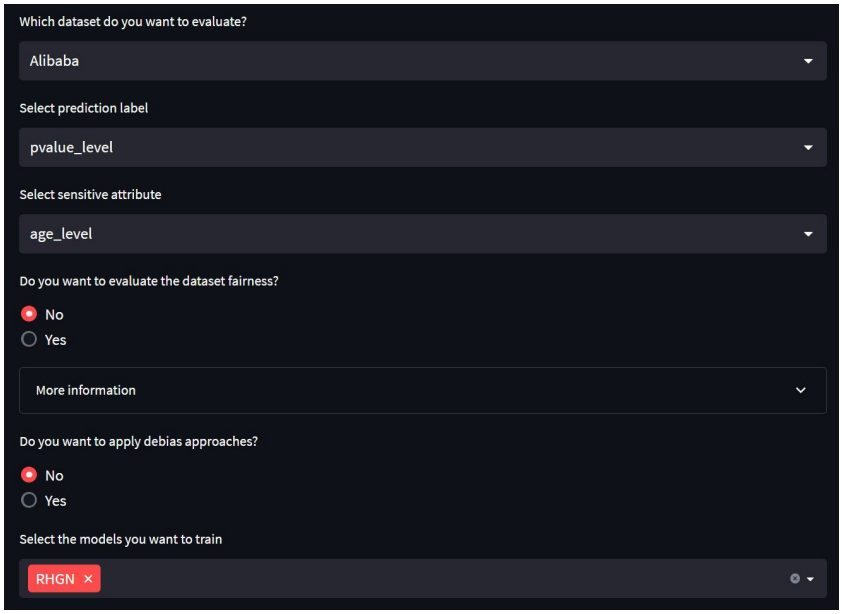


Figure 7.2. FAIRUP UI: Selection of the dataset, input parameters, and pre-processing fairness functionalities.

The opening page of the FAIRUP UI shows the description of the framework components, along with an image of its logical architecture. After clicking on the "*Framework*" button present in the navigation sidebar, the user will be directed to the main page (Figure 7.2), from which it is possible to choose the input dataset, the associated target class, and the sensitive attribute for use in the fairness experiments from a dropdown menu that appears after selecting a dataset. On the same page, we can also choose whether or

[3] `https://streamlit.io/`. Accessed March 30, 2025.
[4] `https://link.erasmopurif.com/FairUP-source-code/`.
[5] `https://link.erasmopurif.com/FairUP/`.
[6] `https://link.erasmopurif.com/FairUP-demo-video/`.

not to implement the pre-processing fairness features. Users can opt for a *preset* configuration for each dataset in FairUP to facilitate quick usage instead of having to manually select every attribute.

We have made four pre-defined datasets available for the first prototype version of the framework, which have also been utilized in other research studies presented in this dissertation: Alibaba and JD, adopted and illustrated in Chapters 4 and 5, NBA, used in Chapter 5, and Pokec-z, utilized in Chapters 5 and 6.

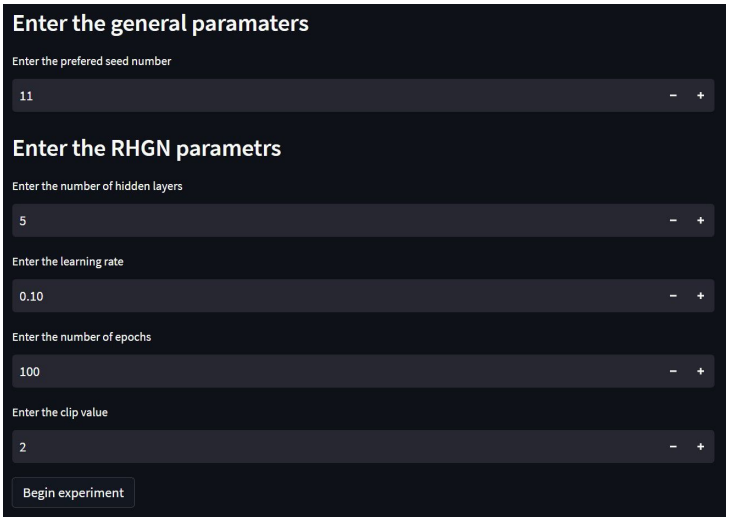


Figure 7.3. FairUP UI: Selection of the training parameters for the chosen GNN model(s). In the displayed example, RHGN parameters are set.

As previously discussed, the users have the option to choose the number of models to be trained. Once the decision is made, the framework requires the input of training parameters for each model (Figure 7.3). After completing the process, the users can initiate the experiment and review a chart that presents the classification and fairness outcomes.

7.3 GNNFairViz

This section will describe GNNFairViz, a visual analytics framework for GNN fairness, developed in collaboration with the Fudan University, based in Shanghai, China. The tool provides the end-users with the following capabilities:

- supporting the customization and examination of fairness from different perspectives, i.e., by exploiting different fairness concepts and adopting different fairness metrics (see Section 2.2.5);
- offering hints and options for interaction to select nodes from the input graph and examine their impact on model bias;

- enabling the interactive diagnosis of fairness issues in GNNs, as static explanations using graph data subsets are often challenging for humans to comprehend.

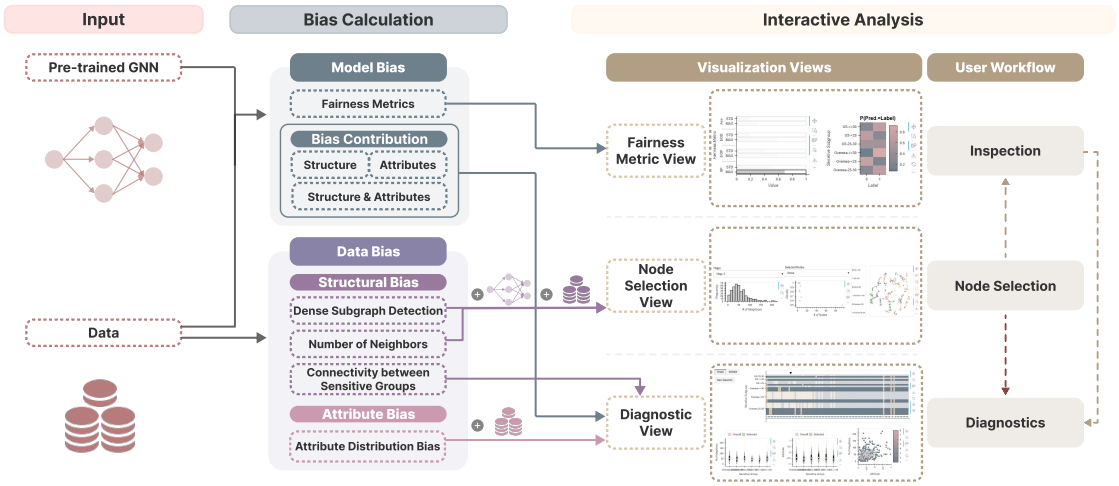


Figure 7.4. Logical architecture of the GNNFAIRVIZ framework.

The GNNFAIRVIZ framework consists of three main components, and its logical architecture is displayed in Figure 7.4. Initially, the process involves providing a pre-trained GNN model and a dataset (either in graph or tabular data format) as the tool's *input*. In the *bias calculation* component, after receiving a defined set of nodes, the *model bias* module evaluates various fairness metrics scores by utilizing the model and data. Additionally, it estimates the impact of the graph structure and attributes on model bias both separately and combined, which we refer to as *bias contribution* in this section. The *structural bias* module analyzes the computational graph (i.e., detecting nodes' neighbors, connectivity between sensitive groups, and dense subgraphs) to uncover potential discrimination in the data structure. The *attribute bias* module adopts a set of statistical tests to assess fairness within attribute distributions. The *interactive analysis* component takes the results from the bias calculation to enable users to concretely get insights about the fairness assessment previously conducted and also to allow bias mitigation procedures.

In the rest of this section, we will describe in detail the key components of the proposed framework, their functionalities, and a use case derived from a real-world scenario.

7.3.1 Bias Calculation

This section describes the techniques used to calculate model bias and data bias, including structural bias and attribute bias.

Model Bias As already explained in previous chapters, model bias occurs when the predictions are discriminatory toward specific sensitive groups. GNNFAIRVIZ incorporates binary fairness metrics (Section 2.2.5), as well as their extension to multiclass and

multigroup scenarios, as defined in Section 5.3, given that binary configurations often fail to capture the complexities of the real world.

Bias Contribution We evaluate the impact of both the graph structure and attributes of a certain group of nodes on model bias by utilizing counterfactual reasoning. This allows us to consider how the result could have changed if the input had been altered in a specific manner.

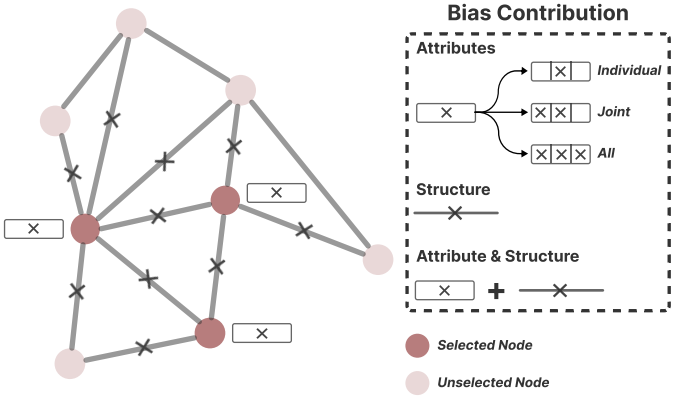


Figure 7.5. GNNFairViz: Bias contribution process

We evaluate the bias contribution of five different factors (i.e., individual characteristics, a group of characteristics, all characteristics, network structure, and the combination of all characteristics and network structure) for a specific set of nodes. The impact is determined by the change in model bias output before and after eliminating specific information from the input data, as exemplified in Figure 7.5.

Data Bias Data bias happens when discrimination is present in the data structures, regardless of the GNN model or the training algorithm.

Structural Bias This may inherently emphasize or de-emphasize specific nodes or groups based on their topological properties. Three functionalities are implemented:

- **Dense Subgraph Detection**: as highlighted in Section 4.1, neighboring nodes sharing the same sensitive attributes can amplify structural biases; thus, we employ a community detection algorithm [66] to reveal dense subgraphs in the input graph.
- **Number of Neighbors**: Different nodes might have different impacts in introducing model bias. To depict the influence of each node in the message-passing process (Section 2.3.4), we calculate the number of neighbors each node possesses in its computational graph. This number directly represents the frequency with which a node's information is propagated.
- **Connectivity between Sensitive Groups**: Given a set of nodes, we measure the impact of each sensitive group within the selected nodes on groups across the

entire dataset by their connectivity. This is quantified by summing the number of neighbors of nodes in their computational graphs for each sensitive group, directly revealing the structural bias patterns.

Attribute Bias This module leverages a multifaceted statistical analysis to detect bias in the attributes, identifying two cases: (1) for *one-hot* or *binary encoded attributes* derived from categorical variables, we apply the *chi-square test* [252] to first analyze the independence between sensitive groups and attribute values, and then to assess whether the distributions of attribute values within each sensitive group shows a significant difference when compared to the rest of the nodes; (2) for *non-binary attributes*, the difference in distributions of attribute values across the sensitive groups is tested with the *Kruskal-Wallis H-test* [193], followed by multiple *Mann-Whitney U tests* [227, 358] to detect disparities in attribute distributions.

7.3.2 Interactive Analysis

This section illustrates the GNNFairViz's modules, referred to as views, that allow the users to interact with the framework. GNNFairViz is mainly developed exploiting the Python packages *Bokeh*[7] and *HoloViz*[8].

Node Selection View This view allows users to select nodes effectively, analyze their influence on model bias, and provides various settings for customization, such as projection algorithms and sampling sizes. It includes: *node embeddings*, which displays spatial relationships and clustering patterns of node embeddings, helping to see differential treatment across sensitive groups; *number of neighbors*, which shows the distribution of neighbors within computational graphs for each node, illustrating their impact during the message-passing process; *dense subgraphs*, which identifies and displays the densely connected subgraphs to suggest potential structural bias.

Fairness Metrics View This view supports users in evaluating the fairness of GNN models by offering detailed insights into the metrics and their implications on model bias. It displays various fairness metrics scores to allow users to inspect model bias comprehensively and provides in-depth details for each selected metric. Heat maps and bar charts show the distribution of metrics across different labels and sensitive groups.

Diagnostic View This view provides tools for analyzing model bias from the perspective of data bias, including both structural and attribute biases. In particular, it displays the contributions of attributes, structure, and their combination to model bias; offers an overview of attribute bias and supports exploration through interactive tools; shows the connectivity patterns among sensitive groups to reveal potential structural bias; visualizes how sensitive groups are distributed to understand specific forms of attribute bias;

[7] `https://bokeh.org/`. Accessed March 30, 2025.
[8] `https://holoviz.org/`. Accessed March 30, 2025.

