PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

SETIF 1 UNIVERSITY-FERHAT ABBAS



Faculty of Sciences

Department of Computer Science

Dissertation

Submitted in partial fulfillment of the requirements for Doctorate degree in Computer Science

Option: Data, Text and Web Mining

Presented by **Sabrina Karboua**

Algorithmic Bias and Fairness in Recommendation Systems

PhD defense board

President	Houceme Mansouri	Professor	Setif 1 University-Ferhat ABBAS
Director	Fouzi Harrag	Professor	Setif 1 University-Ferhat ABBAS
Co-Director	Farid Meziane	Professor	University of Derby, UK
Examiner	Makhlouf Derdour	Professor	University Of Oum El Bouaghi
Examiner	Hichem Talbi	MCA.	University of Constantine 2
Examiner	Samir Fenanir	MCA.	Setif 1 University-Ferhat ABBAS

Algeria, July 04,2026

Acknowledgements

بسم الله الرحمن الرحيم

والصلاة والسلام على أشرف المرسلين، سيدنا محمد وعلى آله وصحبه أجمعين
وما توفيقي إلا بالله عليه توكلت وإليه أنيب

The completion of this doctoral thesis marks the culmination of a challenging yet profoundly rewarding intellectual journey, one that would not have been possible without the guidance, support, and encouragement of numerous individuals to whom I owe an immense debt of gratitude.

First and foremost, I express my deepest appreciation to my supervisor, Professor Fouzi Harrag, whose support, guidance have been instrumental throughout this research journey.

I am equally indebted to my co-supervisor, Professor Farid Meziane, whose expertise and thoughtful insights have greatly enriched this work, and have significantly enhanced the quality of this thesis.

I extend my sincere gratitude to Professor Mohamed Deriche for his valuable contributions to this research. His constructive feedback has strengthened the theoretical foundations and practical implications of this work.

I thank the members of my thesis committee for their time, expertise, and valuable feedback. Their insightful comments have helped refine my work and broaden my understanding of the field.

My gratitude extends to my colleagues who have provided intellectual companionship throughout this journey. The discussions and collaborative exchanges we have shared have been sources of inspiration and motivation.

الحمد لله الذي بنعمته تتم الصالحات

Dedication

I dedicate this work to my parents: May they find here the testimony of my deep gratitude and acknowledgment

To my brother and my sisters, and my family who give love and liveliness.

To all those who have helped me - directly or indirectly - and those who shared with me the emotional moments during the accomplishment of this work and who warmly supported and encouraged throughout my journey.

To all my friends who have always encouraged me, and to whom I wish more success.

Thanks!

Sabrina Karboua

Abstract

Recommender systems (RSs) have become essential components of modern digital platforms, enabling personalized delivery of content, products, and services across domains such as e-commerce, streaming, social media, and online news. By leveraging large-scale user interaction data, they substantially influence information discovery and user engagement. However, RSs often inherit and amplify biases embedded in historical data, raising significant fairness concerns. These biases arise along two principal dimensions. On the user side, recommendations may become correlated with sensitive demographic attributes, such as gender, leading to stereotypical or discriminatory outcomes. On the item side, popularity bias reinforces the overexposure of highly popular items while systematically limiting the visibility of long-tail content, thereby reducing catalog diversity.

As inherently multi-stakeholder systems, recommender systems must balance the interests of users seeking relevant recommendations and item providers requiring equitable exposure. Consequently, fairness is fundamentally a two-sided problem. Existing approaches typically address either user-side or item-side fairness in isolation or apply fairness constraints only to recommendation outputs, leaving biased latent representations unchanged. Such representations continue to encode sensitive information, allowing bias to re-emerge. This dissertation argues that effective and durable two-sided fairness must be achieved at the representation level by learning unbiased user and item embeddings before recommendation generation.

This research proposes a representation-centric framework for two-sided fairness that directly debiases latent user and item representations. The proposed approach simultaneously mitigates user-side demographic bias and item-side popularity bias while preserving recommendation quality. Extensive experiments on multiple benchmark datasets demonstrate significant improvements in both fairness dimensions without compromising recommendation accuracy. The proposed approach consistently outperforms state-of-the-art fairness-aware recommender systems across a broad range of fairness metrics, highlighting the effectiveness of representation-level debiasing for developing fair, accurate, and balanced recommender systems.

Keywords: Recommender systems, algorithmic fairness, two-sided fairness, bias mitigation, disentangled representation learning, adversarial debiasing, variational autoencoders, orthogonal regularization, multi-stakeholder fairness, popularity bias.

ملخص

أصبحت أنظمة التوصية مكوّنًا أساسيًا في المنصات الرقمية الحديثة، حيث تُمكّن من تقديم محتوى ومنتجات وخدمات مخصّصة في مجالات متعددة، مثل التجارة الإلكترونية، ومنصات البث، ووسائل التواصل الاجتماعي، والأخبار الرقمية. ومن خلال الاستفادة من بيانات تفاعل المستخدمين على نطاق واسع، تؤثر هذه الأنظمة بصورة كبيرة في اكتشاف المعلومات وتعزيز تفاعل المستخدمين. إلا أنها غالبًا ما ترث التحيزات الكامنة في البيانات التاريخية وتُعيد إنتاجها أو تضخيمها، مما يثير تحديات جوهرية تتعلق بالعدالة. وتتجلى هذه التحيزات في بُعدين رئيسيين: فمن جهة المستخدم، قد ترتبط التوصيات بسمات ديموغرافية حساسة، مثل النوع الاجتماعي، بما يؤدي إلى نتائج نمطية أو تمييزية. ومن جهة العناصر الموصى بها، يؤدي تحيز الشعبية إلى تعزيز ظهور العناصر الأكثر شيوعًا على حساب عناصر الذيل الطويل، مما يحدّ من تنوع المحتوى.

وباعتبار أن أنظمة التوصية تعمل بطبيعتها ضمن بيئات متعددة أصحاب المصلحة، فإنها مطالبة بالموازنة بين مصالح المستخدمين الساعين إلى توصيات دقيقة وملائمة، ومصالح مزودي العناصر الذين يتطلعون إلى فرص عرض عادلة. ومن ثم، تُعد العدالة في هذه الأنظمة مسألة ثنائية الجانب بطبيعتها. وتعالج معظم الدراسات السابقة عدالة المستخدم أو عدالة العناصر كلٌّ على حدة، أو تفرض قيودًا على مخرجات التوصية فقط، مع الإبقاء على التمثيلات الكامنة المتحيزة دون معالجة. ونتيجة لذلك، تستمر هذه التمثيلات في ترميز المعلومات الحساسة، مما يسمح بعودة التحيزات. وتُظهر هذه الأطروحة أن تحقيق عدالة ثنائية مستدامة يتطلب إزالة التحيز مباشرةً من التمثيلات الكامنة للمستخدمين والعناصر قبل توليد التوصيات.

تقدم هذه الأطروحة إطارًا يرتكز على التمثيلات الكامنة لتحقيق العدالة ثنائية الجانب في أنظمة التوصية من خلال تعلم تمثيلات منزوعة التحيز لكل من المستخدمين والعناصر. ويسهم هذا النهج في الحدّ بالتزامن من التحيزات الديموغرافية المتعلقة بالمستخدم ومن تحيز الشعبية المرتبط بالعناصر، مع الحفاظ على جودة التوصيات. وتُظهر التجارب المكثفة على عدة مجموعات بيانات مرجعية تحقيق تحسينات ملحوظة في كلا بُعدي العدالة دون التأثير سلبًا في دقة التوصية. كما تتفوق النتائج باستمرار على أحدث الأساليب المعتمدة في الأدبيات وفقًا لمختلف مقاييس العدالة، مما يؤكد فعالية إزالة التحيز على مستوى التمثيلات في تطوير أنظمة توصية أكثر عدالة ودقة وتوازنًا.

الكلمات المفتاحية: أنظمة التوصية، العدالة الخوارزمية، العدالة ثنائية الجانب، تخفيف الانحياز، تعلم التمثيل، التعلم التنافسي، أجهزة التشفير التلقائي التباينية

Résumé

Les systèmes de recommandation (SR) sont devenus des composantes essentielles des plateformes numériques modernes, assurant la diffusion personnalisée de contenus, de produits et de services dans des domaines tels que le commerce électronique, la diffusion en continu, les réseaux sociaux et les médias en ligne. En exploitant des données d'interaction à grande échelle, ils influencent fortement la découverte de l'information et l'engagement des utilisateurs. Toutefois, ces systèmes héritent et amplifient souvent les biais présents dans les données historiques, soulevant d'importantes préoccupations en matière d'équité. Ces biais se manifestent selon deux dimensions principales. Du côté des utilisateurs, les recommandations peuvent être corrélées à des attributs démographiques sensibles, tels que le genre, conduisant à des résultats stéréotypés ou discriminatoires. Du côté des items, le biais de popularité favorise les contenus les plus populaires au détriment des items de longue traîne, réduisant ainsi la diversité du catalogue.

En tant que systèmes intrinsèquement multi-parties prenantes, les systèmes de recommandation doivent concilier les intérêts des utilisateurs, qui recherchent des recommandations pertinentes, et ceux des fournisseurs d'items, qui nécessitent une exposition équitable. L'équité constitue ainsi un problème fondamentalement bilatéral. Les approches existantes traitent généralement l'équité du côté des utilisateurs ou des items de manière indépendante, ou imposent des contraintes uniquement au niveau des recommandations produites, sans corriger les représentations latentes sous-jacentes. Ces représentations continuent alors d'encoder des informations sensibles, favorisant la réapparition des biais. Cette thèse montre qu'une équité bilatérale durable nécessite un débiaisage directement au niveau des représentations apprises.

Cette recherche propose un cadre centré sur les représentations pour assurer l'équité bilatérale dans les systèmes de recommandation en apprenant des représentations latentes débiaisées des utilisateurs et des items. L'approche proposée réduit simultanément les biais démographiques côté utilisateur et les biais de popularité côté item tout en préservant la qualité des recommandations. Des évaluations expérimentales menées sur plusieurs jeux de données de référence démontrent des améliorations significatives des deux dimensions de l'équité sans compromettre les performances de recommandation. Les résultats surpassent systématiquement les méthodes de l'état de l'art selon les différentes métriques d'équité, confirmant l'efficacité du débiaisage au niveau des représentations pour concevoir des systèmes de recommandation plus équitables, précis et équilibrés.