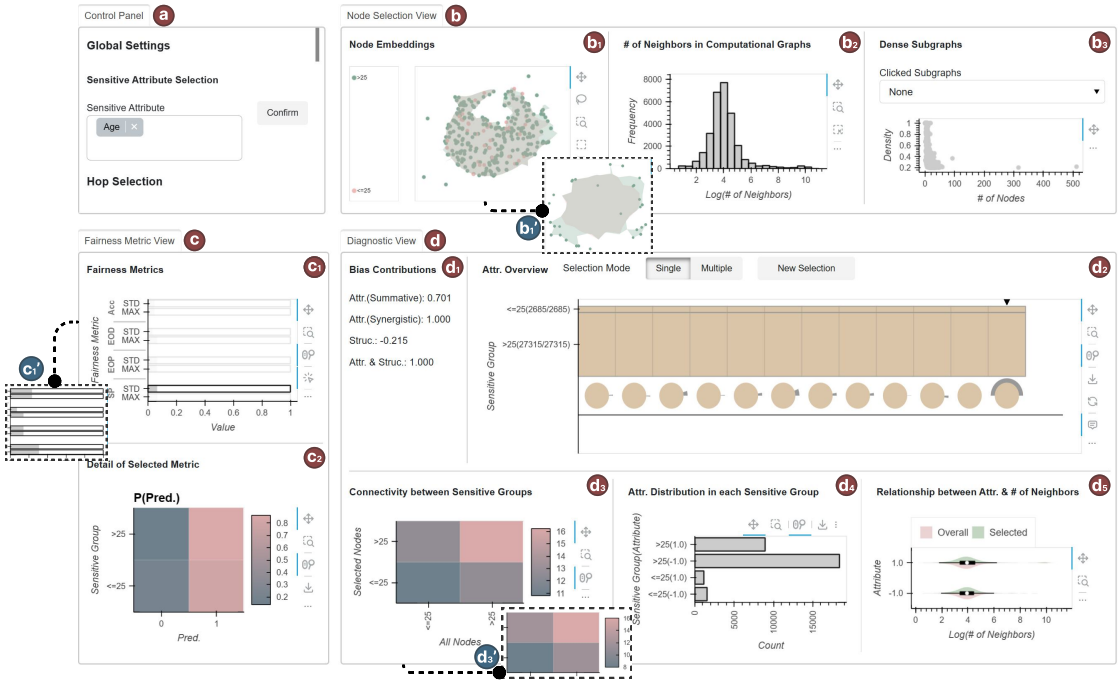


Figure 7.6. Overview of the plots generated by the execution of the GNNFairViz Use Case: *Age fairness in default prediction*

and helps inspect interactions between graph structure and attributes, showing how edge composition may amplify or mitigate bias.

7.3.3 Use Case: Age fairness in default prediction

This section explores how Sam, a GNN expert, uses GNNFairViz to analyze age fairness in default prediction using the Credit dataset[9] [377], which has already been used in our study described in the previous chapter (see Section 6.2.3). This dataset comprises 30 000 nodes, each representing a credit card user, with edges indicating connections based on the similarity of their purchase and payment patterns. The sensitive attribute examined is *age*, specifically distinguishing users older than 25 years from those younger.

Figure 7.6 displays the overview of the plots generated by the execution of the use case described in this section. Within the specific paragraphs, we will refer to each plot by means of a typewriter-style notation (e.g., the plot labeled as "a" in the figure will be indicated as `[a]`).

Model training and initial analysis Sam begins by training a Graph Attention Network (GAT, see Section 2.3.5) model on the dataset to predict future credit card

[9] `https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients`. Accessed March 30, 2025.

payment defaults. After training, Sam uses the *Fairness Metrics View* ([c1]) to perform an initial analysis, which reveals a bias against younger users (≤ 25 years) in the model's predictions. This bias is measured using different metrics, such as *statistical parity* ([c2]).

Node embeddings and dense subgraphs To dive deeper, Sam examines the *Node Embeddings View* ([b2]), which shows the distribution of neighbors for each node. This plot helps Sam identify nodes with many connections that might disproportionately influence the model's fairness. He can filter nodes by the number of neighbors to focus on highly connected ones, which could be contributing to bias.

Sam then moves to the *Dense Subgraph View* ([b3]), which identifies densely connected subgraphs. He selects these subgraphs to examine the local structure of the graph and determines clusters that could affect fairness due to their specific topology. This helps Sam pinpoint specific areas in the graph where structural biases might be present.

Attribute and structural bias analysis Using the *Diagnostic View* ([d]), Sam performs a detailed analysis of attribute and structural biases. He first looks at the *Bias Contributions* plot ([d1]) and discovers that the attribute "*HistoryOfOverduePayments*" significantly contributes to model bias. Sam drills down into this attribute in the *Attribute Overview* plot ([d2]) and sees that younger users have a different distribution of overdue payments compared to older users.

Sam then examines the *Connectivity between Sensitive Groups* plot ([d3]). He finds that nodes in the group with age ≤ 25 frequently connect with nodes in the group with age $>$ 25. This inter-group connectivity suggests that the graph structure might help in reducing differences in node embeddings between age groups, thus promoting fairness.

To further investigate, Sam looks at the *Attribute Distribution in Each Sensitive Group* plot ([d4]). This plot shows that the distribution of "*HistoryOfOverduePayments*" is indeed different across age groups, confirming his previous findings. Sam also checks the *Relationship between Attributes and Number of Neighbors* plot ([d5]), which shows no significant interaction between this attribute and the number of neighbors, demonstrating that the bias is mainly due to the attribute itself rather than its interaction with the graph structure.

Fairness improvements Sam decides to compare different GNN architectures to find a fairer model. He switches to a Graph Convolutional Network (GCN, see Section 2.3.5) model and uses the *Fairness Metrics View* ([c1]) to evaluate its performance. The results indicate that the GCN model provides fairer predictions compared to the GAT model. Sam further explores the *Detail of Selected Metric* plot ([c2]) to see how these metrics vary across different labels and sensitive groups, confirming the GCN model's superior fairness.

7.4 Summary

In this chapter, we presented two innovative frameworks aimed at addressing fairness in GNN-based user profiling models. The first framework, FairUP, is designed for fairness analysis and bias mitigation. It enables users to analyze GNN model prediction results and evaluate fairness metrics scores. FairUP also allows for the mitigation of bias through various pre-processing debiasing approaches prior to training. A user-friendly interface was developed to facilitate interaction with the complex framework, offering end-users the ability to understand, compare, and explore different GNN models as well as various bias detection and mitigation strategies using a standardized tool. The second framework introduced, a visual analytics framework for GNN fairness named GNNFairViz, focuses on analyzing model bias from the perspectives of attribute bias and structural bias. This general and flexible framework includes a visual analysis tool that seamlessly integrates into the working environment and workflows of target users. The evaluation demonstrates GNNFairViz's ability to provide valuable insights for bias mitigation in GNN models.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Chapter 8

Conclusion and Future Research Directions

The only impossible journey is the one you never begin.

Tony Robbins

At the beginning of this dissertation (Chapter 1), we introduced and motivated the four research challenges addressed during the presented doctoral project:

1. Designing and developing automated decision-making systems that reflect Human-Centered Artificial Intelligence (HCAI) and Responsible AI principles and respect European regulations;
2. Assessing and mitigating fairness in behavioral user modeling applications employing Graph Neural Networks (GNNs);
3. Extending existing algorithmic fairness metrics from binary to multiclass and multi-group scenarios to tackle the observed limitations of binary metrics, which often distort real-world contexts;
4. Implementing unified frameworks for a standardized fairness evaluation and visualization.

In this chapter, we will summarize the contributions provided for each of these challenges, including scientific outputs not specifically related to the core topic of the dissertation, highlight potential limitations detected during the work, and discuss future research directions.

8.1 Developing Responsible AI Systems

Ethical considerations and implications around the implementation and usage of AI systems have become central in the last few years. Relatedly, we have witnessed the

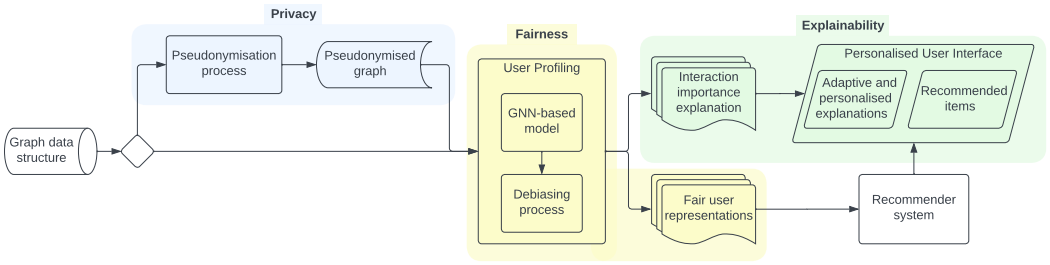


Figure 8.1. Logical architecture of the originally-planned general framework to develop during the doctoral project.

emergence of research fields that prioritize the rights of individuals and, more broadly, of society rather than just the performance of the developed models. In particular, as illustrated in Section 2.1, HCAI and Responsible AI aim to amplify, augment, and enhance human performance to develop AI systems that are ethical, reliable, safe, and transparent, thus trustworthy.

Our preliminary effort centered on creating a specific Responsible AI framework designed for loan approval procedures (Chapter 3). This system adheres to the policies established by European regulations (e.g., *Ethics Guidelines for Trustworthy AI* and *AI Act*), highlighting the importance of interpretability and fairness. The goal of this study was to demonstrate how transparency and equity could improve decision-making processes by instilling trust in the end-users.

Continuing on this research line, before focusing on algorithmic fairness as our core topic, the initial plan for this doctoral project was to address three different beyond-accuracy perspectives related to HCAI, particularly *privacy*, *fairness*, and *explainability*. The idea of building a general framework, whose components are displayed in Figure 8.1, was presented at the UMAP '22 Doctoral Consortium [260].

Along with the extensive work on fairness, several contributions have been produced in order to (partially) cover the other perspectives. For instance, we developed a dynamic privacy-preserving approach for recommendations in an academic environment using pseudonymisation techniques to protect personal data while retaining recommendation performance [276]. Regarding explainability, we specifically focused on the implementation of *explainable user interfaces* (XUIs), which are defined as interfaces that enhance the transparency of AI systems by making their decision-making processes more understandable to users [45, 272]. In our Responsible AI framework (Chapter 3), the XUI aims to explain to bank customers why a certain loan grant request is approved or rejected, to meet the right of the end-users to receive clear motivations for decisions made by an automated system, established by the European regulations. The objective of the GNNFairViz's XUI (Section 7.3) is instead to provide researchers and practitioners with a visual analytics tool to better comprehend the rationale behind a particular outcome in their experiments about fairness in GNN models. With the same goal but in

a different domain, we proposed FACADE [274, 275], a fake news detection system with an innovative XUI designed to assist fact-checkers and content managers in understanding the detailed motivation behind each prediction.

In this context, the most interesting research direction to follow is represented by the *adaptive XUIs* that try to address a typical problem we encounter in terms of explainability, that is, finding UIs following the *one-fits-all* paradigm without considering the different characteristics of individuals. As a preliminary analysis, within the same academic environment as the aforementioned contribution to privacy, we conducted a study to evaluate user perception of different XUIs based on the different researcher profiles [271].

8.2 Fairness Assessment of User Modeling Applications Employing Graph Neural Networks

The constant interaction of the users with automated decision-making systems produces an extensive quantity of data on a daily basis. As defined in our survey [267] and also reported in Section 2.4, acquiring, extracting, and representing user features and personal characteristics is the process of *user modeling* (or *user profiling*) used to construct accurate user models (or user profiles). This process involves drawing inferences about personality traits and behaviors from user-generated data.

It is precisely on user behaviors that our research focused on, as determining the peculiarities of groups of individuals through the study of their actions has become one of the main research areas in modern personalization systems, as underlined in Section 2.4.3, where we described how user modeling topic shifted perspectives over the years. Specifically, our emphasis has been on modeling such behaviors through graph structures, taking into account their centrality in the current scientific landscape, in which, for each domain, entities are considered interconnected [332] and, therefore, analyzable by means of graphs.

The technological choice fell on GNNs (illustrated in detail in Section 2.3), which are specialized architectures created to handle graph-structured data. GNNs are designed to naturally incorporate the dependencies between nodes and edges in a graph. This means they can preserve the graph's inherent structure and topology, which is crucial for many applications where the connections and interactions between entities matter, such as indeed behavioral user modeling, where nodes and edges represent, respectively, users (and items) and the relationships between them.

Although GNNs have shown effectiveness in classifying user models, it is crucial to recognize that, similar to all machine learning (ML) systems trained on historical data, they can be impacted by biases present in the input datasets and may subsequently manifest these biases in their results. The presence of hidden unfair procedures in these models leads to sensible hazards. Even in cases where the models do not intentionally create discrimination (i.e., *disparate impact*, see Section 2.2.4) by exclusively concentrating on behavioral data, ML models can still exhibit systematic biases for certain demographic

groups. Therefore, detecting (Chapter 4) and mitigating (Chapter 6) unfairness in behavioral user modeling is paramount in this field.

Our research on algorithmic fairness (Section 2.2) started from these intuitions. Several contributions have been published about fairness analysis in classical ML models over the last years, as also reported in Section 4.1, while just a few addressed this challenge considering GNN architectures, and in particular, at the beginning of our investigation, none of them assessed fairness in behavioral user modeling applications leveraging GNNs. In this scenario, we carried out two user modeling tasks by conducting binary classification on two real-world datasets (i.e., ALIBABA and JD, see Section 4.2.2), using the most efficient GNNs in this field (i.e., CATGCN and RHGN, see Section 4.2.1). After that, we assessed *disparate impact* and *disparate mistreatment* (see Section 2.2.4) in GNNs designed for behavioral user modeling, employing four binary algorithmic fairness metrics, namely *statistical parity* (Equation (2.2)), *equal opportunity* (Equation (2.4)), *overall accuracy equality* (Equation (2.6)), and *treatment equality* (Equation (2.8)). Through a comprehensive set of experiments (see Section 4.3), we have recognized three important insights into the models under investigation, connecting their distinctive user modeling paradigms to the metric scores:

1. A multiple interaction modeling achieves better fairness scores compared to a model relying on binary associations between nodes;

2. Despite the better results gained by specific models, every GNN designed for user modeling tasks necessitates bias mitigation approaches to concretely produce fair outcomes;

3. Disparate mistreatment assessment is needed to provide a complete fairness evaluation, especially in contexts where there is a high cost for misclassification.

Concerning the previously mentioned biases generated by GNN-based models, several studies (see Section 4.1) revealed that this phenomenon can be mainly attributed to the topology of graph structures and the message-passing procedure, illustrated in Section 2.3.4, typical of these neural architectures. This procedure is crucial for GNN training, allowing nodes to exchange and aggregate information from their neighborhood. However, it can worsen discriminatory impacts since nodes with the same sensitive characteristics are more likely to be linked together than those with different characteristics. As documented in Section 6.1, reducing bias in GNNs is more complex than standard ML models. Breaking the connections between sensitive attributes and other attributes in graph data is not enough. Proximity to other nodes in the same protected group could indirectly suggest membership, impacting node representations. Thus, addressing unfairness during the model's foundational training operations by intervening in the message-passing process presents a promising approach to reducing bias propagation.

With this purpose in mind, our study introduces a new in-processing bias mitigation method, FAME (Fairness-Aware MEssages), which aims to promote equitable outcomes by directly altering the message-passing mechanism. Our approach differs from existing

methods as it integrates a bias correction parameter directly into the standard message-passing procedure rather than solely focusing on balancing neighborhood aggregation or applying debiasing techniques post-aggregation. This parameter serves to modify the impact of sensitive attributes on adjacent nodes, thus guaranteeing fair representations of graph data. Being a *GNN-specific* technique, meaning that it is specifically designed for a particular GNN type, we developed two variants: FAME (Section 6.3), tailored for Graph Convolutional Networks (GCNs, see Section 2.3.5), and A-FAME (Attention-based Fairness-Aware Messages, or simply Attention-FAME, Section 6.3.2), tailored for Graph Attention Networks (GATs, see Section 2.3.5). Our proposed methods' effectiveness has been demonstrated through a comprehensive set of experiments (see Section 6.4) conducted on three datasets (i.e., German, Credit, and Pokec-z, see Section 6.2.3). We evaluated the performance employing two algorithmic fairness metrics (i.e., *statistical parity* and *equal opportunity*, respectively defined in Equations (2.2) and (2.4)) and compared it to six state-of-the-art baselines (Section 6.4.1). These baselines include vanilla GNNs, GNN-agnostic methods, and other GNN-specific bias mitigation techniques. The following insights emerged from the evaluation:

1. Incorporating a correction term that addresses disparities in sensitive attributes between linked nodes during aggregation effectively limits the influence of nodes with equal sensitive attribute values and mitigates biases;

2. Mitigating in-processing bias by directly modifying the message-passing mechanism proves superior performance compared to several state-of-the-art strategies, particularly GNN-specific approaches.

Given the promising results obtained, especially by the proposed bias mitigation approach, future investigations should be extended to different GNN architectures and evaluated in different domains or applications. In particular, preliminary developments started to enhance the method in two aspects: on the one hand, we are implementing a third variant for models based on Graph Isomorphism Networks (GINs) [368], which will probably be named I-FAME; on the other hand, to improve the overall technique, we are experimenting with the introduction of a *fairness-loss function* in order to further optimize the GNN training for specific fairness metrics.

8.3 From Binary to Multiclass and Multigroup Fairness Metrics

According to our analysis depicted in Section 4.1, the majority of measures that target classification parity typically seek to detect and mitigate unfairness in binary scenarios. There are two main reasons why this practice has become so widespread: many tasks involving ML models are naturally binary (such as hiring processes, loan granting procedures, and spam detection), and it is more appropriate to measure fairness mathematically on a binary outcome variable. While these two reasons are technically valid, our research seeks to thoroughly examine and assess the potential consequences

of using such binary measures in user modeling, particularly in real-life situations where enforcing binarization can be questionable from an ethical perspective. Another challenge usually present in bias detection studies is the adoption of the absolute difference between the values of the two sensitive groups being analyzed in the fairness score computation. This methodology creates significant difficulties in identifying disadvantaged demographic groups across various models, datasets, and fairness criteria combinations.

To overcome these limitations, after investigating the experimental results of the fairness assessment of binary behavioral user modeling tasks described before and identifying the need for a more nuanced understanding of algorithmic fairness metrics in real-world applications (see the second part of Chapter 4), we expand the range of classification fairness metrics to cover scenarios in which both the target classes and the sensitive attributes are multi-valued, establishing our work as one of the first thorough initiatives in this area (Chapter 5). In particular, we extended the definition of the four standard fairness metrics belonging to the category of *disparate impact* and *disparate mistreatment* (see Section 2.2.4) already adopted in the binary fairness assessment described in Chapter 4, i.e., *statistical parity* (Equation (2.2)), *equal opportunity* (Equation (2.4)), *overall accuracy equality* (Equation (2.6)), and *treatment equality* (Equation (2.8)). We introduced two sets of metrics:

- *Multigroup fairness metrics* (Equations (5.1) to (5.4)), to assess potential discrimination in contexts where the class to predict is binary, but the analyzed sensitive attribute is multi-valued;

- *Multiclass and multigroup fairness metrics* (Equations (5.5) to (5.8)), to analyze fairness in scenarios where both the target class and the sensitive attribute are multi-valued.

For the first time in the field, we performed an extensive analysis to evaluate the impact of using the proposed generalized metrics instead of their binary counterparts. Specifically, we assessed these metrics on four real-world datasets (i.e., ALIBABA, JD, POKEC, and NBA, see Section 5.2.1) to determine the presence of potential unfairness in binary and multiclass/multigroup scenarios in behavioral user modeling tasks. Our evaluation (see Section 5.4) led to the following insights:

1. Employing multigroup metrics for fairness analysis reveals discrimination against sensitive groups that may be obscured by binary evaluation, crucial for identifying hidden biases and correcting misconceptions about the status of underprivileged groups;

2. In contexts where an explicit positive category is absent (e.g., in classifying a user's consumption grade), using multiclass and multigroup fairness metrics to assess both binary and multi-valued scenarios reveals the full extent of model discrimination. This approach provides a comprehensive understanding of group discrimination and disparities, essential for developing effective bias mitigation strategies.

Our research uncovers the risks and potential distortions that arise when reducing attributes into binary categories, highlighting the potential for creating a false perception of equity. The results of our study highlight the critical necessity of integrating more comprehensive and context-sensitive approaches in order to effectively recognize and address unfair practices in behavioral user modeling. This underscores the valuable insights obtained through a rigorous analytical investigation. In future investigations, in addition to examining general ML models with the same objectives, we will expand our research to disentangle these interactions, particularly examining the influence of dataset characteristics and the structure of binary groups/classes on fairness scores. This will involve a rigorous analysis to comprehend potential correlations, further improving our comprehension of fairness in automated decision-making systems. Moreover, as already designed in a preliminary in-progress experiment, we will study the effect of applying our in-processing bias mitigation approach FAME (Chapter 6) to multiclass and multigroup fairness metrics.

8.4 Unified Frameworks for Fairness Evaluation

The critical need for fair and unbiased GNN models in behavioral user modeling applications was extensively underlined in Chapters 4 and 5. Existing bias detection and mitigation approaches vary widely (see Sections 2.2.4 to 2.2.6), and general benchmarking frameworks for effectively assessing and addressing these biases, at the time of beginning our research, only exist for generic ML models, as motivated in Section 7.1. To address the imperative need for fair and unbiased GNN models applied in behavioral user modeling tasks, which has been the core of the presented doctoral project, it is essential to develop specific standardized frameworks for computing performance and evaluating fairness for these particular models. Such unified frameworks are necessary not only to mitigate biases but also to provide consistent, comparable tools for evaluating GNNs. Consequently, the establishment of robust, standardized fairness assessment frameworks is paramount to achieving methodological coherence and enhancing the rigor of evaluations in the field of GNN-based behavioral user modeling.

In this light, we presented the development and usage of two innovative systems in Chapter 7. The goal was to standardize the evaluation of fairness in applications leveraging GNN-based models.

The first framework described is **FairUP** (Section 7.2). It was built to evaluate the fairness of GNN-based models for user modeling tasks and is based on our first-of-its-kind fairness analysis of binary behavioral user modeling (Chapter 4). Specifically, we developed a flexible and comprehensive tool to train advanced GNN-based user modeling architectures (i.e., CatGCN, RHGN, and FairGNN) on various graph datasets. This framework can detect biases during pre-processing (through the *disparate impact* metric, Equation (2.3)) and post-processing (employing four fairness metrics already adopted in other studies of our research, i.e., *statistical parity*, Equation (2.2), *equal opportunity*, Equation (2.4), *overall accuracy equality*, Equation (2.6), and *treatment equality*, Equation (2.8)), and mitigate potential unfairness in the original datasets (employing three

different approaches, namely *sampling*, *reweighting*, and *disparate impact remover*). We implemented an easy-to-use and accessible UI to enable users to engage with and explore the framework's capabilities (Section 7.2.4).

GNNFAIRVIZ (Section 7.3) is the second framework presented in this dissertation, and as already mentioned, it has been developed in collaboration with the Fudan University of Shanghai, China. It presents a visual analytics framework for examining GNN fairness through a data-centric lens, providing users with insights into bias within their models (see Section 7.3.1). This tool is *GNN-agnostic*, meaning it is compatible with various GNN architectures. It also provides interactive visualizations (see Section 7.3.2) for exploring model bias, allows flexible node selection, and supports fairness diagnostics from a data bias perspective. Following the description of the developed tool, which incorporates a human-in-the-loop approach to probe fairness issues in GNNs, we presented a detailed usage scenario to showcase its usability and effectiveness (i.e., the fairness analysis in a default prediction context considering the customer age as the sensitive attribute, see Section 7.3.3).

Beyond the positive features of the proposed frameworks, we identified the potential limitations described below, which will be analyzed and addressed in future research:

- *Scalability*: training large graph datasets is a well-known computational problem. We have tried to tackle this issue by using multiple GPUs to speed up calculations, refining methods for handling sparse matrices in data processing, and incorporating sampling methods for visualization purposes. As a result, users can enjoy a seamless experience when working with graphs containing up to 20 000 nodes using the current system versions. Although this covers many widely used benchmark datasets, dealing with larger graphs still poses a challenge. The main reason behind these scalability issues is the quadratic growth in the time and space complexity of graph data. Potential future enhancements may involve managing larger graphs by integrating graph database technologies and making use of advanced hardware.

- *Generalization*: both frameworks are designed to be model-agnostic, which means they can be used with a variety of architectures and allow users to specify any combination of sensitive attributes. Currently, they are mainly used for analyzing node classification tasks and do not have support for other tasks like link prediction. Furthermore, although the fairness metrics are based on our research on GNN fairness, users may have specific requirements that call for further customization of these metrics. Our future work will concentrate on expanding our approach to meet these needs.

- *Learning curve*: this drawback applies in particular to GNNFAIRVIZ. Even if it lacks complex visualization charts, it can present a challenge for domain experts, especially those with no prior visual analytics knowledge. The high level of coordination between the visualization charts might not be familiar to GNN experts, who are more used to traditional static visualizations. Providing a tutorial document for users can be beneficial. However, addressing the challenge from the source involves making GNNFAIRVIZ more user-friendly for domain experts to get started while

ensuring it remains powerful for in-depth analysis. This highlights the necessity for continuous improvement and innovation in visual analysis tools.

8.5 Final Remarks

In conclusion, this dissertation has tackled critical challenges in fairness analysis in behavioral user modeling, with a particular focus on ethical AI development. Initially, the study involved designing an automated decision-making system that aligns with HCAI and Responsible AI principles, adhering to European regulations.

The core technical contribution of this work is the innovative evaluation of fairness in user modeling applications employing GNNs. This research introduced novel methods to assess and mitigate biases within these models, demonstrating significant improvements in ensuring equitable AI-driven outcomes.

Furthermore, this dissertation has made a substantial ethical contribution by extending traditional binary algorithmic fairness metrics to multiclass and multigroup scenarios. This addresses the inadequacies of binary metrics, providing a more accurate and contextually relevant approach for fairness evaluation in real-world contexts.

A key advancement is the development of unified frameworks for standardized fairness evaluation and visualization. These structures not only make it easier to conduct consistent and transparent fairness assessments but also provide a more thorough understanding of the moral implications and effectiveness of AI models.

The presented contributions collectively highlight the importance of integrating fairness into AI design and evaluation. By advancing both the technical and ethical aspects of fairness in behavioral user modeling, this dissertation sets a solid foundation for responsible and human-centered AI technologies, paving the way for future research and applications in this crucial field.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Appendix A

Trust & Reliance Scale

The following evaluation scale for explainability system is contextualized in chapter 3 and published in the corresponding journal article [269].

1. I am **confident** in the tool. I feel it works well.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

2. I **trust** the tool's output.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

3. The tool is **reliable**. I can count on it to be correct all the time.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

4. The tool can **perform** the task **better** than a novice human user.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

5. If need be, I feel **confident** in considering changing my decision by taking the tool's output.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

6. I feel **safe** that when I rely on the tool I will get the right answer.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

7. I **like** using the tool for decision making.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

8. The explanations let me judge when I should **trust** and **not trust** the tool.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

Appendix B

Usability Test Questionnaire

This questionnaire is adopted to evaluate the user interface of the Responsible AI framework presented in chapter 3, and it is based on the usability test conducted in a pre-doctoral work [273].