Mots-clés : Systèmes de recommandation, équité algorithmique, équité bilatérale, atténuation des biais, apprentissage de représentations, apprentissage adversarial, autoencodeurs variationnels, régularisation orthogonale, DS-DVAE, FUIREC, ORAF.

Contents

List of Figures 11

List of Tables 13

List of Abbreviations 15

Introduction 1

Context and Problem Statement . . . 1
Research Questions and Objectives . . . 3
Main Contributions . . . 6
Thesis Organization . . . 7

I State of the art 9

1 Theoretical and Methodological Foundations of Fair Recommendation Systems 10

1.1 Definition and Role of Recommender Systems . . . 10
1.2 Classical Approaches . . . 11
1.2.1 Collaborative Filtering . . . 12
1.2.2 Hybrid Approaches . . . 13
1.3 Deep-Learning–Based Recommender Systems . . . 13
1.3.1 Neural Collaborative Filtering . . . 14
1.3.2 The Wide & Deep Architecture . . . 14
1.3.3 Autoencoder-Based Recommenders . . . 15
1.3.4 Cold-Start Recommendation: DropoutNet, NLBA, MELU . . . 16
1.3.5 Other Deep Architectures . . . 18
1.4 Latent Representation Learning: The Embedding Space as a Fairness-Critical Object 18
1.4.1 What latent representations encode . . . 18
1.4.2 Latent entanglement and its consequences . . . 19
1.4.3 Latent representation learning architectures and their fairness implications . 19

1.4.4 Representation learning as the unifying framework . . . 20
1.5 Defining Bias and Fairness in Recommendation . . . 21
1.6 Sources, Causes, and a Taxonomy of Biases . . . 21
1.6.1 Data biases . . . 22
1.6.2 Model and algorithmic biases . . . 24
1.6.3 Result biases . . . 25
1.6.4 Feedback-loop bias . . . 26
1.7 Perspectives on Fairness . . . 26
1.7.1 Level perspective . . . 26
1.7.2 Subject perspective . . . 29
1.8 Fairness-Aware Methods: A Three-Stage View . . . 30
1.8.1 Pre-Processing Methods . . . 31
1.8.2 In-Processing Methods . . . 34
1.8.3 Post-Processing Methods . . . 42
1.9 Domains Affected by Bias and Unfairness . . . 43
1.9.1 Music recommendation . . . 44
1.9.2 Movie and book recommendation . . . 44
1.9.3 Employment and job recommendation . . . 44
1.9.4 News and content recommendation . . . 44
1.9.5 Educational (learning) recommendation . . . 45
1.9.6 Other high-stakes domains . . . 45
1.10 Summary . . . 45

2 Related Works on Two-Sided Fairness in Recommender Systems 47
2.1 Related Works on Mitigating User and Item Provider Demographic Bias . . . 47
2.1.1 Mitigating User Demographic Bias . . . 48
2.1.2 Mitigating Item Provider Demographic Bias . . . 50
2.2 Related Works on Mitigating Item Popularity Bias . . . 51
2.3 Two-Sided Stakeholder Fairness Frameworks . . . 53
2.3.1 Comparative Synthesis . . . 56
2.4 Gaps in the State of the Art and Thesis Motivations . . . 58
2.4.1 Absence of Representation-Level Two-Sided Stakeholder Fairness . . . 60
2.4.2 Dominance of Output-Level Intervention . . . 60
2.4.3 The Insufficiency of Adversarial Learning Alone and the Case for Hybrid Frameworks . . . 61
2.5 Summary . . . 63

II Contributions **64**

3 Dual-Sided Fair Recommendation via Variational Disentanglement **65**

3.1 Introduction . . . 65
3.2 Problem Formulation . . . 66
3.3 Fairness Definitions . . . 67
3.4 DS-DVAE Proposed Architecture . . . 67
3.4.1 User Disentangled Encoder . . . 68
3.4.2 Item Disentangled Encoder . . . 69
3.4.3 Task-Specific Recommendation Decoder . . . 70
3.4.4 Auxiliary Sensitive Attribute Decoders . . . 70
3.4.5 Training Objective . . . 71
3.4.6 Inference and Fair Recommendation Generation . . . 73
3.5 Experimental Setup . . . 73
3.5.1 Dataset Description . . . 73
3.5.2 VAE Backbone Models . . . 75
3.5.3 Evaluation Metrics . . . 77
3.6 Results . . . 78
3.6.1 Comparative Performance Evaluation . . . 78
3.6.2 Backbone-Agnostic Evaluation . . . 79
3.7 Group Fairness Analysis . . . 82
3.8 Embedding Visualization . . . 84
3.9 Summary . . . 86

4 A Unified Two-Sided Fairness Framework for Recommender Systems via Joint Adversarial Learning and Explicit Regularization **87**

4.1 Introduction . . . 88
4.2 Proposed Architecture . . . 92
4.2.1 Recommendation Model . . . 92
4.2.2 **Adversarial Discriminators** . . . 94
4.2.3 **Fairness Constraints Terms** . . . 95
4.2.4 **Objective Function of FUIREC** . . . 97
4.2.5 Adversarial Training: The MinMax Game . . . 97
4.3 Experiments . . . 99
4.3.1 Dataset Description . . . 100
4.3.2 Evaluation Metrics . . . 103
4.3.3 Baseline models . . . 104
4.4 Results and Discussion . . . 105

4.4.1 RQ1: Comparative Performance of FUIREC Across Fairness and Accuracy Metrics . . . 105
4.4.2 RS2: Impact of the Dual Adversarial Components . . . 107
4.4.3 RQ3: Impact of Popularity Regularization . . . 108
4.4.4 RQ4: Impact of the User Fairness Constraint . . . 109
4.4.5 Visualization of Learned Representations . . . 109
4.4.6 Hyperparameter Sensitivity and Trade-off Analysis . . . 111
4.5 Summary . . . 112

5 ORAF: An Orthogonal Regularization-Based Adversarial Framework for Achieving Two-Sided Fairness in Recommender Systems 113
5.1 Introduction . . . 113
5.2 The Proposed ORAF Framework . . . 115
5.2.1 Recommender Module . . . 116
5.2.2 Adversarial Discriminators . . . 121
5.2.3 Attribute Predictors for Bias-Explicit Embeddings . . . 122
5.2.4 Orthogonal Regularization . . . 124
5.3 ORAF Training . . . 125
5.3.1 Theoretical Framework and Objective . . . 125
5.3.2 Adversarial Training Analysis . . . 128
5.3.3 Optimization Dynamics and Convergence Analysis . . . 129
5.4 Experimental Details . . . 130
5.5 Evaluation Metrics . . . 132
5.5.1 Recommendation Quality . . . 132
5.5.2 User-side Fairness Metrics . . . 132
5.5.3 Item-side Fairness Metrics . . . 134
5.6 System Comparison Setup . . . 135
5.6.1 User-Side Fairness . . . 135
5.6.2 Item-side Fairness . . . 136
5.7 Experimental Results and Analysis . . . 136
5.8 Summary . . . 161

General Conclusion and Future Perspectives 162

Bibliography 3

List of Figures

1.1 Content-based filtering pipeline. . . . 11
1.2 Matrix factorization model . . . 13
1.3 Neural collaborative filtering . . . 14
1.4 Wide & Deep architecture . . . 15
1.5 VAE-based recommender . . . 16
1.6 DropoutNet Network . . . 17
1.7 MELU Architecture . . . 18
1.8 Taxonomy of biases in recommender systems . . . 22
1.9 Fairness perspectives in recommender systems. . . . 27
1.10 The three-stage view of fairness-aware methods. . . . 31
1.11 Pre-processing fairness-aware methods. . . . 32
1.12 Data re-sampling. . . . 32
1.13 Data re-labeling. . . . 33
1.14 Data augmentation. . . . 34
1.15 In-processing fairness-aware methods. . . . 34
1.16 Regularization-based fairness. . . . 36
1.17 Causal inference for debiasing. . . . 36
1.18 Adversarial learning architecture. . . . 37
1.19 Disentanglement-based debiasing approach. . . . 38
1.20 Orthogonal regularization approach . . . 39
1.21 Reinforcement-learning-based fairness-aware recommendation. . . . 40
1.22 Post-processing fairness-aware methods. . . . 42

3.1 Proposed Dual-Sided Disentangled VAE (DS-DVAE) architecture. . . . 68
3.2 User group fairness before and after DS-DVAE. . . . 83
3.3 Provider exposure before and after DS-DVAE (artists for Last.fm-360K, directors for MovieLens-1M). . . . 84

3.4 Embedding visualization before and after applying DS-DVAE on the MovieLens-1M dataset. . . . 85
3.5 Embedding visualization before and after applying DS-DVAE on the Last.fm-360K dataset. Top row: user latent space colored by user gender (blue = male, red = female). Bottom row: item latent space colored by artist gender (green = male, purple = female). . . . 85

4.1 Simplified Movie Recommendation Scenario Illustrating the Two-Sided Fairness Concept in RSs . . . 89
4.2 The overall architecture of FUIREC. . . . 93
4.3 Training dynamics of FUIREC across epochs. . . . 99
4.4 Rating Frequency and Popularity Threshold in Datasets. . . . 101
4.5 The effect of the individual fairness constraint across three benchmark datasets. . . . 110
4.6 **t-SNE visualization of learned representations using the ML100K dataset.** 110
4.7 Hyperparameter sensitivity analysis. . . . 112

5.1 Overview of ORAF Architecture . . . 117
5.2 ORAF Training Process. . . . 128
5.3 Gradient norm evolution per loss component across 50 training epochs on ML-100K. 129
5.4 Impact of Two-sided Fairness Mechanisms on Recommender Systems' Performance . 137
5.5 Performance comparison of the proposed ORAF framework across diverse recommendation architectures on the ML100k dataset . . . 149
5.6 Impact of discriminator update frequency on fairness-utility trade-off . . . 150
5.7 Learning rate optimization landscape for accuracy-fairness trade-offs on ML-100K dataset . . . 152
5.8 Hyperparameter sensitivity analysis . . . 153
5.9 ML-1M dataset . . . 156
5.10 ML-100K dataset . . . 157
5.11 BookCrossing dataset . . . 158
5.12 Study case of female user 366 . . . 160

List of Tables

2.1 Comparison of two-sided stakeholder fairness frameworks. 57
2.2 Datasets used in the evaluation of two-sided stakeholder fairness frameworks. 57
2.3 Evaluation metrics employed by two-sided stakeholder fairness frameworks. 58
2.4 Fairness criteria adopted by two-sided stakeholder fairness frameworks, decomposed by stakeholder side. 59

3.1 DS-DVAE hyperparameter configuration. 75
3.2 Comprehensive Performance Comparison: Baseline VAE vs. DS-DVAE at K=10. . . 78
3.3 Experimental results of different VAE backbones with (w) and without (w/o) the proposed DS-DVAE disentanglement applied. 81

4.1 Dataset Statistics 100
4.2 Performance comparison on ML-1M, BookCrossing, and ML-100k datasets. ↑: higher better, ↓: lower better. 106
4.3 Ablation study on dual adversarial discriminators (G: gender, P: popularity). Bold: best values. 108
4.4 Impact of the Popularity Regularization term. 109

5.1 Notation used in the ORAF framework. 118
5.2 Statistical Summary of Datasets Used 130
5.3 Comparison of User Side Performance on ML-1M, BookCrossing, and ML-100k. . . . 138
5.4 Ablation Study: Component-wise Analysis of User-Side Debiasing Mechanisms. . . . 140
5.5 Adversarial probing study: attribute inference AUC under nonlinear attack models (mean ± std over 5 seeds). 142
5.6 Performance comparison across demographic groups across five runs. 143
5.7 Item-side fairness performance comparison for popularity bias mitigation on ML-100K dataset. 144
5.8 Popularity-threshold sensitivity analysis: for ML-100K and ML-1M. 145
5.9 Ablation Study: Component-wise Analysis of Item-Side Debiasing Mechanisms. . . . 146

5.10 An ablation study on the role of attribute predictors in the orthogonal regularization. p-values from paired t-tests across 5 seeds. 148
5.11 Computational complexity comparison across model variants. 154
5.12 Quantitative embedding separation metrics for bias-aware and bias-free representations across all datasets. 159
A.1 Per-user deviation analysis over 20 randomly sampled users from ML-100K. 2
A.2 Aggregated deviation analysis over 20 randomly sampled users. 3
B.3 Component-wise ablation study on ML-100K over five independent seeds. p-val (vs Full): paired t-test against Full ORAF. p-val (vs Prev): paired t-test against immediately preceding variant. $p < 0.05$ (significant). n.s.: $p \geq 0.05$ (not significant). . . 2

List of Abbreviations

ASS : Attribute Substitution Stability
AUC : Area Under the Curve
BCE : Binary Cross-Entropy
CF : Collaborative Filtering
CNN : Convolutional Neural Network
CPFair : Consumer/Producer Fairness Re-ranking
DBI : Davies-Bouldin Index
DeepFM : Deep Factorization Machine

DL : Deep Learning
DNN : Deep Neural Network
DP : Demographic Parity
DS-DVAE : Dual-Sided Disentangled VAE
FDA : Fairness-oriented Data Augmentation
GB : Gradient Boosting
GF : Group Fairness
ICD : Inter-Centroid Distance
IPW : Inverse Propensity Weighting
ITFR : Intersectional Two-sided Fairness
LLM : Large Language Model
MAE : Mean Absolute Error

MF : Matrix Factorization
MLP : Multi-Layer Perceptron
MMS : Maximin Share
NCF : Neural Collaborative Filtering
NDCG : Normalized Discounted Cumulative Gain
NeuMF : Neural Matrix Factorization
NLBA : Non-Linear Bias Adaptation
NLP : Natural Language Processing
NN : Neural Network
ORAF : Orthogonal Reg based Adversarial Framework
PMF : Probabilistic Matrix Factorization
PPR : Pairwise Preference Regression
RecSys : Recommender Systems
REO : Ranking-based Equal Opportunity
RF : Random Forest
RL : Reinforcement Learning
RMSE : Root Mean Square Error
RS : Recommendation System
RSP : Ranking-based Statistical Parity
SS : Silhouette Score
SVM : Support Vector Machine
t-SNE : t-Distributed Stochastic Neighbor Embedding
TFROM :Two-sided Fairness-aware Rec. Model
UGF : User Group Fairness

Introduction

Context and Problem Statement

Recommender systems (RSs) have become an indispensable component of contemporary digital infrastructures. By exploiting interaction histories, content features, and contextual signals, they distill massive item catalogs into compact personalized lists tailored to each user's preferences [1, 2]. Their footprint extends across e-commerce platforms, online learning environments, news outlets, social networks, and streaming services, where they actively shape what users discover, learn, purchase, and consume. Beyond information filtering, modern recommender pipelines increasingly integrate deep representation learning, self-attention, and graph-based propagation, which have substantially improved predictive accuracy but have also deepened the opacity of the underlying decision processes [3, 4].

This rise in scale and influence has been accompanied by an increasingly sharp awareness that recommendation outputs are far from neutral. A growing body of recent work demonstrates that RSs systematically encode, reproduce, and at times amplify the biases present in the training data, in the optimization objective, and in the closed-loop dynamics that drive long-term retraining [1, 5, 6]. The consequences of such biases are wide-ranging. Economically, marginalized content providers see their visibility and revenue diminish; socially, certain demographic groups receive narrower or more stereotyped suggestions that reinforce pre-existing inequities; ecologically, the catalog diversity of the platform contracts under the pressure of self-reinforcing popularity dynamics [7, 8, 9].

Bias in recommendation pipelines decomposes naturally along the two-sided structure of the platforms in which RSs operate, and is therefore conventionally studied from the user side and the item side [10, 4].

On the user side, bias arises whenever the quality of a recommendation depends systematically on a user's demographic attributes, such as gender, age, ethnicity, geography, or socio-economic position, rather than purely on their declared preferences. The most extensively documented manifestation is user demographic bias, in which protected groups, most commonly female users or users from minority backgrounds, receive recommendations of lower relevance, narrower scope, or stronger stereotype intensity than users in the dominant group, even when the two cohorts express equivalent preferences [11]. The consequences extend well beyond degraded recommendation quality.

At the individual level, users from protected groups are systematically steered away from relevant opportunities, whether career prospects, educational resources, or financial products, reinforcing pre-existing socioeconomic disadvantages and eroding user trust in the platform [12]. At the societal level, these patterns constitute a form of algorithmic discrimination that scales with platform reach: as RSs mediate access to information for hundreds of millions of users, the cumulative effect of individually small demographic disparities amounts to a structural entrenchment of inequality that persists and intensifies across retraining cycles [13].

On the item side, bias concerns the exposure that items, products, or content providers obtain through the recommendation channel. The most pervasive phenomenon is item popularity bias [2, 8], which is the systematic over-recommendation of a small set of already-popular items at the expense of the long tail, regardless of whether the suppressed items would have been relevant to the user. The consequences are equally severe. For content providers, sustained low visibility translates directly into lost revenue and market exclusion, creating a structurally unequal competitive environment that favors incumbents over newcomers regardless of quality [9]. For users, the homogenization of recommendation lists reduces serendipity and constrains content diversity, progressively narrowing the informational horizon of the platform [14]. At the ecosystem level, the winner-takes-all dynamic that popularity bias engenders causes catalog diversity to collapse over successive feedback cycles, ultimately undermining the long-term sustainability and richness of the platform itself [15, 16].
Despite the conceptual symmetry between the two sides, the vast majority of fairness-aware recommendation methods address only one of them at a time. Approaches that enforce demographic parity on the consumer side typically leave item exposure untouched, and approaches that mitigate popularity bias on the provider side rarely examine how the redistributed exposure interacts with the demographic groups of users that ultimately receive it [1, 3]. This one-sided perspective sits uneasily with the fundamental architecture of modern recommendation platforms, which operate as multi-stakeholder ecosystems whose long-term health depends jointly on the satisfaction of users and providers [10, 17]. Recent work has formalized this by framing the platform as a two-sided market in which producer visibility and consumer relevance are jointly determined by the recommendation policy, and has shown that treating fairness as a unilateral objective is both theoretically unsound and practically counterproductive [18]. Worse, recent empirical analyzes show that improving fairness on a single side frequently shifts the harm to the other: correcting user-group disparities tends to amplify popularity concentration, and re-balancing item exposure can introduce new disparities between user groups [8, 19]. The difficulty is compounded over time: longitudinal analyzes of interactive recommendation demonstrate that accuracy and fairness objectives that appear compatible in a single training round progressively diverge as feedback loops accumulate, so that a system that satisfies both constraints at deployment may violate one of them after several retraining cycles [20].

Existing two-sided fairness frameworks have predominantly addressed this problem either through post-processing re-ranking, which adjusts the output list after the model has already generated its

predictions, or through constrained optimization formulations that incorporate fairness penalties into the training objective [11, 21]. While these approaches have demonstrated improvements in fairness metrics, they share a fundamental limitation: they intervene on the recommendation output without modifying the underlying representations, meaning that the learned embeddings remain entangled with sensitive attributes and popularity signals. As a result, the bias is suppressed at the surface level but persists structurally in the latent space, from which it can re-emerge under distributional shift or after retraining [22, 23]. In this thesis, we argue that durable two-sided fairness requires intervening at the representation level itself, so that the learned user and item embeddings are structurally purged of sensitive information before any recommendation is generated.

These observations motivate the central perspective adopted in this thesis: two-sided fairness. Two-sided fairness designates a recommendation regime that simultaneously enforces (i) equitable treatment of users across demographic groups, so that no protected group is structurally disadvantaged, and (ii) equitable exposure opportunities for items across popularity tiers and provider groups, so that relevant but less popular items receive a share of visibility proportional to their merit. The objective is not the mechanical equalization of outcomes but the construction of a recommendation regime in which neither stakeholder side is penalized by the optimization process itself.

Research Questions and Objectives

Despite the increasing attention devoted to fairness-aware recommendation, existing approaches remain limited in several important respects. First, many studies focus exclusively on a single fairness perspective, typically emphasizing either user-side fairness or item-side fairness independently. Second, conventional debiasing strategies often suppress observable bias patterns without fully eliminating sensitive information leakage from latent embeddings. Third, the majority of existing frameworks fail to jointly optimize recommendation utility and multi-stakeholder fairness within a unified representation learning framework.