1. Overall, I am satisfied with how easy it is to use this system.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

2. It was simple to use this system.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

3. I was able to complete the tasks quickly using this system.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

4. It was easy to learn to use this system.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

5. It was easy to find the information I needed.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Bibliography

[1] Mohamed Abdelrazek, Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. FairUP: A Framework for Fairness Analysis of Graph Neural Network-Based User Profiling Models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 3165–3169, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–18, New York, NY, USA, April 2018. Association for Computing Machinery.

[3] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver, editors, *User Modeling, Adaption and Personalization*, Lecture Notes in Computer Science, pages 1–12, Berlin, Heidelberg, 2011. Springer.

[4] Sara Abri, Rayan Abri, and Salih Cetin. A Classification on Different Aspects of User Modelling in Personalized Web Search. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '20, pages 194–199, New York, NY, USA, February 2021. Association for Computing Machinery.

[5] Iman I. M. Abu Sulayman and Abdelkader Ouda. User Modeling via Anomaly Detection Techniques for User Authentication. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0169–0176, Vancouver, BC, Canada, October 2019. IEEE.

[6] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[7] Barbara D Adams, Lora E Bruyn, Sébastien Houde, Paul Angelopoulos, Kim Iwasa-Madge, and Carol McCann. Trust in automated systems. *Ministry of National Defence*, 2003.

[8] Gediminas Adomavicius and Alexander Tuzhilin. Context-Aware Recommender Systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 217–253. Springer US, Boston, MA, 2011.

[9] Abien Fred Agarap. Deep Learning using Rectified Linear Units (ReLU), February 2019.

[10] 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*, pages 2114–2124. PMLR, December 2021.

[11] Erfan Aghasian, Saurabh Garg, Longxiang Gao, Shui Yu, and James Montgomery. Scoring Users' Privacy Disclosure Across Multiple Online Social Networks. *IEEE Access*, 5:13118–13130, 2017.

[12] Yongsu Ahn and Yu-Ru Lin. FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1086–1095, January 2020.

[13] Sara Alaoui, Younès El Bouzekri El Idrissi, and Rachida Ajhoun. Building Rich User Profile Based on Intentional Perspective. *Procedia Computer Science*, 73:342–349, 2015.

[14] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, P. Winston Michalak, Shahab Asoodeh, and Flavio P. Calmon. Beyond Adult and COMPAS: Fairness in Multi-Class Prediction, June 2022.

[15] Cliff Allen, Beth Yaeckel, and Deborah Kania. *Internet world guide to one-to-one web marketing*. John Wiley & Sons, Inc., 1998.

[16] Jürgen Allgayer, Karin Harbusch, Alfred Kobsa, Carola Reddig, Norbert Reithinger, and Dagmar Schmauks. XTRA: a natural-language access system to expert systems. *International Journal of Man-Machine Studies*, 31(2):161–195, August 1989.

[17] Gordon Willard Allport, Kenneth Clark, and Thomas Pettigrew. The nature of prejudice. 1954.

[18] Giuseppe Amato and Umberto Straccia. User Profile Modeling and Applications to Digital Libraries. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Serge Abiteboul, and Anne-Marie Vercoustre, editors, *Research and Advanced Technology for Digital Libraries*, volume 1696, pages 184–197. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[19] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural News Recommendation with Long- and Short-term User Representations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy, July 2019. Association for Computational Linguistics.

[20] Vito Walter Anelli, Tommaso Di Noia, Eugenio Di Sciascio, Antonio Ferrara, and Alberto Carlo Maria Mancino. Sparse Feature Factorization for Recommender Systems with Knowledge Graphs. In *Fifteenth ACM Conference on Recommender Systems*, pages 154–165, Amsterdam Netherlands, September 2021. ACM.

[21] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. In *Ethics of Data and Analytics*. Auerbach Publications, 2022.

[22] Maria Rosa Antognazza. *Leibniz: An Intellectual Biography*. Cambridge University Press, October 2008.

[23] Liliana Ardissono and Dario Sestero. Using dynamic user models in the recognition of the plans of the user. *User Modeling and User-Adapted Interaction*, 5(2):157–190, June 1995.

[24] Mozhdeh Ariannezhad, Ming Li, Sebastian Schelter, and Maarten De Rijke. A Personalized Neighborhood-based Model for Within-basket Recommendation in Grocery Shopping. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 87–95, Singapore Singapore, February 2023. ACM.

[25] E Babakus and W G Mangold. Adapting the SERVQUAL scale to hospital services: an empirical investigation. *Health Services Research*, 26(6):767–786, February 1992.

[26] Krisztian Balog, Toine Bogers, Leif Azzopardi, Maarten de Rijke, and Antal van den Bosch. Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 551–558, New York, NY, USA, 2007. Association for Computing Machinery.

[27] Krisztian Balog and Maarten de Rijke. Determining Expert Profiles (With an Application to Expert Finding). In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007.

[28] Krisztian Balog, Yi Fang, Maarten De Rijke, Pavel Serdyukov, Luo Si, et al. Expertise retrieval. *Foundations and Trends® in Information Retrieval*, 6(2–3):127–256, 2012.

[29] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 265–274, Paris France, July 2019. ACM.

[30] Krisztian Balog and ChengXiang Zhai. User Simulation for Evaluating Information Access Systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, pages 302–305, New York, NY, USA, November 2023. Association for Computing Machinery.

[31] Hubert Baniecki, Wojciech Kretowicz, Piotr Piątyszek, Jakub Wiśniewski, and Przemysław Biecek. dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. *Journal of Machine Learning Research*, 22(214):1–7, 2021.

[32] Tim Barnett, Allison W Pearson, Rodney Pearson, and Franz W Kellermanns. Five-factor model personality traits as predictors of perceived and actual usage of technology. *European Journal of Information Systems*, 24(4):374–390, July 2015.

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

[34] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, July 2019.

[35] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations, July 2017.

[36] Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. Contrastive Curriculum Learning for Sequential User Behavior Modeling via Data Augmentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3737–3746, Virtual Event Queensland Australia, October 2021. ACM.

[37] Dan Biddle. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Routledge, London, 2 edition, March 2017.

[38] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, Kathleen Walker, and Allovus Design. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical report, 2020.

[39] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[40] Alan F. Blackwell. Interacting with an inferred world: the challenge of machine learning for humane computer interaction. In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*, CA '15, pages 169–180, Aarhus N, 2015. Aarhus University Press.

[41] Cody Blakeney, Gentry Atkinson, Nathaniel Huish, Yan Yan, Vangelis Metsis, and Ziliang Zong. Measuring Bias and Fairness in Multiclass Classification. In *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–6, October 2022.

[42] Charles Blundell, Lars Buesing, Alex Davies, Petar Veličković, and Geordie Williamson. Towards combinatorial invariance for Kazhdan-Lusztig polynomials. *Representation Theory of the American Mathematical Society*, 26(37):1145–1191, 2022.

[43] Financial Stability Board. Artificial intelligence and machine learning in financial services. Technical report, November 2017.

[44] Dan Bouk. How Our Days Became Numbered: Risk and the Rise of the Statistical Individual. In *How Our Days Became Numbered*. University of Chicago Press, May 2015.

[45] Cassidy Bradley, Dezhi Wu, Hengtao Tang, Ishu Singh, Katelyn Wydant, Brittany Capps, Karen Wong, Forest Agostinelli, Matthew Irvin, and Biplav Srivastava. Explainable Artificial Intelligence (XAI) User Interface Design for Solving a Rubik's Cube. In Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy, editors, *HCI International 2022 – Late Breaking Posters*, pages 605–612, Cham, 2022. Springer Nature Switzerland.

[46] Giorgio Brajnik and Carlo Tasso. A shell for developing non-monotonic user modeling systems. *International Journal of Human-Computer Studies*, 40(1):31–62, January 1994.

[47] John S. Bridle. Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In Françoise Fogelman Soulié and Jeanny Hérault, editors, *Neurocomputing*, pages 227–236, Berlin, Heidelberg, 1990. Springer.

[48] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs, May 2014.

[49] Peter Brusilovski, Alfred Kobsa, and Wolfgang Nejdl. *The Adaptive Web: Methods and Strategies of Web Personalization*. Springer Science & Business Media, April 2007.

[50] Peter Brusilovsky. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11(1):87–110, March 2001.

[51] Peter Brusilovsky. KnowledgeTree: a distributed architecture for adaptive e-learning. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters - WWW Alt. '04*, page 104, New York, NY, USA, 2004. ACM Press.

[52] Peter Brusilovsky, Alfred Kobsa, and Julita Vassileva, editors. *Adaptive Hypertext and Hypermedia*. Springer Netherlands, Dordrecht, 1998.

[53] Erik Brynjolfsson and ANDREW Mcafee. Artificial intelligence, for real. *Harvard business review*, 1:1–31, 2017.

[54] Myles Burnyeat. Plato on why mathematics is good for the soul. In T. Smiley, editor, *Mathematics and Necessity: Essays in the History of Philosophy*, pages 1–81. 2000.

[55] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, June 2016. Publisher: SAGE Publications Ltd.

[56] Alper Caglayan, Magnus Snorrason, Jennifer Jacoby, James Mazzu, Robin Jones, and Krishna Kumar. Learn sesame a learning agent engine. *Applied Artificial Intelligence*, 11(5):393–412, July 1997.

[57] Silvia Calegari and Gabriella Pasi. Ontology-Based Information Behaviour to Improve Web Search. *Future Internet*, 2(4):533–558, December 2010.

[58] Bagher Rahimpour Cami, Hamid Hassanpour, and Hoda Mashayekhi. User preferences modeling using dirichlet process mixture model for a content-based recommender system. *Knowledge-Based Systems*, 163:644–655, January 2019.

[59] Longbing Cao. AI in Finance: Challenges, Techniques, and Opportunities. *ACM Comput. Surv.*, 55(3):64:1–64:38, February 2022.

[60] Yue Cao, Xiaojiang Zhou, Jiaqi Feng, Peihao Huang, Yao Xiao, Dayao Chen, and Sheng Chen. Sampling Is All You Need on Modeling Long-Term User Behaviors for CTR Prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2974–2983, Atlanta GA USA, October 2022. ACM.

[61] Tara Capel and Margot Brereton. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–23, New York, NY, USA, April 2023. Association for Computing Machinery.

[62] Giovanna Castellano, A. Maria Fanelli, Corrado Mencar, and M. Alessandra Torsello. Similarity-Based Fuzzy Clustering for User Profiling. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 75–78, November 2007.

[63] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):166:1–166:38, April 2024.

[64] Junyi Chai, Hao Zeng, Anming Li, and Eric W. T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, December 2021.

[65] Haonan Chen, Zhicheng Dou, Yutao Zhu, Zhao Cao, Xiaohua Cheng, and Ji-Rong Wen. Enhancing User Behavior Sequence Modeling by Generative Tasks for Session Search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 180–190, Atlanta GA USA, October 2022. ACM.

[66] Jie Chen and Yousef Saad. Dense Subgraph Extraction with Application to Community Detection. *IEEE Transactions on Knowledge and Data Engineering*, 24(7):1216–1230, July 2012.

[67] Richard J. Chen, Judy J. Wang, Drew F. K. Williamson, Tiffany Y. Chen, Jana Lipkova, Ming Y. Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7(6):719–742, June 2023.

[68] 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*, pages 1–1, 2021.

[69] 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 Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2116–2122, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization.

[70] Weixin Chen, Mingkai He, Yongxin Ni, Weike Pan, Li Chen, and Zhong Ming. Global and Personalized Graphs for Heterogeneous Sequential Recommendation by Learning Behavior Transitions and User Intentions. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 268–277, Seattle WA USA, September 2022. ACM.

[71] Zhiyong Cheng, Sai Han, Fan Liu, Lei Zhu, Zan Gao, and Yuxin Peng. Multi-Behavior Recommendation with Cascading Graph Convolution Networks. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pages 1181–1189, New York, NY, USA, April 2023. Association for Computing Machinery.

[72] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. Matroids, Matchings, and Fairness. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2212–2220. PMLR, April 2019.

[73] Junsu Cho, Dongmin Hyun, Dong won Lim, Hyeon jae Cheon, Hyoung-iel Park, and Hwanjo Yu. Dynamic Multi-Behavior Sequence Modeling for Next Item Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4199–4207, June 2023.

[74] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017.

[75] Yun-Wei Chu, Seyyedali Hosseinalipour, Elizabeth Tenorio, Laura Cruz, Kerrie Douglas, Andrew Lan, and Christopher Brinton. Mitigating Biases in Student Performance Prediction via Attention-Based Personalized Federated Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3033–3042, Atlanta GA USA, October 2022. ACM.

[76] Philip R. Cohen and C. Raymond Perrault. Elements of a Plan-Based Theory of Speech Acts. *Cognitive Science*, 3(3):177–212, 1979.

[77] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1):312:14730–312:14846, March 2024.

[78] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, fourth edition.* MIT Press, April 2022.

[79] Luca Luciano Costanzo, Yashar Deldjoo, Maurizio Ferrari Dacrema, Markus Schedl, and Paolo Cremonesi. Towards Evaluating User Profiling Methods Based on Explicit Ratings on Item Features. In *Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with 13th ACM Conference on Recommender Systems(RecSys 2019)*, Copenhagen, Denmark, September 2019.

[80] Nick Craswell, Arjen de Vries, and Ian Soboroff. Overview of the TREC-2005 enterprise track. In *TREC*, volume 5, pages 1–7, January 2005.