Motivated by these limitations, this thesis investigates how advanced representation learning and adversarial optimization strategies can contribute to the development of recommendation systems that are simultaneously accurate, robust, and fair across multiple stakeholders.

To guide this investigation, the thesis addresses the following research questions.

Main Research Question

MRQ

How can recommendation systems achieve effective multi-stakeholder fairness while preserving recommendation quality through disentangled and adversarial representation learning frameworks?

Research Questions

RQ1 — Bias Representation and Propagation How do latent representations learned by modern recommendation systems encode, propagate, and amplify demographic, behavioral, and popularity-related biases across users and items?

RQ2 — Disentangled Fair Representation Learning To what extent can disentangled representation learning separate recommendation-relevant information from sensitive attributes in order to improve fairness across multiple stakeholders?

RQ3 — Adversarial and Orthogonal Debiasing Can adversarial learning combined with orthogonal regularization provide stronger protection against sensitive information leakage than conventional fairness-aware recommendation approaches?

RQ4 — Multi-Stakeholder Fairness Optimization How can user-side fairness and item-side fairness be jointly optimized without significantly degrading recommendation accuracy and personalization effectiveness?

RQ5 — Fairness–Utility Trade-Offs and Generalization What trade-offs emerge between recommendation utility, fairness, robustness, and representation disentanglement across different datasets and recommendation scenarios?

Research Objectives

To address these research questions, this thesis pursues the following objectives:

Objective 1 — Analytical Investigation of Bias in Recommendation Systems To conduct a comprehensive analysis of algorithmic bias in recommendation systems by examining:

- the origins of unfairness,
- the mechanisms of bias propagation,
- the impact of latent representation learning on fairness outcomes.

This objective establishes the theoretical and conceptual foundations of the thesis.

Objective 2 — Development of a Dual-Sided Fairness Framework To design and develop a fairness-aware recommendation architecture capable of simultaneously addressing user-side fairness and item-side fairness through disentangled latent representation learning. This objective motivates the proposed **DS-DVAE** framework.

Objective 3 — Representation Disentanglement for Fair Recommendation To investigate how disentangled variational learning can isolate sensitive demographic information from recommendation-relevant latent factors in order to reduce indirect discrimination and hidden bias leakage.

Objective 4 — Adversarial Orthogonal Debiasing To develop an adversarial fairness framework integrating orthogonal regularization mechanisms capable of enforcing stronger separation between sensitive and task-relevant representations. This objective forms the theoretical foundation of the **ORAF** framework.

Objective 5 — Experimental Evaluation and Comparative Analysis To evaluate the proposed fairness-aware recommendation frameworks using:

- recommendation quality metrics,
- fairness evaluation measures,
- robustness analysis,
- ablation studies,
- latent-space visualization techniques.

Objective 6 — Advancement of Multi-Stakeholder Fairness in Recommendation Systems To contribute toward the development of more transparent, equitable, and trustworthy recommendation systems by proposing unified fairness-aware learning strategies applicable to real-world recommendation environments.

Building upon these research questions and objectives, the thesis introduces a set of methodological and theoretical contributions aimed at advancing fairness-aware recommendation through disentangled and adversarial representation learning. These contributions collectively establish the foundation for the proposed multi-stakeholder fairness frameworks developed throughout the dissertation.

Main Contributions

To address these questions, the thesis develops three hybrid in-processing frameworks. All three operate directly on the learned user and item embedding spaces, rather than on the input data or on the output list, so that fairness becomes a structural property of the representations themselves and not a corrective patch applied after the fact.

First contribution (Chapter 3): DS-DVAE (Dual-Sided Disentangled Variational Autoencoder). This framework ensures two-sided fairness for both consumer and item provider by eliminating the influence of sensitive attributes from the recommendation process through architectural design. DS-DVAE explicitly factorises user and item embeddings into two orthogonal subspaces: a task subspace that captures preference-relevant information, and a sensitive subspace that isolates demographic information (user gender and provider gender). Recommendations are generated solely from the task subspace, so sensitive attributes from both consumers and providers have no path to influence the output. The framework is instantiated as a variational autoencoder with factorised posteriors, auxiliary attribute decoders that supervise the sensitive subspace, and an orthogonality constraint that reinforces separation.

Second contribution (Chapter 4): FUIREC. This framework combines adversarial learning with fairness-constraint-based regularization. Two adversarial discriminators are deployed to recover the sensitive attributes (user gender on the consumer side, item popularity tier on the provider side) from the learned embeddings, while the recommender simultaneously minimizes the recommendation loss and a set of explicit fairness regularizers that bound demographic disparity on the user side and exposure imbalance on the item side. The framework provides a lightweight and easily tunable mechanism for simultaneous two-sided debiasing.

Third contribution (Chapter 5): ORAF. This framework pairs adversarial learning with orthogonal regularization. In addition to the adversarial debiasing game, every user and item embedding is factorized into a bias-aware component and a debiased component, and an orthogonality penalty drives the two components into mutually perpendicular subspaces. Because this orthogonality term is a closed-form, differentiable quantity, it offers a verifiable geometric guarantee that sensitive information cannot leak from the bias-aware subspace into the debiased subspace, an assurance that adversarial training alone, which only enforces distributional indistinguishability, cannot provide [24]. ORAF serves as an enhancement to FUIREC, replacing explicit output-level constraints with orthogonal regularization that operates directly within the latent space.

All three frameworks are evaluated comprehensively on benchmark datasets and against state-of-the-art two-sided fairness baselines, and are shown to reduce both user- and item-side bias substantially while incurring only a marginal cost in predictive accuracy.

Thesis Organization

The remainder of the dissertation is organized as follows. Chapter 1 introduces recommender systems formally and reviews the algorithmic families, including classical collaborative filtering, deep neural recommenders, and variational autoencoders, on which the proposed contributions are built, then develops the foundations of bias and fairness in recommendation: it presents an expanded taxonomy of bias types and locates each one on the recommendation lifecycle, articulates the individual/group and consumer/provider/two-sided fairness perspectives, and surveys the three families of mitigation methods. Chapter 2 presents a focused comparative review of two-sided stakeholder fairness frameworks, organized around consumer demographic bias, item popularity bias, and joint two-sided fairness. The chapter identifies three structural gaps in the existing literature — the absence of representation-level two-sided fairness, the dominance of output-level post-processing, and the insufficiency of adversarial learning alone, that directly motivate the three proposed contributions

Chapter 3 presents the first contribution, DS-DVAE, a dual-sided disentangled variational autoencoder that achieves two-sided fairness by architectural design through subspace factorisation, including formal problem formulation, architectural design, and experimental evaluation on Last.fm-360K and and MovieLens-1M.

Chapter 4 presents the second contribution, FUIREC, combining adversarial learning with explicit fairness constraints (individual user fairness and popularity regularizer), including formal problem formulation, architectural design, training procedure, and full experimental evaluation on ML-1M, ML-100K, and BookCrossing.

Chapter 5 presents the third contribution, ORAF, integrating adversarial learning with orthogonal regularization as an enhancement to FUIREC, including theoretical analysis of the geometric guarantee, complete loss formulation, ablation studies, and comparative evaluation against multiple two-sided fairness baselines across three datasets and nine recommendation architectures.

Chapter 6 concludes the thesis, summarizes the contributions, discusses limitations and ethical considerations, and outlines directions for future work on multi-stakeholder fairness in algorithmic decision systems.

Publications

The research presented in this thesis has resulted in the following publications:

- Karboua, S., Harrag, F., & Meziane, F. (2026). A Dual-Adversarial Learning Framework with Explicit Constraints for Two-Sided Fairness in Recommendation Systems. Open Information Science. (published)

- Karboua, S., Harrag, F., & Meziane, F. (2024, November). Multi-Head Self-Attention Based Arabic News Recommendation System.
- Karboua, S., Harrag, F., Meziane, F., & Boutadjine, A. (2022, December). Mitigation of Popularity Bias in Recommendation Systems. In TACC (pp. 69-80). (published)
- Karboua, S., Harrag, F., Meziane, F., & Deriche, M.(2026). ORAF : An Orthogonal Regularization-Based Adversarial Framework for Achieving Two-Sided Fairness in Recommendation Systems. Scientific Reports journal (published).
- Karboua, S., Harrag, F., & Meziane, F. (2026) Dual-Sided Fair Recommendation via Variational Disentanglement. 5th International Conference on Artificial Intelligence Science and Applications in Industry and Society 2026 (CAISAIS). (accepted).
- Karboua, S., Harrag, F., Meziane, F., & Deriche, M.(2026) (under review). Mitigating Bias and Ensuring Fairness in LLM-Based Recommendation Systems.Applied Intelligence journal (under review).

Part I

State of the art

Chapter 1

Theoretical and Methodological Foundations of Fair Recommendation Systems

This chapter establishes the theoretical and methodological foundations upon which the contributions of this thesis are built. It is organized into two complementary parts. The first part introduces recommender systems formally, surveying the principal algorithmic families and devoting particular attention to the latent embedding space as a fairness-critical object, arguing that learned representations inherently encode sensitive information alongside task-relevant preference signals, and that this entanglement constitutes the root cause of the bias phenomena addressed throughout this dissertation. The second part systematically examines bias and fairness in recommender systems, presenting a taxonomy of bias types organized across the recommendation lifecycle, articulating the principal fairness perspectives, and surveying the three families of fairness-aware mitigation methods. Together, these two parts provide the conceptual vocabulary and technical background necessary to situate the proposed contributions within the broader landscape of fair recommendation research.

1.1 Definition and Role of Recommender Systems

A recommender system (RS) is an information-filtering tool whose purpose is to identify, from a typically very large catalog of items, a small subset that is most likely to be of interest to a given user [25]. Formally, given a set of users $U = \{u_1, \ldots, u_n\}$ and a set of items $I = \{i_1, \ldots, i_m\}$, a recommender system learns a scoring function $f : UI \to \mathbb{R}$ that estimates the relevance of each item to each user. The recommendation task is then to return, for every user u, a top-K list of items ranked by $f(u, \cdot)$. This abstraction is deceptively simple , its instantiations span an enormous design space, from purely heuristic neighbourhood methods to deep neural architectures with billions of

parameters.

Recommender systems have become indispensable components of modern digital platforms, mediating user access to products, audiovisual content, social media feeds, educational materials, and even employment opportunities [26]. Their economic and social influence is now widely acknowledged: large platforms attribute a substantial fraction of their user engagement and revenue to algorithmically curated recommendations [12].

1.2 Classical Approaches

Content-Based Filtering

Content-based filtering (CBF) recommends items to a user based on the characteristics of items that the user has previously interacted with, rather than on the behavior of other users [25]. The underlying assumption is that a user who liked a particular item in the past will tend to like items that are similar to it in terms of observable features. Figure 1.1 illustrates the full content-based filtering pipeline.

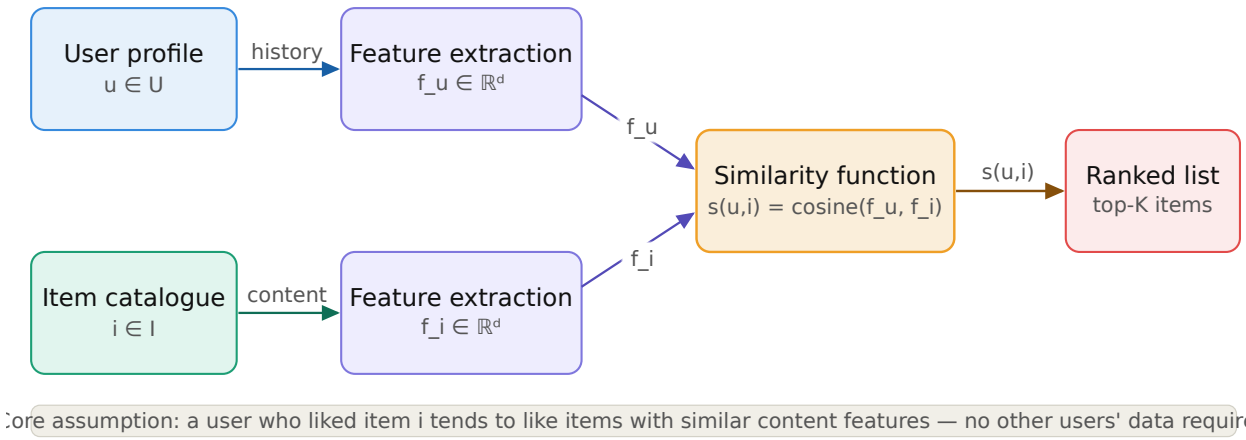


Figure 1.1: Content-based filtering pipeline.

Each item $i \in \mathcal{I}$ is described by a feature vector $\mathbf{f}_i \in \mathbb{R}^p$, where p is the number of features. In text-based domains such as news, books, and research papers, these features are typically TF-IDF-weighted term frequencies or dense sentence embeddings; in music recommendation they may include acoustic descriptors, genre tags, or artist metadata; in movie recommendation they include genre, director, cast, and plot-derived keywords. A user profile $\mathbf{p}_u \in \mathbb{R}^p$ is then constructed by aggregating the feature vectors of items the user has rated positively, commonly as a weighted

average:

$$\mathbf{p}_u = \frac{\sum_{i \in \mathcal{I}_u} r_{ui}\, \mathbf{f}_i}{\sum_{i \in \mathcal{I}_u} r_{ui}}, \tag{1.1}$$

where $\mathcal{I}_u$ is the set of items rated by user u and r_{ui} is the corresponding rating. The predicted relevance of an unseen item j for user u is then estimated via cosine similarity:

$$\hat{r}_{uj} = \cos(\mathbf{p}_u,\, \mathbf{f}_j) = \frac{\mathbf{p}_u^\top \mathbf{f}_j}{\|\mathbf{p}_u\|\, \|\mathbf{f}_j\|}. \tag{1.2}$$

1.2.1 Collaborative Filtering

Collaborative filtering (CF), the most influential family of recommendation algorithms, exploits the principle that users who agreed in the past will tend to agree in the future. CF methods divide into two main subfamilies [27].

Memory-based CF

Memory-based (neighbourhood-based) CF computes similarities between users (user-based CF) or between items (item-based CF) and predicts ratings as weighted averages over the k nearest neighbours. The user-based prediction takes the form

$$\hat{r}_{u,i} = \frac{\sum_{v \in \mathcal{N}_k(u)} \mathrm{sim}(u, v)\, r_{v,i}}{\sum_{v \in \mathcal{N}_k(u)} \mathrm{sim}(u, v)}, \tag{1.3}$$

where $\mathcal{N}_k(u)$ is the set of k nearest neighbours of user u. While interpretable, memory-based approaches scale poorly with the number of users and items and degrade rapidly under data sparsity.

Model-based CF: matrix factorization

Model-based CF, exemplified by matrix factorization (MF), learns low-dimensional latent representations of users and items by decomposing the partial user–item rating matrix $R \in \mathbb{R}^{nm}$ into the product of two low-rank matrices: $R \approx UV^\top$, where $U \in \mathbb{R}^{nd}$ and $V \in \mathbb{R}^{md}$ encode user and item embeddings respectively in a shared d-dimensional latent space [28]. The predicted score for user u and item i is

$$\hat{r}_{u,i} = \mathbf{u}_u^\top \mathbf{v}_i + b_u + b_i + \mu, \tag{1.4}$$

where b_u and b_i are user and item biases and μ is the global mean rating. Parameters are learned by minimizing a regularized squared-error loss over observed entries

$$\mathcal{L} = \sum_{(u,i) \in \Omega} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda \left(\|\mathbf{u}_u\|^2 + \|\mathbf{v}_i\|^2 \right), \tag{1.5}$$

typically through stochastic gradient descent or alternating least squares. Variants such as Probabilistic Matrix Factorization [29] and Bayesian Personalized Ranking [30] have established matrix

factorization as the de-facto baseline of modern recommender research. Figure 1.2 illustrates the matrix factorization principle. Critically for this thesis, the latent embeddings produced by matrix factorization can encode sensitive demographic information about users and popularity-related signals about items , these are precisely the dimensions of the embedding space that our fairness frameworks intervene upon.

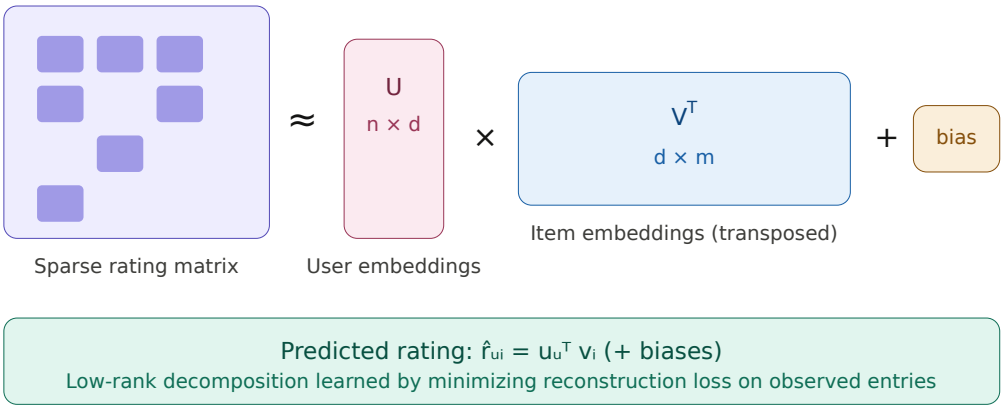


Figure 1.2: Matrix factorization model

1.2.2 Hybrid Approaches

Hybrid recommenders combine content-based and collaborative signals to mitigate the weaknesses of each: cold-start of CF and over-specialization of content-based filtering [31]. Combinations may take the form of weighted ensembles, switching policies, feature augmentation, or unified probabilistic models such as factorization machines [32], which generalize matrix factorization to arbitrary feature interactions.

1.3 Deep-Learning–Based Recommender Systems

The mid-2010s rise of deep neural networks brought a step-change in recommender system research [33]. By replacing the linear inner product of matrix factorization with non-linear neural architectures, deep models can capture complex high-order interactions between users, items, and context. We focus here on the four families that are most relevant to this thesis: neural collaborative filtering, the Wide & Deep architecture, autoencoder/VAE-based models, and meta-learning models for cold-start recommendation.

1.3.1 Neural Collaborative Filtering

Neural Collaborative Filtering (NCF) [34] replaces the inner product of matrix factorization with a multi-layer perceptron (MLP) applied to the concatenation of user and item embeddings. Each user u is represented by an embedding $\mathbf{e}_u \in \mathbb{R}^d$ and each item i by $\mathbf{e}_i \in \mathbb{R}^d$, both initialized from one-hot identifier inputs. The predicted interaction score is

$$\hat{r}_{u,i} = \sigma(W_o \mathbf{h}_n + \mathbf{b}_o), \quad \mathbf{h}_k = \mathrm{ReLU}(W_k \mathbf{h}_{k-1} + \mathbf{b}_k), \quad \mathbf{h}_0 = [\mathbf{e}_u \,\|\, \mathbf{e}_i]. \tag{1.6}$$

The model is trained end-to-end by minimizing a binary cross-entropy or pairwise ranking loss. NCF and its variants , NeuMF (which fuses generalized matrix factorization with the MLP path), MLP-CF, and DeepFM, have become standard backbones in fairness-aware recommendation. Importantly, the MLP-based architecture used in our **FUIREC** contribution (Chapter 4) follows directly from this paradigm: the recommendation backbone is an NCF-style model, while the dual-adversarial discriminators and explicit fairness constraints are appended without altering the underlying NCF prediction pipeline. Figure 1.3 depicts the canonical NCF architecture.