[81] Mihaela Curmei, Andreas A. Haupt, Benjamin Recht, and Dylan Hadfield-Menell. Towards Psychologically-Grounded Dynamic Preference Models. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 35–48, Seattle WA USA, September 2022. ACM.

[82] Enyan Dai and Suhang Wang. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pages 680–688, New York, NY, USA, March 2021. Association for Computing Machinery.

[83] Lorraine Daston. Classical Probability in the Enlightenment. In *Classical Probability in the Enlightenment.* Princeton University Press, May 1988.

[84] DataRobot. Intelligence briefing: How banks are winning with ai and automated machine learning. White paper, 2019.

[85] Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74, 2021.

[86] Paul De Bra, Peter Brusilovsky, and Geert-Jan Houben. Adaptive hypermedia: from systems to framework. *ACM Computing Surveys*, 31(4es):12, December 1999.

[87] Luis M. De Campos, Juan M. Fernandez-Luna, Juan F. Huete, and Eduardo Vicente-Lopez. Using Personalization to Improve XML Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 26(5):1280–1292, May 2014.

[88] Ernesto William De Luca, Till Plumbaum, Jérôme Kunegis, and Sahin Albayrak. Multilingual ontology-based user profile enrichment. In *MSW*, pages 41–42, 2010.

[89] Joey De Pauw, Koen Ruymbeek, and Bart Goethals. Who do you think I am? Interactive User Modelling with Item Metadata. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 640–643, Seattle WA USA, September 2022. ACM.

[90] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification, March 2023.

[91] William Deringer. *Calculated Values: Finance, Politics, and the Quantitative Age.* Harvard University Press, February 2018.

[92] William Deringer. Just Fines: Mathematical Tables, Church Landlords, and Algorithmic Fairness circa 1628, October 2021.

[93] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, Peter W. Battaglia, Vishal Gupta, Ang Li, Zhongwen Xu, Alvaro Sanchez-Gonzalez, Yujia Li, and Petar Velickovic. ETA Prediction with Graph Neural Networks in Google Maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 3767–3776, New York, NY, USA, 2021. Association for Computing Machinery.

[94] Susan J. Devlin, H. K. Dong, and Marbue Brown. Selecting a scale for measuring quality. *Marketing Research*, 5(3):12–17, 1993.

[95] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 1. Springer, 2019.

[96] Rui Ding, Ruobing Xie, Xiaobo Hao, Xiaochun Yang, Kaikai Ge, Xu Zhang, Jie Zhou, and Leyu Lin. Interpretable User Retention Modeling in Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 702–708, Singapore Singapore, September 2023. ACM.

[97] Yushun Dong, Jian Kang, Hanghang Tong, and Jundong Li. Individual Fairness for Graph Neural Networks: A Ranking based Approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 300–310, New York, NY, USA, 2021. Association for Computing Machinery.

[98] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. EDITS: Modeling and Mitigating Data Bias for Graph Neural Networks. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 1259–1269, New York, NY, USA, April 2022. Association for Computing Machinery.

[99] Paul Dourish. *Where the Action is: The Foundations of Embodied Interaction*. MIT Press, 2001.

[100] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[101] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[102] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, August 2014.

[103] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib, and Henry Friday Nweke. A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions. *IEEE Access*, 7:144907–144924, 2019.

[104] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval*, 16(1-2):1–177, July 2022.

[105] Paul Erickson. The World the Game Theorists Made. In *The World the Game Theorists Made*. University of Chicago Press, November 2015.

[106] Jerry Alan Fails and Dan R. Olsen. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 39–45, New York, NY, USA, 2003. Association for Computing Machinery.

[107] Zhifang Fan, Dan Ou, Yulong Gu, Bairan Fu, Xiang Li, Wentian Bao, Xin-Yu Dai, Xiaoyi Zeng, Tao Zhuang, and Qingwen Liu. Modeling Users' Contextualized Page-wise Feedback for Click-Through Rate Prediction in E-commerce Search. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 262–270, Virtual Event AZ USA, February 2022. ACM.

[108] Marina Farid, Rania Elgohary, Ibrahim Moawad, and Mohamed Roushdy. User Profiling Approaches, Modeling, and Personalization, October 2018.

[109] Ghazal Fazelnia, Eric Simon, Ian Anderson, Benjamin Carterette, and Mounia Lalmas. Variational User Modeling with Slow and Fast Features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 271–279, Virtual Event AZ USA, February 2022. ACM.

[110] Mohd Fazil, Amit Kumar Sah, and Muhammad Abulaish. DeepSBD: A Deep Neural Network Model With Attention Mechanism for SocialBot Detection. *IEEE Transactions on Information Forensics and Security*, 16:4211–4223, 2021.

[111] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney NSW Australia, August 2015. ACM.

[112] Tim Finin and David Drager. GUMS: A General User Modeling System. *Proceedings of the workshop on Strategic computing natural language of the Human Language Technology Conference*, pages 224–230, May 1986.

[113] Josef Fink and Alfred Kobsa. A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web. *User Modeling and User-Adapted Interaction*, 10(2):209–249, June 2000.

[114] Jay W. Forrester. Counterintuitive behavior of social systems. *Theory and Decision*, 2(2):109–140, December 1971.

[115] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein Interface Prediction using Graph Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[116] P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, 9(5):768–786, September 1998.

[117] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996.

[118] Min Gao, Kecheng Liu, and Zhongfu Wu. Personalisation in web computing and informatics: Theories, techniques, applications, and future research. *Information Systems Frontiers*, 12(5):607–629, November 2010.

[119] Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. Personalizing Search Results Using Hierarchical RNN with Query-aware Attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 347–356, Torino Italy, October 2018. ACM.

[120] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. PMLR, July 2017.

[121] Daniela Godoy and Analia Amandi. User profiling in personal information agents: a survey. *The Knowledge Engineering Review*, 20(4):329–361, December 2005.

[122] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable AI: The New 42? In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 295–303, Cham, 2018. Springer International Publishing.

[123] Lin Gong, Lu Lin, Weihao Song, and Hongning Wang. JNET: Learning User Representations via Joint Network Embedding and Topic Embedding. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 205–213, New York, NY, USA, 2020. Association for Computing Machinery.

[124] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016.

[125] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2, July 2005.

[126] Liang Gou, Michelle X. Zhou, and Huahai Yang. KnowMe and ShareMe: understanding automatically discovered personality traits from social media and user sharing preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 955–964, New York, NY, USA, April 2014. Association for Computing Machinery.

[127] Francesca Greco and Alessandro Polli. Emotional Text Mining: Customer profiling in brand management. *International Journal of Information Management*, 51:101934, April 2020.

[128] Aditya Grover and Jure Leskovec. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA, 2016. Association for Computing Machinery.

[129] Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. Partner Matters! An Empirical Study on Fusing Personas for Personalized Response Selection in Retrieval-Based Chatbots. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574, Virtual Event Canada, July 2021. ACM.

[130] Jie Gu, Feng Wang, Qinghui Sun, Zhiquan Ye, Xiaoxiao Xu, Jingmin Chen, and Jun Zhang. Exploiting Behavioral Consistence for Universal User Representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4063–4071, May 2021.

[131] Yulong Gu, Wentian Bao, Dan Ou, Xiang Li, Baoliang Cui, Biyu Ma, Haikuan Huang, Qingwen Liu, and Xiaoyi Zeng. Self-Supervised Learning on Users' Spontaneous Behaviors for Multi-Scenario Ranking in E-commerce. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3828–3837, Virtual Event Queensland Australia, October 2021. ACM.

[132] Yulong Gu, Zhuoye Ding, Shuaiqiang Wang, and Dawei Yin. Hierarchical User Profiling for E-commerce Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 223–231, New York, NY, USA, January 2020. Association for Computing Machinery.

[133] David Gunning and David Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, June 2019. Number: 2.

[134] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yinglong Wang, Jun Ma, and Mohan Kankanhalli. Attentive Long Short-Term Preference Modeling for Personalized Product Search. *ACM Transactions on Information Systems*, 37(2):19:1–19:27, 2019.

[135] 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 *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1808–1812, New York, NY, USA, 2021. Association for Computing Machinery.

[136] Ian Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press, July 2006.

[137] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2125–2126, New York, NY, USA, 2016. Association for Computing Machinery.

[138] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[139] William L. Hamilton. *Graph Representation Learning*. Morgan & Claypool Publishers, September 2020.

[140] Jinkun Han, Wei Li, Zhipeng Cai, and Yingshu Li. Multi-Aggregator Time-Warping Heterogeneous Graph Neural Network for Personalized Micro-Video Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 676–685, Atlanta GA USA, October 2022. ACM.

[141] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[142] Eduardo Hargreaves, Claudio Agosti, Daniel Menasche, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. Fairness in Online Social Network Timelines: Measurements, Models and Mechanism Design. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):68–69, 2019.

[143] Steve Harrison, Deborah Tatar, and Phoebe Sengers. The three paradigms of hci. In *Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA*, pages 1–18, 2007.

[144] Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics.

[145] Taha Hassan, Bob Edmison, Timothy Stelter, and D. Scott McCrickard. Learning to Trust: Understanding Editorial Authority and Trust in Recommender Systems for Education. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 24–32, Utrecht Netherlands, June 2021. ACM.

[146] Winston Haynes. Bonferroni Correction. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 154–154. Springer, New York, NY, 2013.

[147] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong 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 '20, pages 639–648, New York, NY, USA, 2020. Association for Computing Machinery.

[148] J. B. Heaton, N. G. Polson, and J. H. Witte. Deep Learning in Finance, January 2018.

[149] Richard Heiberger and Naomi Robbins. Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications. *Journal of Statistical Software*, 57:1–32, April 2014.

[150] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[151] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[152] Robert D Hof, Heather Green, and Linda Himelstein. Now it's your web. *Business Week*, pages 68–74, 1998.

[153] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects, February 2019.

[154] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

[155] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–16, New York, NY, USA, 2019. Association for Computing Machinery.

[156] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain?, December 2017.

[157] Caley Horan. Insurance Era: Risk, Governance, and the Privatization of Security in Postwar America. In *Insurance Era*. University of Chicago Press, June 2021.

[158] Max Hort, Rebecca Moussa, and Federica Sarro. Multi-objective search for gender-fair and semantically correct word embeddings. *Applied Soft Computing*, 133:1–13, January 2023.

[159] Linmei Hu, Chen Li, Chuan Shi, Cheng Yang, and Chao Shao. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management*, 57(2):102142, March 2020.

[160] Xiaowen Huang, Quan Fang, Shengsheng Qian, Jitao Sang, Yan Li, and Changsheng Xu. Explainable Interaction-driven User Modeling over Knowledge Graph for Sequential Recommendation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 548–556, New York, NY, USA, 2019. Association for Computing Machinery.

[161] Mikella Hurley and Julius Adebayo. Credit Scoring in the Era of Big Data. *Yale Journal of Law and Technology*, 18:148, 2016.

[162] Alex Hämäläinen, Mustafa Mert Çelikok, and Samuel Kaski. Differentiable user models. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 798–808. PMLR, July 2023.

[163] Jim Isaak and Mina J. Hanna. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, 51(8):56–59, August 2018.

[164] Lokesh Jain, Rahul Katarya, and Shelly Sachdeva. Opinion Leaders for Information Diffusion Using Graph Neural Network in Online Social Networks. *ACM Transactions on the Web*, 17(2):13:1–13:37, April 2023.

[165] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, September 2021.

[166] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, July 2018.

[167] Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. Chasing Fairness in Graphs: A GNN Architecture Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21214–21222, March 2024.

[168] Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. Multi-behavior Recommendation with Graph Convolutional Networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 659–668, Virtual Event China, 2020. Association for Computing Machinery.

[169] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019.

[170] Brittany Johnson and Yuriy Brun. Fairkit-learn: a fairness evaluation and comparison toolkit. In *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, ICSE '22, pages 70–74, New York, NY, USA, 2022. Association for Computing Machinery.

[171] Daniel Kahneman and Amos Tversky. On the reality of cognitive illusions. *Psychological Review*, 103(3):582–591, 1996.

[172] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012.

[173] Sumitkumar Kanoje, Sheetal Girase, and Debajyoti Mukhopadhyay. User Profiling Trends, Techniques and Applications, March 2015.

[174] Gerrit Kasper, Diego de Siqueira Braga, Denis Mayr Lima Martins, and Bernd Hellingrath. User profile acquisition: A comprehensive framework to support personal information agents. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–6, November 2017.

[175] Judy Kay. The um toolkit for reusable, long term user models. *User Modeling and User-Adapted Interaction*, 4(3):149–196, 1995.

[176] Judy Kay, Bob Kummerfeld, and Piers Lauder. Personis: A Server for User Models. In Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen, Paul De Bra, Peter Brusilovsky, and Ricardo Conejo, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 2347, pages 203–212. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[177] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*. MIT Press, October 2020.

[178] Ciara Kennefick. The Contribution of Contemporary Mathematics to Contractual Fairness in Equity, 1751–1867. *The Journal of Legal History*, 39(3):307–339, September 2018.

[179] Sein Kim, Namkyeong Lee, Donghyun Kim, Minchul Yang, and Chanyoung Park. Task Relation-aware Continual User Representation Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 1107–1119, New York, NY, USA, 2023. Association for Computing Machinery.

[180] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[181] Alfred Kobsa. Modeling the user's conceptual knowledge in BGP-MS, a user modeling shell system1. *Computational Intelligence*, 6(4):193–208, 1990.