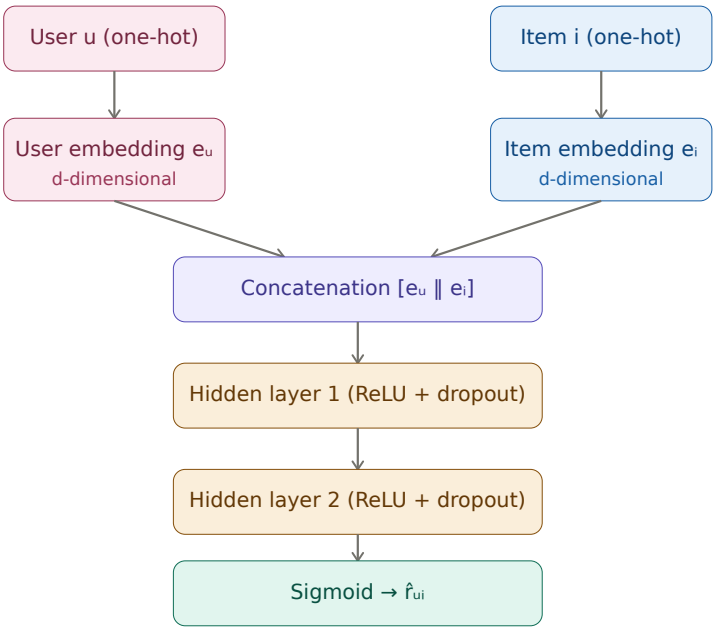


Figure 1.3: Neural collaborative filtering

1.3.2 The Wide & Deep Architecture

Wide & Deep [35] is a hybrid neural recommender that combines a memorization branch (the Wide component) with a generalization branch (the Deep component). The Wide component is a generalized linear model operating on sparse cross-product features, designed to memorize frequent

user–item co-occurrences. The Deep component is a feed-forward network operating on dense embeddings of categorical features, designed to generalize to unseen feature combinations. The two branches are jointly trained to produce a single prediction

$$P(y = 1 \mid \mathbf{x}) = \sigma\Big(\mathbf{w}_w^\top \phi(\mathbf{x}) + \mathbf{w}_d^\top \mathbf{h}_n + b\Big)\,, \tag{1.7}$$

where $\phi(\mathbf{x})$ is the cross-product feature transformation (Wide branch) and $\mathbf{h}_n$ is the last hidden layer of the Deep branch. Wide & Deep is widely deployed at industrial scale (notably by Google for the Play Store). Figure 1.4 illustrates the architecture.

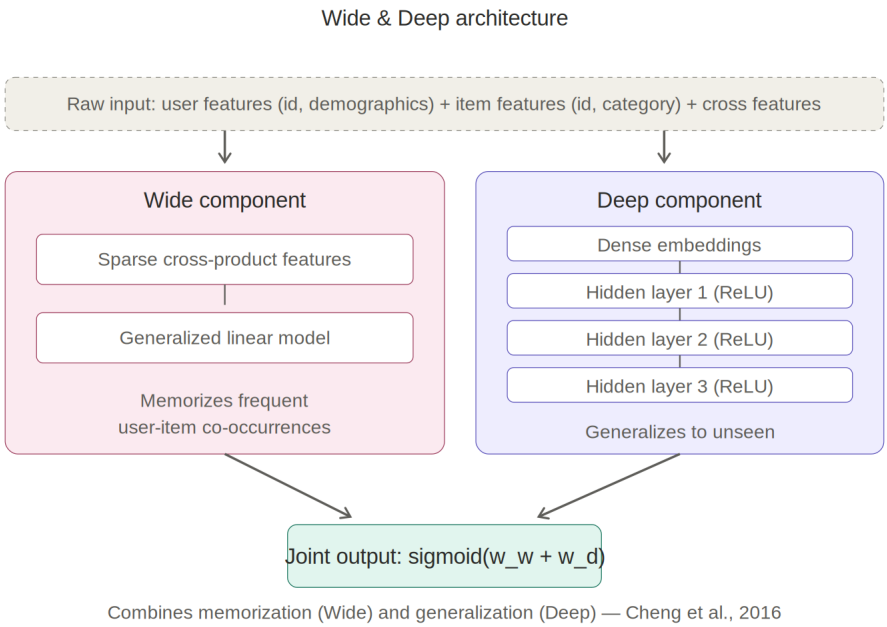


Figure 1.4: Wide & Deep architecture

1.3.3 Autoencoder-Based Recommenders

Autoencoders learn compressed representations of user–item interaction vectors. AutoRec [36] trains a feed-forward autoencoder to reconstruct the rating vector of each user (or item), and Collaborative Denoising Autoencoders [37] extend this idea by injecting noise into the input vector to improve robustness. The most influential variant for our purposes is the Multinomial Variational Autoencoder for collaborative filtering [38], which models user representations probabilistically through a latent Gaussian distribution.

In a VAE-based recommender, each user u is associated with an interaction vector $\mathbf{x}_u \in \mathbb{R}^m$. An encoder network maps $\mathbf{x}_u$ to the parameters of a Gaussian posterior $q_\phi(\mathbf{z} \mid \mathbf{x}_u) = \mathcal{N}(\boldsymbol{\mu}_u, \mathrm{diag}(\boldsymbol{\sigma}_u^2))$, and a decoder reconstructs item probabilities $p_\theta(\mathbf{x} \mid \mathbf{z})$. The model is trained by maximizing the

evidence lower bound

$$\text{ELBO} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta\, \text{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z}))\,, \tag{1.8}$$

where $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$ is a standard Gaussian prior. The reparameterization trick $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I)$ enables end-to-end gradient-based training [39]. Figure 1.5 illustrates the canonical VAE recommender architecture.

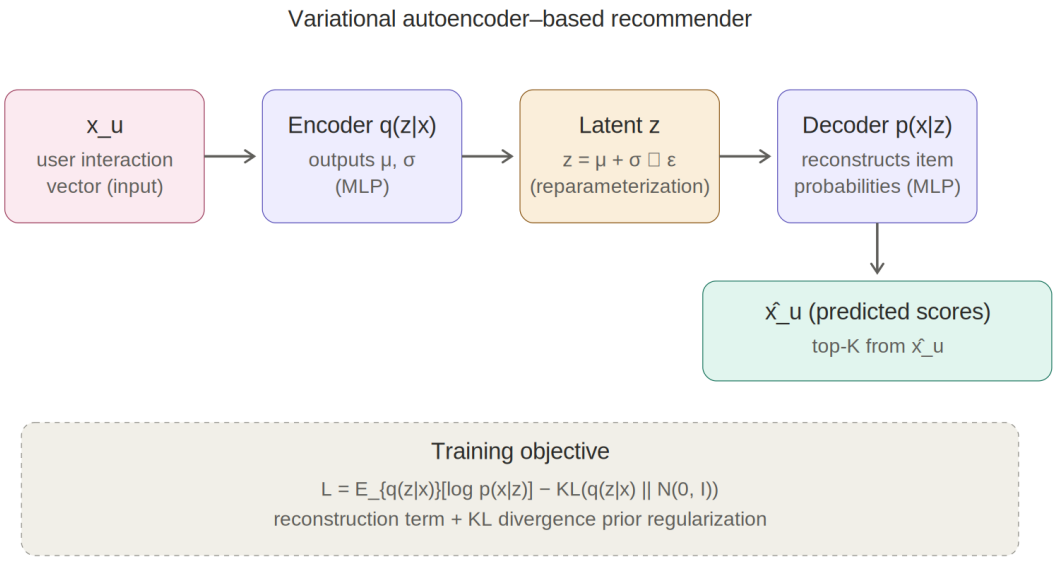


Figure 1.5: VAE-based recommender

VAEs are particularly relevant to this thesis because they form the architectural foundation of our **DS-DVAE** contribution (Chapter 3). DS-DVAE extends the standard VAE by introducing a *factorized* posterior that decomposes the latent space into a task subspace (preferences / item characteristics) and a sensitive subspace (user gender / provider gender). Recommendations are generated only from the task subspace, achieving fairness by architectural design rather than by post-hoc correction.

1.3.4 Cold-Start Recommendation: DropoutNet, NLBA, MELU

Cold-start recommendation , producing useful recommendations for new users or new items with little or no interaction history, is a long-standing challenge [40].

DropoutNet [41]. DropoutNet treats cold-start as a problem of training-time data augmentation. The model takes pre-trained CF embeddings of users (and items) together with side information (demographics, content features) and applies dropout to the CF embeddings during training,

randomly zeroing them out. This forces the encoder to learn to reconstruct the CF representation purely from the side information whenever the CF input is missing , exactly the cold-start regime at inference time. Figure 1.6 illustrates the principle.

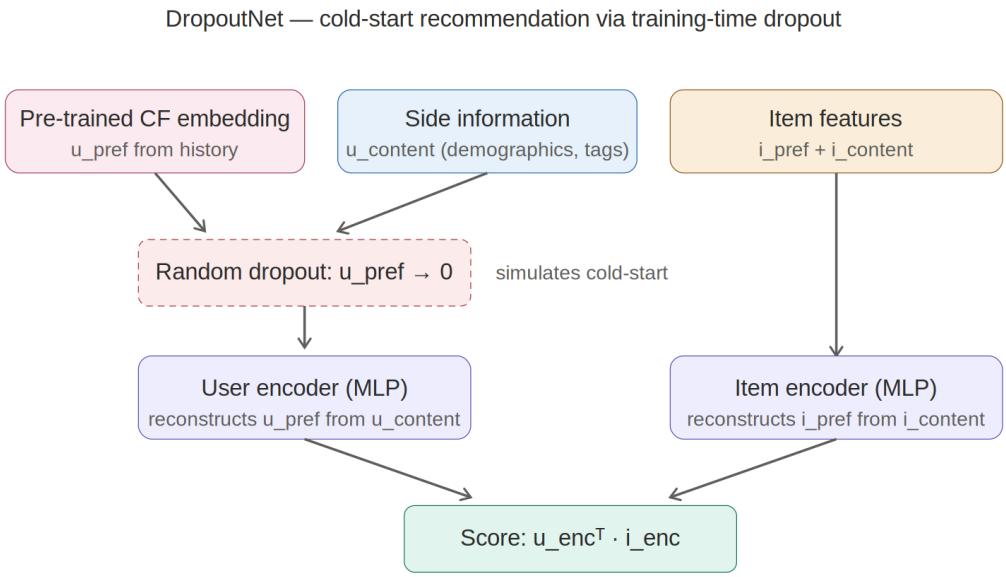


Figure 1.6: DropoutNet Network

NLBA, Non-Linear Bias Adaptation [42]. Neural Network with Bias Adaptation (NLBA) is a cold-start architecture in which the weights of the base network are shared across all users, while only the bias terms of the output and hidden layers are personalized for each user. The shared weights act as a global preference model, while user-specific biases provide a low-capacity, fast-adapting personalization mechanism. Because only a small number of parameters needs to be estimated per user, NLBA can produce useful recommendations even for users with very few interactions.