[182] Alfred Kobsa. Generic User Modeling Systems. *User Modeling and User-Adapted Interaction*, 11(1):49–63, March 2001.

[183] Alfred Kobsa and Josef Fink. An LDAP-based User Modeling Server and its Evaluation. *User Modeling and User-Adapted Interaction*, 16(2):129–169, May 2006.

[184] Alfred Kobsa, Jürgen Koenemann, and Wolfgang Pohl. Personalised hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review*, 16(2):111–155, March 2001.

[185] Alfred Kobsa and Wolfgang Pohl. The user modeling shell system BGP-MS. *User Modeling and User-Adapted Interaction*, 4(2):59–106, June 1994.

[186] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, March 1997.

[187] Joseph A. Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1):101–123, April 2012.

[188] Ivica Kostric, Krisztian Balog, and Filip Radlinski. Soliciting User Preferences in Conversational Recommender Systems via Usage-related Questions. In *Fifteenth ACM Conference on Recommender Systems*, pages 724–729, Amsterdam Netherlands, September 2021. ACM.

[189] Nagaraj Kota, Venkatesh Duppada, Ashvini Jindal, and Mohit Wadhwa. Understanding Job Seeker Funnel for Search and Discovery Personalization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3888–3897, Virtual Event Queensland Australia, October 2021. ACM.

[190] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3):1083–1094, March 2022.

[191] Gokul S Krishnan and S Sowmya Kamath. Dynamic and temporal user profiling for personalized recommenders using heterogeneous data sources. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7, July 2017.

[192] Bruce Krulwich. LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data. *AI Magazine*, 18(2):37–37, June 1997.

[193] William H. Kruskal and W. Allen Wallis. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, December 1952.

[194] Alina Köchling, Shirin Riazy, Marius Claus Wehner, and Katharina Simbeck. Highly Accurate, But Still Discriminatory. *Business & Information Systems Engineering*, 63(1):39–54, February 2021.

[195] Alina Köchling and Marius Claus Wehner. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, 13(3):795–848, November 2020.

[196] William G La Cava. Optimizing fairness tradeoffs in machine learning with multiobjective meta-models. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '23, pages 511–519, New York, NY, USA, 2023. Association for Computing Machinery.

[197] Charlotte Laclau, Christine Largeron, and Manvi Choudhary. A Survey on Fairness for Machine Learning on Graphs, February 2024.

[198] Kleanthi Lakiotaki, Nikolaos F. Matsatsinis, and Alexis Tsoukias. Multicriteria User Modeling in Recommender Systems. *IEEE Intelligent Systems*, 26(2):64–76, March 2011.

[199] Pat Langley. User Modeling in Adaptive Interfaces. In *Proceedings of the Seventh International Conference on User Modeling*, 1999.

[200] Arash Habibi Lashkari, Min Chen, and Ali A. Ghorbani. A Survey on User Profiling Model for Anomaly Detection in Cyberspace. *Journal of Cyber Security and Mobility*, pages 75–112, 2019.

[201] Josh Lauer. Creditworthy: A History of Consumer Surveillance and Financial Identity in America. In *Creditworthy*. Columbia University Press, September 2017.

[202] Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli. An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*, 470:443–456, January 2022.

[203] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):14, 1995.

[204] Jonathan Levy. Freaks of Fortune: The Emerging World of Capitalism and Risk in America. In *Freaks of Fortune*. Harvard University Press, October 2012.

[205] James R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, January 1995.

[206] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9):177:1–177:46, 2023.

[207] Lin Li, Zhenglu Yang, Botao Wang, and Masaru Kitsuregawa. Dynamic Adaptation Strategies for Long-Term and Short-Term User Profile to Personalize Search. In Guozhu Dong, Xuemin Lin, Wei Wang, Yun Yang, and Jeffrey Xu Yu, editors, *Advances in Data and Web Management*, Lecture Notes in Computer Science, pages 228–240, Berlin, Heidelberg, 2007. Springer.

[208] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1023–1031, New York, NY, USA, 2012. Association for Computing Machinery.

[209] Sheng Li and Handong Zhao. A Survey on Representation Learning for User Modeling. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4997–5003, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization.

[210] Shangsong Liang and Maarten De Rijke. Formal language models for finding groups of experts. *Information Processing & Management*, 52(4):529–549, July 2016.

[211] Lizi Liao, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. Attributed Social Network Embedding. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2257–2270, December 2018.

[212] Xiao Lin, Jian Kang, Weilin Cong, and Hanghang Tong. BeMap: Balanced Message Passing for Fair Graph Neural Network. In *Proceedings of the Second Learning on Graphs Conference*, pages 37:1–37:25. PMLR, April 2024. ISSN: 2640-3498.

[213] Jessica Liu, Huaming Chen, Jun Shen, and Kim-Kwang Raymond Choo. FairCompass: Operationalising Fairness in Machine Learning. *IEEE Transactions on Artificial Intelligence*, pages 1–10, 2023.

[214] Jiahao Liu, Dongsheng Li, Hansu Gu, Tun Lu, Peng Zhang, Li Shang, and Ning Gu. Triple Structural Information Modelling for Accurate, Explainable and Interactive Recommendation, April 2023.

[215] Jinbo Liu, Yunliang Chen, Xiaohui Huang, Jianxin Li, and Geyong Min. GNN-based long and short term preference modeling for next-location prediction. *Information Sciences*, 629:1–14, June 2023.

[216] Qi Liu, Jinze Wu, Zhenya Huang, Hao Wang, Yuting Ning, Ming Chen, Enhong Chen, Jinfeng Yi, and Bowen Zhou. Federated User Modeling from Hierarchical Information. *ACM Transactions on Information Systems*, 41(2):46:1–46:33, April 2023.

[217] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. Calibrated Fairness in Bandits, July 2017.

[218] R. Logesh, V. Subramaniyaswamy, V. Vijayakumar, and Xiong Li. Efficient User Profiling Based Intelligent Travel Recommender System for Individual and Group of Users. *Mobile Networks and Applications*, 24(3):1018–1033, June 2019.

[219] Hongyu Lu, Weizhi Ma, Min Zhang, Maarten De Rijke, Yiqun Liu, and Shaoping Ma. Standing in Your Shoes: External Assessments for Personalized Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1523–1533, Virtual Event Canada, July 2021. ACM.

[220] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[221] Sichun Luo, Yuanzhang Xiao, and Linqi Song. Personalized Federated Recommendation via Joint Representation Learning, User Clustering, and Model Adaptation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4289–4293, Atlanta GA USA, October 2022. ACM.

[222] Chenglong Ma, Yongli Ren, Pablo Castells, and Mark Sanderson. NEST: Simulating Pandemic-like Events for Collaborative Filtering by Modeling User Needs Evolution. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1430–1440, Atlanta GA USA, October 2022. ACM.

[223] Zhengyi Ma, Zhicheng Dou, Guanyue Bian, and Ji-Rong Wen. PSTIE: Time Information Enhanced Personalized Search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1075–1084, Virtual Event Ireland, October 2020. ACM.

[224] Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. One Chatbot Per Person: Creating Personalized Chatbots based on Implicit User Profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564, Virtual Event Canada, July 2021. ACM.

[225] François Mairesse and Marilyn A. Walker. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction*, 20(3):227–278, August 2010.

[226] Daniele Malitesta, Claudio Pomo, and Tommaso Di Noia. Graph Neural Networks for Recommendation: Reproducibility, Graph Topology, and Node Representation, November 2023.

[227] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[228] Judith Masthoff. Group Recommender Systems: Combining Individual Models. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 677–702. Springer US, Boston, MA, 2011.

[229] Melissa D. McCradden, Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, May 2020.

[230] Jennifer McIntosh, Xiaojiao Du, Zexian Wu, Giahuy Truong, Quang Ly, Richard How, Sriram Viswanathan, and Tanjila Kanij. Evaluating Age Bias In E-commerce. In *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 31–40, May 2021.

[231] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):115:1–115:35, 2021.

[232] Arpit Merchant and Carlos Castillo. Disparity, Inequality, and Accuracy Tradeoffs in Graph Neural Networks for Node Classification. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 1818–1827, New York, NY, USA, 2023. Association for Computing Machinery.

[233] Stuart E. Middleton, Nigel R. Shadbolt, and David C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, 22(1):54–88, 2004.

[234] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 397–407, New York, NY, USA, March 2019. Association for Computing Machinery.

[235] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019.

[236] Sein Minn, Jill-Jênn Vie, Koh Takeuchi, Hisashi Kashima, and Feida Zhu. Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12810–12818, June 2022.

[237] Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nazi, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

[238] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, 2021.

[239] Cataldo Musto, Fedelucio Narducci, Marco Polignano, Marco De Gemmis, Pasquale Lops, and Giovanni Semeraro. MyrrorBot: A Digital Assistant Based on Holistic User Models for Personalized Access to Online Services. *ACM Transactions on Information Systems*, 39(4):46:1–46:34, 2021.

[240] Cataldo Musto, Christoph Trattner, Alain Starke, and Giovanni Semeraro. Towards a Knowledge-aware Food Recommender System Exploiting Holistic User Models. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, pages 333–337, New York, NY, USA, 2020. Association for Computing Machinery.

[241] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning Convolutional Neural Networks for Graphs. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2014–2023. PMLR, June 2016.

[242] Mehrbakhsh Nilashi, Parveen Fatemeh Rupani, Mohammad Mobin Rupani, Hesam Kamyab, Weilan Shao, Hossein Ahmadi, Tarik A. Rashid, and Nahla Aljojo. Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. *Journal of Cleaner Production*, 240:118162, December 2019.

[243] Roger Nkambou, Janie Brisson, Ange Tato, and Serge Robert. Learning Logical Reasoning Using an Intelligent Tutoring System: A Hybrid Approach to Student Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):15930–15937, September 2023.

[244] Martha C Nussbaum. *Sex and social justice*. Oxford University Press, 1999.

[245] Rodrigo Ochigame. The Long History of Algorithmic Fairness, January 2020.

[246] Jon Orwant. Heterogeneous learning in the Doppelgänger user modeling system. *User Modeling and User-Adapted Interaction*, 4(2):107–130, June 1994.

[247] Sara Ouaftouh, Ahmed Zellou, and Ali Idri. User profile model: A user dimension based classification. In *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, pages 1–5, Rabat, October 2015. IEEE.

[248] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, 7(1):15, March 2023.

[249] Ana Paiva and John Self. TAGUS — A user and learner modeling workbench. *User Modeling and User-Adapted Interaction*, 4(3):197–226, September 1994.

[250] Alexandros Paramythis, Stephan Weibelzahl, and Judith Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20(5):383–453, December 2010.

[251] Marco Pavan, Thebin Lee, and Ernesto William De Luca. Semantic enrichment for adaptive expert search. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 1–4, New York, NY, USA, 2015. Association for Computing Machinery.

[252] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, July 1900.

[253] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. Association for Computing Machinery.

[254] C. Raymond Perrault, James F. Allen, and Philip R. Cohen. Speech Acts as a Basis for Understanding Dialogue Coherence. *American Journal of Computational Linguistics*, pages 32–39, December 1978.

[255] Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3):51:1–51:44, February 2022.

[256] Guangyuan Piao and John G. Breslin. Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction*, 28(3):277–329, August 2018.

[257] Till Plumbaum, Songxuan Wu, Ernesto William De Luca, and Sahin Albayrak. User modeling for the social semantic web. In *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation - Volume 781*, SPIM'11, pages 78–89, Aachen, DEU, 2011. CEUR-WS.org.

[258] D. Poo, B. Chng, and Jie-Mein Goh. A hybrid approach for user profiling. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, January 2003.

[259] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–52, New York, NY, USA, 2021. Association for Computing Machinery.

[260] Erasmo Purificato. Beyond-Accuracy Perspectives on Graph Neural Network-Based Models for Behavioural User Profiling. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22, pages 311–315, New York, NY, USA, 2022. Association for Computing Machinery.

[261] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact and Mistreatment in Behavioural User Profiling. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pages 4399–4403, New York, NY, USA, 2022. Association for Computing Machinery.

[262] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Leveraging Graph Neural Networks for User Profiling: Recent Advances and Open Challenges. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 5216–5219, New York, NY, USA, 2023. Association for Computing Machinery.

[263] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Recent advances in fairness analysis of user profiling approaches in e-commerce with graph neural networks. In *Proceedings of the Discussion Papers - 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023 DP)*, pages 47–56. CEUR, 2023.

[264] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Tutorial on User Profiling with Graph Neural Networks and Related Beyond-Accuracy Perspectives. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, pages 309–312, New York, NY, USA, 2023. Association for Computing Machinery.

[265] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Paradigm Shifts in User Modeling: A Journey from Historical Foundations to Emerging Trends. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, pages 13–16, New York, NY, USA, 2024. Association for Computing Machinery.

[266] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks. *Minds and Machines*, 34(3):33, July 2024.

[267] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. User Modeling and User Profiling: A Comprehensive Survey, February 2024.

[268] Erasmo Purificato and Ernesto William De Luca. What Are We Missing in Algorithmic Fairness? Discussing Open Challenges for Fairness Analysis in User Profiling with Graph Neural Networks. In Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, editors, *Advances in Bias and Fairness in Information Retrieval*, Communications in Computer and Information Science, pages 169–175, Cham, 2023. Springer Nature Switzerland.

[269] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *International Journal of Human-Computer Interaction*, pages 1543–1562, April 2023.

[270] Erasmo Purificato, Hannan Mahadik, Ludovico Boratto, and Ernesto William De Luca. Gnn's fame: Fairness-aware messages for graph neural networks. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '25, pages 1–5, New York, NY, USA, 2025. Association for Computing Machinery.