MELU, Meta-Learned User Preference Estimator [43]. MELU casts cold-start recommendation as a meta-learning problem in the spirit of Model-Agnostic Meta-Learning (MAML). An outer loop trains a global initialization θ_0 across many users, and an inner loop adapts θ_0 to each new user u via a small number of gradient steps on a support set $\mathcal{S}_u$. Unlike NLBA, MELU adapts the full set of task-network parameters per user; the outer loop ensures that θ_0 is positioned in parameter space such that these few-shot updates produce accurate user-specific predictions. Figure 1.7 illustrates the two-loop optimization.

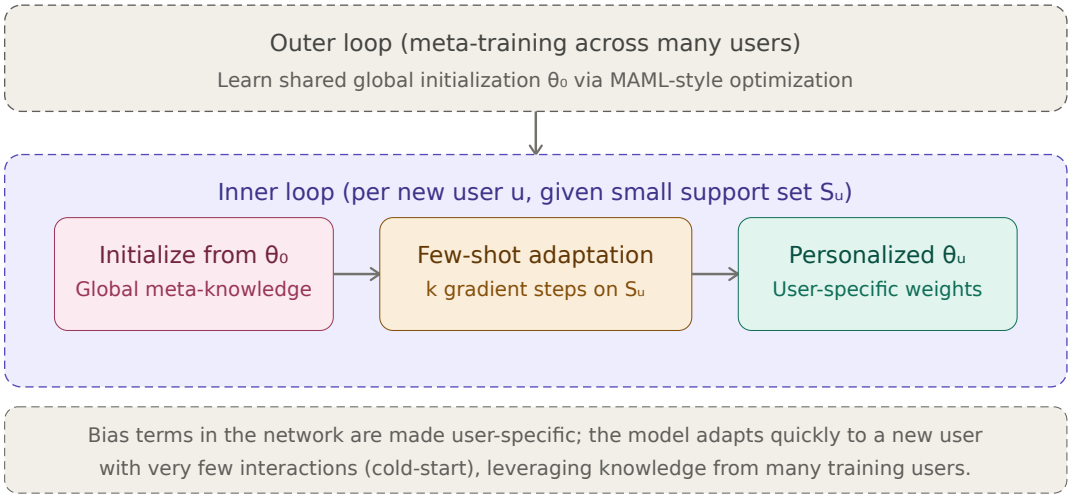


Figure 1.7: MELU Architecture

1.3.5 Other Deep Architectures

Several deep recommendation architectures fall outside the scope of our backbone choices but are worth mentioning. Recurrent [44] and self-attention [45, 46] models capture the temporal ordering of user interactions and dominate session-based recommendation. Graph Neural Network recommenders such as NGCF [47], LightGCN [48] and PinSage [49] treat the user-item bipartite graph as a first-class object, propagating embeddings through message passing.

1.4 Latent Representation Learning: The Embedding Space as a Fairness-Critical Object

The preceding sections have established the principal algorithmic families that populate the recommendation landscape. A theme that runs across all of them is the construction of a *latent representation*: a low-dimensional embedding of users and items in a shared vector space from which recommendation scores are derived. This section examines the latent space not merely as an engineering convenience but as a first-class theoretical object, and in particular as the space in which the fairness problems addressed by this thesis arise and must ultimately be resolved.

1.4.1 What latent representations encode

A recommendation embedding $\mathbf{z}_u \in \mathbb{R}^d$ is optimized to preserve enough information about user u's interaction history to predict future preferences accurately. In practice this means that the optimization will exploit any statistical regularity in the training data, including regularities that are correlated with sensitive attributes such as gender, age, or geographic origin. Because sensitive

demographics correlate with cultural taste in most large-scale interaction datasets [50, 51], the embedding implicitly encodes a mixture of two distinct information types.

1. **Task-relevant factors**: preferences, genre affinities, listening habits, and item characteristics that are directly useful for ranking.

2. **Sensitive factors**: demographic attributes that are morally irrelevant to the ranking decision but statistically predictive of interaction patterns in the biased training data.

The key observation is that these two information types are *entangled* in a standard embedding: they co-inhabit the same d-dimensional vector without explicit separation. The model has no incentive to disentangle them because both improve prediction accuracy on the observed, biased data. As a result, any downstream recommendation generated from the entangled embedding inherits both the useful preference signal and the harmful demographic signal.

1.4.2 Latent entanglement and its consequences

Formally, let $\mathbf{z}_u = f_\theta(\mathbf{x}_u)$ be the embedding produced by encoder f_θ from interaction vector $\mathbf{x}_u$, and let $a_u \in \{0, 1\}$ denote a binary sensitive attribute (e.g., user gender). Latent entanglement is the condition

$$I(\mathbf{z}_u;\, a_u) \;>\; 0, \tag{1.9}$$

where $I(\cdot;\cdot)$ denotes mutual information. Equation (1.9) states that the embedding carries statistically recoverable information about the sensitive attribute. When entanglement holds, a linear classifier trained on $\mathbf{z}_u$ alone can predict a_u at above-chance accuracy, a fact repeatedly confirmed experimentally [23, 22, 52].

Entanglement has three direct consequences for recommendation fairness. First, because $\mathbf{z}_u$ correlates with a_u, the scoring function $f(\mathbf{z}_u, \mathbf{z}_i)$ will produce systematically different score distributions for users in different demographic groups, even when their true preferences are equivalent. Second, because item embeddings $\mathbf{z}_i$ are learned jointly with user embeddings, provider-side demographic information (e.g., artist gender) contaminates the item embedding by the same mechanism. Third, because the entanglement is *geometric* (encoded in the directions of the embedding space) rather than merely statistical, post-hoc re-ranking cannot remove it: even after re-ranking, the scores that drive future retraining cycles retain the demographic shortcut, perpetuating the bias.

1.4.3 Latent representation learning architectures and their fairness implications

Different recommendation architectures have different degrees of susceptibility to latent entanglement.

Matrix factorization. In the MF formulation (Section 1.2.1), each user embedding $\mathbf{u}_u$ is optimized to predict all observed ratings for user u. Because interaction patterns are correlated with demographics in the training data, the d-dimensional user embedding aligns itself with demographic structure. Zhang et al. [52] showed that MF embeddings are in fact linearly separable by gender on standard benchmarks, and Yao et al. [53] confirmed that pairwise accuracy differences across demographic groups are large and consistent. Critically, regularizing the MF objective toward demographic parity reduces but does not eliminate this linear separability [54], suggesting that the entanglement is a structural feature of the MF objective.

Neural collaborative filtering. NCF embeddings (Section 1.3.1) pass through nonlinear activation functions that can, in principle, store sensitive information in directions that are not linearly accessible. This makes the entanglement harder to detect by linear probes but does not in general reduce the mutual information $I(\mathbf{z}_u; a_u)$. Elazar and Goldberg [22] demonstrated precisely this failure mode in the natural language processing setting: adversarial debiasing that succeeds in fooling a linear probe leaves the representation substantially predictive of the sensitive attribute to a nonlinear probe. The same concern applies directly to NCF.

Variational autoencoders. VAE-based recommenders (Section 1.3.3) learn a stochastic latent space governed by a posterior $q_\phi(\mathbf{z} \mid \mathbf{x}_u)$. The KL penalty in the ELBO encourages $\mathbf{z}$ to resemble a standard Gaussian prior, which provides a form of information compression but does *not* enforce statistical independence between $\mathbf{z}$ and a_u. Louizos et al. [55] showed that a standard VAE faithfully reproduces demographic structure in its posterior, and Moyer et al. [56] proposed a conditional-prior approach that explicitly pressures the posterior toward invariance to sensitive attributes. The DS-DVAE contribution of this thesis (Chapter 3) implements a factorized posterior $(\mathbf{z}^t, \mathbf{z}^s)$ that separates task and sensitive factors at the architectural level, addressing this limitation directly.

1.4.4 Representation learning as the unifying framework

The survey of methods in the subsequent sections will demonstrate that all major fairness-aware approaches can be interpreted as interventions on the latent representation at different levels of directness. Pre-processing methods correct the data that the representation is trained on. Regularization and adversarial in-processing methods constrain the representation at training time. Post-processing methods operate on scores derived from the representation but leave the representation itself unchanged. Only disentanglement-based and orthogonal-regularization methods intervene directly on the *geometry* of the latent space, enforcing structural separation between sensitive and task factors. This observation is the principal motivation for the in-processing, representation-level approach adopted throughout this thesis.

1.5 Defining Bias and Fairness in Recommendation

In the recommendation literature the term *bias* denotes a systematic deviation of the model's predictions or rankings from a notional ground truth that captures users' actual preferences and items' actual relevance. Mehrabi et al. [57] provide a foundational definition in the general machine-learning context, characterizing bias as any systematic and unfair discrimination against certain individuals or groups of individuals in favour of others. Wang et al. [1] operationalize this notion for recommender systems as the discrepancy between the empirically observed outcome distribution and the normatively desired one. Because user-interaction data are observational rather than experimental, they are subject to multiple distortions self-selection, platform exposure policies, temporal drift, and social influence each of which introduces systematic deviation between what is observed and what the model should learn.

Fairness, by contrast, is a normative concept: it specifies how a model *should* behave with respect to individuals or groups identified by sensitive attributes such as gender, age, race, or popularity tier. As Deldjoo et al. [4] emphasized in their survey of a decade of research on the subject, fairness in recommendation is multidimensional and admits no single universally accepted formal definition; the same recommender may be considered fair under one criterion and unfair under another. Li et al. [3] further distinguish between individual and group fairness and demonstrate that the two can be mutually incompatible, reinforcing the need for a careful specification of the fairness objective before any mitigation strategy is designed.

The relationship between the two notions is asymmetric. Biases in the data and in the algorithm are *causes*; unfair recommendation outcomes are their *consequences*. Mitigating bias is therefore typically necessary, but not always sufficient, to ensure fairness: even a model trained on perfectly balanced data may still violate fairness criteria when protected groups have different baseline characteristics or when the evaluation metric itself is biased [15].