[271] Erasmo Purificato, Baalakrishnan Aiyer Manikandan, Prasanth Vaidya Karanam, Mahantesh Vishvanath Pattadkal, and Ernesto William De Luca. Evaluating Explainable Interfaces for a Knowledge Graph-Based Recommender System. In *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with 15th ACM Conference on Recommender Systems (RecSys 2021)*, volume 2948, pages 73–88, Amsterdam, the Netherlands, September 2021. CEUR Workshop Proceedings.

[272] Erasmo Purificato, Cataldo Musto, Pasquale Lops, and Ernesto William De Luca. First Workshop on Adaptive and Personalized Explainable User Interfaces (APEx-UI 2022). In *27th International Conference on Intelligent User Interfaces*, IUI '22 Companion, pages 1–3, New York, NY, USA, March 2022. Association for Computing Machinery.

[273] Erasmo Purificato and Antonio M. Rinaldi. A Multimodal Approach for Cultural Heritage Information Retrieval. In *Computational Science and Its Applications – ICCSA 2018*, pages 214–230, Cham, 2018. Springer International Publishing.

[274] Erasmo Purificato, Saijal Shahania, and Ernesto William De Luca. Tell Me Why It's Fake: Developing an Explainable User Interface for a Fake News Detection System. In *Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022)*, volume 3277, pages 51–63, Udine, Italy, November 2022. CEUR Workshop Proceedings.

[275] Erasmo Purificato, Saijal Shahania, Marcus Thiel, and Ernesto William De Luca. FACADE: Fake Articles Classification and Decision Explanation. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 294–299, Cham, 2023. Springer Nature Switzerland.

[276] Erasmo Purificato, Sabine Wehnert, and Ernesto William De Luca. Dynamic Privacy-Preserving Recommendations on Academic Graph Data. *Computers*, 10(9):107, September 2021.

[277] Preston Putzel and Scott Lee. Blackbox Post-Processing for Multiclass Fairness, January 2022.

[278] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. News Recommendation with Candidate-aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1917–1921, Madrid Spain, July 2022. ACM.

[279] Hongjin Qian, Zhicheng Dou, Yutao Zhu, Yueyuan Ma, and Ji-Rong Wen. Learning Implicit User Profile for Personalized Retrieval-Based Chatbot. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1467–1477, Virtual Event Queensland Australia, October 2021. ACM.

[280] T. S. Raghu, P. K. Kannan, H. R. Rao, and A. B. Whinston. Dynamic profiling of consumers for customized offerings over the Internet: a model and analysis. *Decision Support Systems*, 32(2):117–134, December 2001.

[281] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Semi-supervised User Geolocation via Graph Convolutional Networks, May 2018.

[282] Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: Towards Fair Graph Embedding. August 2019.

[283] John Rawls. *A Theory of Justice*. Harvard University Press, 1971.

[284] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[285] Kan Ren, Jiarui Qin, Yuchen Fang, Weinan Zhang, Lei Zheng, Weijie Bian, Guorui Zhou, Jian Xu, Yong Yu, Xiaoqiang Zhu, and Kun Gai. Lifelong Sequential Modeling with Personalized Memorization for User Response Prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574, Paris France, July 2019. ACM.

[286] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[287] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

[288] Elaine Rich. User Modeling via Stereotypes. *Cognitive Science*, 3(4):329–354, 1979.

[289] Carlotta Rigotti and Eduard Fosch-Villaronga. Fairness, AI & recruitment. *Computer Law & Security Review*, 53:105966, July 2024.

[290] Agapi Rissaki, Bruno Scarone, David Liu, Aditeya Pandey, Brennan Klein, Tina Eliassi-Rad, and Michelle A. Borkin. BiaScope: Visual Unfairness Diagnosis for Graph Embeddings. In *2022 IEEE Visualization in Data Science (VDS)*, pages 27–36, October 2022.

[291] Cristobal Romero and Sebastian Ventura. Data mining in education. *WIREs Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[292] Somya Ranjan Sahoo and B. B. Gupta. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100:106983, March 2021.

[293] Mohammad Sajib Al Seraj. A Survey on User Modeling in HCI. *Computer Applications: An International Journal*, 5(1):21–28, February 2018.

[294] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Nature, September 2019.

[295] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. FairSNA: Algorithmic Fairness in Social Network Analysis. *ACM Computing Surveys*, 56(8):213:1–213:45, April 2024.

[296] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence*, 283:1–15, June 2020.

[297] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80, January 2009.

[298] Beau G. Schelble, Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J. McNeese, Richard Pak, and Guo Freeman. Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors*, 66(4):1037–1055, April 2024.

[299] Silvia Schiaffino and Analía Amandi. Intelligent User Profiling. In Max Bramer, editor, *Artificial Intelligence An International Perspective: An International Perspective*, Lecture Notes in Computer Science, pages 193–216. Springer, Berlin, Heidelberg, 2009.

[300] Bruno Sguerra, Viet-Anh Tran, and Romain Hennequin. Ex2Vec: Characterizing Users and Items from the Mere Exposure Effect. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 971–977, Singapore Singapore, September 2023. ACM.

[301] Mohammad Ahmad Sheikh, Amit Kumar Goel, and Tapas Kumar. An Approach for Prediction of Loan Approval using Machine Learning Algorithm. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 490–494, July 2020.

[302] Qijie Shen, Wanjie Tao, Jing Zhang, Hong Wen, Zulong Chen, and Quan Lu. SAR-Net: A Scenario-Aware Ranking Network for Personalized Fair Recommendation in Hundreds of Travel Scenarios. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4094–4103, Virtual Event Queensland Australia, October 2021. ACM.

[303] Ben Shneiderman. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 10(4):26:1–26:31, 2020.

[304] Ben Shneiderman. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human-Computer Interaction*, 36(6):495–504, April 2020.

[305] Ben Shneiderman. *Human-Centered AI*. Oxford University Press, January 2022.

[306] Kai Shu, Suhang Wang, and Huan Liu. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435, April 2018.

[307] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 436–439, Vancouver British Columbia Canada, August 2019. ACM.

[308] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 525–534, Lisbon Portugal, November 2007. ACM.

[309] Kerry-Louise Skillen, Liming Chen, Chris D. Nugent, Mark P. Donnelly, William Burns, and Ivar Solheim. Ontological User Profile Modeling for Context-Aware Application Personalization. In José Bravo, Diego López-de Ipiña, and Francisco Moya, editors, *Ubiquitous Computing and Ambient Intelligence*, Lecture Notes in Computer Science, pages 261–268, Berlin, Heidelberg, 2012. Springer.

[310] D. Sleeman. UMFE: A user modelling front-end subsystem. *International Journal of Man-Machine Studies*, 23(1):71–88, July 1985.

[311] Sergey Sosnovsky and Darina Dicheva. Ontological technologies for user modelling. *International Journal of Metadata, Semantics and Ontologies*, 5(1):32, 2010.

[312] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, May 1997.

[313] Riem Spielhaus. Islam and feminism: German and European variations on a global theme. In *Muslim Women and Gender Justice*. Routledge, 2019.

[314] Giuseppe Spillo, Allegra De Filippo, Cataldo Musto, Michela Milano, and Giovanni Semeraro. Towards Sustainability-aware Recommender Systems: Analyzing the Trade-off Between Algorithms Performance and Carbon Footprint. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, pages 856–862, New York, NY, USA, 2023. Association for Computing Machinery.

[315] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[316] Troy Strader, John Rozycki, Thomas Root, and Yu-Hsiang Huang. Machine Learning Stock Market Prediction Studies: Review and Research Directions. *Journal of International Technology and Information Management*, 28(4):63–83, January 2020.

[317] Dirk J Struik. *A concise history of mathematics*. Courier Corporation, 2012.

[318] Lucille Alice Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, November 1987.

[319] K Sudhakar, Boussaadi Smail, Tatireddy Subba Reddy, S Shitharth, Diwakar Ramanuj Tripathi, and Mochammad Fahlevi. Web User Profile Generation and Discovery Analysis using LSTM Architecture. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 371–375, October 2022.

[320] Ke Sun, Tieyun Qian, Tong Chen, Yile Liang, Quoc Viet Hung Nguyen, and Hongzhi Yin. Where to Go Next: Modeling Long- and Short-Term User Preferences for Point-of-Interest Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):214–221, April 2020.

[321] Zhiqing Sun, Jian Tang, Pan Du, Zhi-Hong Deng, and Jian-Yun Nie. DivGraphPointer: A Graph Pointer Network for Extracting Diverse Keyphrases. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 755–764, New York, NY, USA, 2019. Association for Computing Machinery.

[322] Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

[323] Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. User modeling with neural network for review rating prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1340–1346, Buenos Aires, Argentina, 2015. AAAI Press.

[324] Dieudonne Tchuente. User Modeling and Profiling in Information Systems: A Bibliometric Study and Future Research Directions. *Journal of Global Information Management (JGIM)*, 30(1):1–25, January 2022.

[325] Maritzol Tenemaza. User models for recommendation systems. In *Human Factors and Systems Interaction*, volume 52. AHFE Open Acces, 2022.

[326] Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974.

[327] Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. Contestability in Algorithmic Systems. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '19 Companion, pages 523–527, New York, NY, USA, November 2019. Association for Computing Machinery.

[328] Jan-Willem van Dam and Michel van de Velden. Online profiling and clustering of Facebook users. *Decision Support Systems*, 70:60–72, February 2015.

[329] Julita Vassileva. Motivating participation in social computing applications: a user modeling perspective. *User Modeling and User-Adapted Interaction*, 22(1):177–201, April 2012.

[330] Sriram Vasudevan and Krishnaram Kenthapadi. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 2773–2780, New York, NY, USA, 2020. Association for Computing Machinery.

[331] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[332] Petar Veličković. Everything is Connected: Graph Neural Networks. *Current Opinion in Structural Biology*, 79:102538, April 2023.

[333] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[334] Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, and Erik Duval. Context-Aware Recommender Systems for Learning: A Survey and Future Challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, October 2012.

[335] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pages 1–7, Gothenburg Sweden, May 2018. ACM.

[336] Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Soc., August 2021.

[337] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[338] Duc-Vinh Vo, Jessada Karnjana, and Van-Nam Huynh. An integrated framework of learning and evidential reasoning for user profiling using short texts. *Information Fusion*, 70:27–42, June 2021.

[339] Moritz von Zahn, Stefan Feuerriegel, and Niklas Kuehl. The Cost of Fairness in AI: Evidence from E-Commerce. *Business & Information Systems Engineering*, 64(3):335–348, June 2022.

[340] Wolfgang Wahlster and Alfred Kobsa. User Models in Dialog Systems. In Alfred Kobsa and Wolfgang Wahlster, editors, *User Models in Dialog Systems*, Symbolic Computation, pages 4–34, Berlin, Heidelberg, 1989. Springer.

[341] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. Modeling Techniques for Machine Learning Fairness: A Survey, April 2022.

[342] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):35:1–35:27, March 2023.

[343] Putra Wanda and Huang Jin Jie. DeepProfile: Finding fake profile in online social network using dynamic CNN. *Journal of Information Security and Applications*, 52:102465, June 2020.

[344] Chunyang Wang, Yanmin Zhu, Haobing Liu, Wenze Ma, Tianzi Zang, and Jiadi Yu. Enhancing User Interest Modeling with Knowledge-Enriched Itemsets for Sequential Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1889–1898, Virtual Event Queensland Australia, October 2021. ACM.

[345] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, pages 134–147, New York, NY, USA, March 2022. Association for Computing Machinery.

[346] Dongjie Wang, Pengyang Wang, Kunpeng Liu, Yuanchun Zhou, Charles E. Hughes, and Yanjie Fu. Reinforced Imitative Graph Representation Learning for Mobile User Profiling: An Adversarial Training Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4410–4417, May 2021.

[347] Jingkun Wang, Yongtao Jiang, Haochen Li, and Wen Zhao. Improving News Recommendation with Channel-Wise Dynamic Representations and Contrastive User Modeling. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 562–570, Singapore Singapore, February 2023. ACM.

[348] Pengyang Wang, Yanjie Fu, Hui Xiong, and Xiaolin Li. Adversarial Substructured Representation Learning for Mobile User Profiling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 130–138, New York, NY, USA, 2019. Association for Computing Machinery.

[349] Pengyang Wang, Kunpeng Liu, Lu Jiang, Xiaolin Li, and Yanjie Fu. Incremental Mobile User Profiling: Reinforcement Learning with Spatial Knowledge Graph for Modeling Event Streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 853–861, New York, NY, USA, August 2020. Association for Computing Machinery.

[350] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, and Huamin Qu. ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[351] Qianwen Wang, Zhenhua Xu, Chen Zhu-Tian, Yong Wang, Shixia Liu, and Huamin Qu. Visual Analysis of Discrimination in Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1470–1480, February 2021.

[352] Ruiqin Wang, Zongda Wu, Jungang Lou, and Yunliang Jiang. Attention-based dynamic user modeling and Deep Collaborative filtering recommendation. *Expert Systems with Applications*, 188:116036, February 2022.

[353] Honghao Wei, Fuzheng Zhang, Nicholas Jing Yuan, Chuan Cao, Hao Fu, Xing Xie, Yong Rui, and Wei-Ying Ma. Beyond the Words: Predicting User Personality from Heterogeneous Information. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 305–314, New York, NY, USA, February 2017. Association for Computing Machinery.

[354] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. Hierarchical User Intent Graph Network for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 24:2701–2712, 2022.

[355] Hong Wen, Jing Zhang, Fuyu Lv, Wentian Bao, Tianyi Wang, and Zulong Chen. Hierarchically Modeling Micro and Macro Behaviors via Multi-Task Learning for Conversion Rate Prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2187–2191, Virtual Event Canada, July 2021. ACM.

[356] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, January 2020.

[357] Anna Wierzbicka. *English: Meaning and Culture*. Oxford University Press, USA, 2006.

[358] Frank Wilcoxon. Individual Comparisons by Ranking Methods. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer, New York, NY, 1992.

[359] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. User-as-Graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1624–1630, Montreal, Canada, August 2021. International Joint Conferences on Artificial Intelligence Organization.

[360] Chuhan Wu, Fangzhao Wu, Junxin Liu, Shaojian He, Yongfeng Huang, and Xing Xie. Neural Demographic Prediction using Search Query. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 654–662, New York, NY, USA, 2019. Association for Computing Machinery.

[361] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. UserBERT: Pre-training User Model with Contrastive Self-supervision. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2087–2092, New York, NY, USA, 2022. Association for Computing Machinery.

[362] Jinze Wu, Qi Liu, Zhenya Huang, Yuting Ning, Hao Wang, Enhong Chen, Jinfeng Yi, and Bowen Zhou. Hierarchical Personalized Federated Learning for User Modeling. In *Proceedings of the Web Conference 2021*, WWW '21, pages 957–968, New York, NY, USA, 2021. Association for Computing Machinery.

[363] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, October 2016.

[364] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, January 2021.

[365] Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Xiyue Zhang, Hongsheng Yang, Jian Pei, and Liefeng Bo. Knowledge-Enhanced Hierarchical Graph Transformer Network for Multi-Behavior Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4486–4493, May 2021.

[366] Yikun Xian, Tong Zhao, Jin Li, Jim Chan, Andrey Kan, Jun Ma, Xin Luna Dong, Christos Faloutsos, George Karypis, S. Muthukrishnan, and Yongfeng Zhang. EX3: Explainable Attribute-aware Item-set Recommendations. In *Fifteenth ACM Conference on Recommender Systems*, pages 484–494, Amsterdam Netherlands, September 2021. ACM.

[367] Tiankai Xie, Yuxin Ma, Jian Kang, Hanghang Tong, and Ross Maciejewski. FairRankVis: A Visual Analytics Framework for Exploring Algorithmic Fairness in Graph Mining Models. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):368–377, January 2022.

[368] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, New Orleans, Louisiana, United States, 2019.

[369] Wei Xu. Toward human-centered AI: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019.

[370] Hongrui Xuan, Yi Liu, Bohan Li, and Hongzhi Yin. Knowledge Enhancement for Contrastive Multi-Behavior Recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 195–203, Singapore Singapore, February 2023. ACM.

[371] Lyuxin Xue, Deqing Yang, and Yanghua Xiao. Factorial User Modeling with Hierarchical Graph Neural Network for Enhanced Sequential Recommendation. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06, July 2022.

[372] 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*, CIKM '21, pages 3573–3577, New York, NY, USA, 2021. Association for Computing Machinery.

[373] Shaojie Yan, Tao Zhao, and Jinsheng Deng. Interaction-aware Hypergraph Neural Networks for User Profiling. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, Shenzhen, China, October 2022. IEEE.

[374] Boming Yang, Dairui Liu, Toyotaro Suzumura, Ruihai Dong, and Irene Li. Going Beyond Local: Global Graph-Enhanced Personalized News Recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 24–34, Singapore Singapore, September 2023. ACM.

[375] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph Convolutional Networks for Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377, July 2019.

[376] Xinwu Ye, Jielin Feng, Erasmo Purificato, Ludovico Boratto, Michael Kamp, Zengfeng Huang, and Siming Chen. GNNFairViz: Visual Analysis for Graph Neural Network Fairness. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–17, 2025. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[377] I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2016.

[378] 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*, KDD '18, pages 974–983, New York, NY, USA, 2018. Association for Computing Machinery.

[379] Zeping Yu, Jianxun Lian, Ahmad Mahmoody, Gongshen Liu, and Xing Xie. Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4213–4219, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization.

[380] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. Parameter-Efficient Transfer from Sequential Behaviors for User Modeling and Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1469–1478, Virtual Event China, July 2020. ACM.

[381] Fajie Yuan, Guoxiao Zhang, Alexandros Karatzoglou, Joemon Jose, Beibei Kong, and Yudong Li. One Person, One Model, One World: Learning Continual User Representation without Forgetting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 696–705, Virtual Event Canada, July 2021. ACM.

[382] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171–1180, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee.

[383] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in Ranking: A Survey. *ACM Computing Surveys*, 55(6):1–41, July 2023. arXiv:2103.14000 [cs].

[384] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR, May 2013.

[385] Rim Zghal Rebaï, Leila Ghorbel, Corinne Amel Zayani, and Ikram Amous. An Adaptive Method for User Profile Learning. In Barbara Catania, Giovanna Guerrini, and Jaroslav Pokorný, editors, *Advances in Databases and Information Systems*, Lecture Notes in Computer Science, pages 126–134, Berlin, Heidelberg, 2013. Springer.

[386] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 793–803, New York, NY, USA, 2019. Association for Computing Machinery.

[387] Yanfu Zhang, Hongchang Gao, Jian Pei, and Heng Huang. Robust Self-Supervised Structural Graph Neural Network for Social Network Prediction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 1352–1361, New York, NY, USA, April 2022. Association for Computing Machinery.

[388] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021.

[389] Zheng Zhang, Qi Liu, Hao Jiang, Fei Wang, Yan Zhuang, Le Wu, Weibo Gao, and Enhong Chen. FairLISA: Fair User Modeling with Limited Sensitive Attributes Information. *Advances in Neural Information Processing Systems*, 36:41432–41450, December 2023.

[390] Xuanzhi Zheng, Guoshuai Zhao, Li Zhu, and Xueming Qian. PERD: Personalized Emoji Recommendation with Dynamic User Preference. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1922–1926, Madrid Spain, July 2022. ACM.

[391] Zhi Zheng, Zhaopeng Qiu, Tong Xu, Xian Wu, Xiangyu Zhao, Enhong Chen, and Hui Xiong. CBR: Context Bias aware Recommendation for Debiasing User Modeling and Click Prediction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pages 2268–2276, New York, NY, USA, April 2022. Association for Computing Machinery.

[392] 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, January 2020.

[393] Xujuan Zhou, Yue Xu, Yuefeng Li, Audun Josang, and Clive Cox. The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review*, 37(2):119–132, February 2012.

[394] Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. Contrastive Learning of User Behavior Sequence for Context-Aware Document Ranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2780–2791, Virtual Event Queensland Australia, October 2021. ACM.

[395] Philip Zigoris and Yi Zhang. Bayesian adaptive user profiling with explicit & implicit feedback. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 397–404, New York, NY, USA, November 2006. Association for Computing Machinery.

[396] John Zimmerman and Kaushal Kurapati. Exposing profiles to build trust in a recommender. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 608–609, New York, NY, USA, April 2002. Association for Computing Machinery.

[397] Mustafa Mert Çelikok, Pierre-Alexandre Murena, and Samuel Kaski. Modeling needs user modeling. *Frontiers in Artificial Intelligence*, 6, 2023.

[398] Konrad Żołna and Bartłomiej Romański. User Modeling Using LSTM Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), February 2017.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg

Author's Publications

The complete list of the scientific publications produced during the Ph.D. is provided below, in reverse chronological order.

Note: The * symbol indicates the authors' equal contribution to the core part of the specific work. Instead, the alphabetical order of the author list denotes the shared primary contribution.

1. **Purificato, E.**, Mahadik, H., Boratto, L. and De Luca, E.W. (2025). *GNN's FAME: Fairness-Aware MEssages for Graph Neural Networks*. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '25), pp. 1-5.
2. Ye, X., Feng, J., **Purificato, E.**, Boratto, L., Kamp, M., and Huang, Z. (2025). *GNN-FairViz: Visual Analysis for Graph Neural Network Fairness*. In *IEEE Transactions on Visualization and Computer Graphics*, pp. 1-17.
3. **Purificato, E.**, Boratto, L., and De Luca, E.W. (2024). *Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks*. In *Minds and Machines*, 34(33), pp. 1-34.
4. **Purificato, E.**, Boratto, L., and De Luca, E.W. (2024). *Paradigm Shifts in User Modeling: A Journey from Historical Foundations to Emerging Trends*. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '24), pp. 13-16.
5. Boratto, L., Malitesta, D., Marras, M., Medda, G., Musto, C., and **Purificato, E.** (2024). *First International Workshop on Graph-Based Approaches in Information Retrieval (IRonGraphs 2024)*. In *European Conference on Information Retrieval* (ECIR '24), pp. 415-421. Cham: Springer Nature Switzerland.
6. **Purificato, E.**, Boratto, L., and De Luca, E.W. (2023). *Recent Advances in Fairness Analysis of User Profiling Approaches in E-Commerce with Graph Neural Networks*. In *Proceedings of the Discussion Papers - 22nd International Conference of the Italian Association for Artificial Intelligence* (AIxIA '23). CEUR Workshop Proceedings, Vol. 3268, pp. 47-56.
7. Abdelrazek, M.*, **Purificato, E.***, Boratto, L. and De Luca, E.W. (2023). *FairUP: A Framework for Fairness Analysis of Graph Neural Network-based User Profiling Models*. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '23), pp. 3165-3169.

8. **Purificato, E.** and De Luca, E.W. (2023). *What Are We Missing in Algorithmic Fairness? Discussing Open Challenges for Fairness Analysis in User Profiling with Graph Neural Networks*. In *Advances in Bias and Fairness in Information Retrieval* (BIAS '23), pp. 169-175. Cham: Springer Nature Switzerland.
9. **Purificato, E.**, Lorenzo, F., Fallucchi, F., and De Luca, E.W. (2023). *The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes*. In *International Journal of Human-Computer Interaction*, 39(7), pp. 1543-1562.
10. **Purificato, E.**, Boratto, L., and De Luca, E.W. (2023). *Leveraging Graph Neural Networks for User Profiling: Recent Advances and Open Challenges*. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (CIKM '23), pp. 5216-5219.
11. **Purificato, E.**, Boratto, L. and De Luca, E.W. (2023). *Tutorial on User Profiling with Graph Neural Networks and Related Beyond-Accuracy Perspectives*. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '23), pp. 309-312.
12. De Luca, E.W.*, **Purificato, E.***, Boratto, L., Marrone, S., and Sansone, C. (2023). *First Workshop on User Perspectives in Human-Centred Artificial Intelligence (HCAI4U)*. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter* (CHItaly '23), pp. 1-3.
13. **Purificato, E.***, Wehnert, S.*, and De Luca, E.W. (2023). *Usability Studies in Times of Pandemic: Different Solutions for the Remote Usability Tests of Research Digital Tools*. In *International Conference on Human-Computer Interaction* (HCI International '23), pp. 666-673. Cham: Springer Nature Switzerland.
14. Wehnert, S.*, **Purificato, E.***, and De Luca, E.W. (2023). *A Usability Study of a Research Institute Website with Eye-Tracking Devices*. In *International Conference on Human-Computer Interaction* (HCI International '23), pp. 702-711. Cham: Springer Nature Switzerland.
15. **Purificato, E.***, Shahania, S.*, Thiel, M. and De Luca, E.W. (2023). *FACADE: Fake Articles Classification and Decision Explanation*. In *European Conference on Information Retrieval* (ECIR '23), pp. 294-299. Cham: Springer Nature Switzerland.
16. **Purificato, E.***, Shahania, S.*, and De Luca, E.W. (2022). *Tell Me Why It's Fake: Developing an Explainable User Interface for a Fake News Detection System*. In *Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence* (XAI.it '22) co-located with AIxIA '22. CEUR Workshop Proceedings, Vol. 3277, pp. 51-63.
17. **Purificato, E.**, Boratto, L., and De Luca, E.W. (2022). *Do Graph Neural Networks Build Fair User Models? Assessing Disparate Impact and Mistreatment in Behavioural User Profiling*. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (CIKM '22), pp. 4399-4403.
18. **Purificato, E.** (2022). *Beyond-Accuracy Perspectives on Graph Neural Network-Based Models for Behavioural User Profiling*. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '22), pp. 311-315.
19. Fallucchi, F., Di Stabile, R., **Purificato, E.**, Giuliano, R., and De Luca, E.W. (2022). *Enriching Videos with Automatic Place Recognition in Google Maps*. In *Multimedia Tools and Applications*, pp. 1-17.

20. **Purificato, E.**, Musto, C., Lops, P., and De Luca, E.W. (2022). *First Workshop on Adaptive and Personalized Explainable User Interfaces (APEx-UI 2022).* In *Companion Proceedings of the 27th International Conference on Intelligent User Interfaces* (IUI '22), pp. 1-3.

21. **Purificato, E.**, Wehnert, S., and De Luca, E.W. (2021). *Dynamic Privacy-Preserving Recommendations on Academic Graph Data.* In *Computers*, 10(9), pp. 107-132.

22. **Purificato, E.**, Manikandan, B.A., Karanam, P.V., Pattadkal, M.V., and De Luca, E.W. (2021). *Evaluating Explainable Interfaces for a Knowledge Graph-Based Recommender System.* In *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (IntRS '21) co-located with RecSys '21. CEUR Workshop Proceedings, Vol. 2948, pp. 73-88.

Human-Centred Artificial Intelligence
Otto von Guericke University Magdeburg