1.6 Sources, Causes, and a Taxonomy of Biases

Three classes of mechanisms generate bias in recommender systems: distortions in the training data, distortions arising from the model itself, and distortions induced by the closed-loop interaction between the model and its users. We sketch each class and then unfold them into a taxonomy of the most prominent bias types reported in the literature. The taxonomy we adopt is summarized in Figure 1.8; it extends the classical organization of Mehrabi et al. [57] and Wang et al. [1] with two recently studied families, cold-start and temporal/recency bias, that are increasingly recognized as first-order concerns in modern recommender systems.

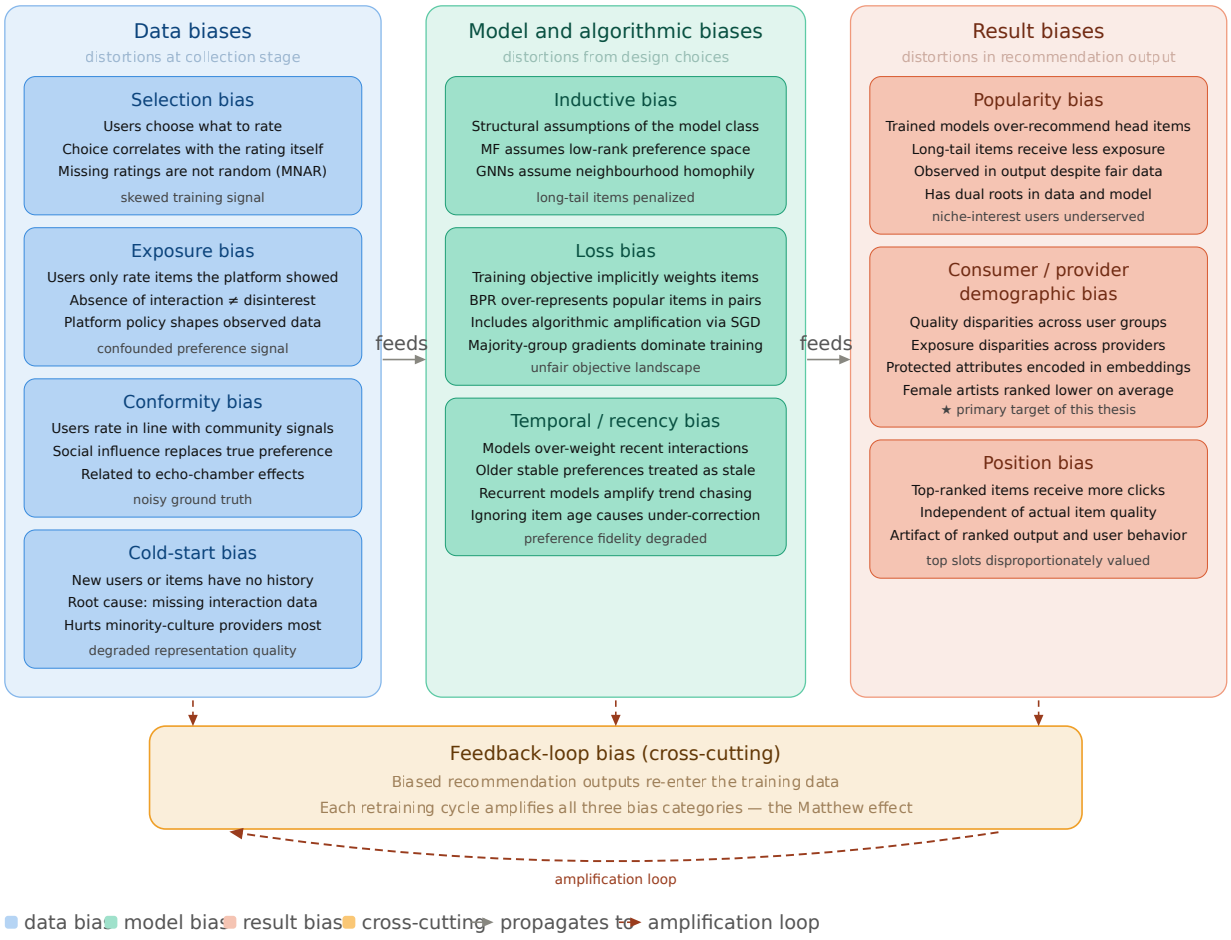


Figure 1.8: Taxonomy of biases in recommender systems

1.6.1 Data biases

Data biases are distortions that enter the recommendation pipeline at the data collection stage, before any model is trained. Because users do not interact with the platform uniformly, they choose which items to rate, they respond only to what the platform shows them, and they are influenced by what others say, the observed interaction matrix is a systematically skewed sample of true preferences. Any model trained on this skewed sample inherits these distortions regardless of how sophisticated its architecture is. The three principal data biases are selection bias, exposure bias, and conformity bias; we also classify cold-start bias here because its root cause is missing interaction data rather than a design choice made by the system architect.

Selection bias arises because users do not interact with items uniformly at random: they choose what to rate, and the choice is itself correlated with the rating. Marlin et al. [58] formally identified this as a missing-not-at-random (MNAR) problem and showed that ordinary matrix-factorization estimators are biased under this regime. Schnabel et al. [59] and Liang et al. [60]

subsequently developed inverse-propensity weighting and exposure modeling to correct for it; the dynamic counterpart of the problem, selection bias that drifts over time together with item age and user preferences, has been recently studied by Huang et al. [61], who introduced the DANCER estimator.

Exposure bias refers to the conflation between absence of preference and absence of opportunity: a user can only express interest in an item after the platform has surfaced it. The absence of an interaction therefore does not in general imply disinterest. Saito et al. [62] estimated propensity scores to correct for exposure in the implicit-feedback setting, and Liu et al. [63] used a knowledge-distillation framework to deconfound observed interactions from the underlying preference signal. We adopt the data-collection framing of Saito et al. [62]; readers who treat platform exposure as a policy decision may alternatively classify this bias under model and algorithmic biases, as in Deldjoo et al. [4].

Conformity bias captures the tendency of users to rate items in line with visible community signals rather than according to their own isolated judgment. Krishnan et al. [64] demonstrated this empirically on Amazon product reviews and proposed a denoising methodology that separates true preference from socially induced shifts. The phenomenon is closely related to homophily and echo-chamber effects studied in social-network analysis [65].

Cold-start bias is a data sparsity limitation that severely disadvantages new users or items with few interactions. Because embedding-based recommenders require sufficient observational signal to learn accurate representations, new entities inherit degraded recommendation quality not from a model design choice but from the absence of training data. The problem was formalized in the meta-learning framework of Lee et al. [43] and the content-conditioned dropout architecture of Volkovs et al. [41]. Although cold-start manifests in model output, its root cause is missing interaction data, which places it squarely in the data category [4]. It disproportionately affects niche or minority-culture content providers and newly joining users from underrepresented demographics, intersecting directly with the fairness concerns developed later in this chapter.

Summary. Data biases share a common origin: they emerge during the data collection stage, before any recommendation model is fit. They are imprinted in the training data itself and therefore propagate to every downstream model trained on it. Mitigating data biases requires intervention at the data level (re-sampling, re-weighting, propensity correction) or at the training-loss level (inverse propensity scoring, debiasing through causal assumptions). Even the most sophisticated in-processing or post-processing fairness method will struggle to recover from severe data biases, which makes their identification and correction the natural first step of any fairness pipeline.

1.6.2 Model and algorithmic biases

Model and algorithmic biases originate in the choices made by the system designer: the model architecture, the training objective, and the optimization procedure. Unlike data biases, which are inherited passively from the collection process, model biases are actively introduced by the learning algorithm itself. A matrix factorization model imposes a low-rank preference geometry; a BPR loss over-weights popular items as positive training examples; a recurrent model that discounts old interactions encodes a preference for recency. These biases cannot be removed by pre-processing the data alone because they re-emerge at every training step; the appropriate remedy is in-processing, that is, modifying the model or its objective to counteract the chosen structural assumptions.

Inductive bias is the set of structural assumptions implicit in the model class. Matrix factorization assumes a low-rank linear preference space; graph neural networks assume neighborhood homophily. Steck [66] showed that the inductive bias of common collaborative-filtering objectives systematically penalizes long-tail items because their reconstruction error contributes less to the overall loss.

Loss bias concerns the implicit weighting that different training objectives place on items and interactions. Pair-wise losses such as BPR over-represent popular items because they appear more frequently as positive examples in the constructed training pairs [66]; point-wise and list-wise objectives induce their own characteristic distortions [67]. A related phenomenon, algorithmic amplification bias, arises when stochastic gradient descent amplifies majority-group signals through gradient magnitude asymmetry, exacerbating demographic disparities beyond what the loss function itself would predict [5].

Temporal / recency bias arises when models weight recent interactions disproportionately and treat older interactions as if they were stale, even when they reflect stable long-term preferences. We adopt the model-level framing of Huang et al. [61] and Beutel et al. [68], who attribute recency bias to the model's weighting decisions at training time; Jin et al. [6] alternatively classify it as a data bias, grounding it in the genuine temporal drift of user preferences in the underlying data distribution. Both views are defensible, and the appropriate mitigation differs accordingly. Huang et al. [61] showed that ignoring item age in debiasing pipelines leads to systematic under-correction of selection bias for older items, and Beutel et al. [68] demonstrated that naive temporal modeling in recurrent recommenders amplifies trend chasing at the expense of preference fidelity.

Summary. Model and algorithmic biases differ from data biases in that they arise from choices made by the system designer, namely the model class, the training objective, and the optimization procedure. They cannot be removed by data-only pre-processing because they are properties of the learning algorithm itself. The mitigation strategy is therefore in-processing: modifying the loss function, the model architecture, or the training schedule so that the chosen biases are explicitly counteracted during learning.

1.6.3 Result biases

Result biases are the biases that users and content providers experience directly in the recommendation output. They are downstream consequences of data and model biases, made visible at the point where the ranked list is presented to the user. Because they are observable, they are also the biases most frequently targeted by post-processing mitigation strategies such as re-ranking. The three principal result biases are popularity bias, consumer and provider demographic bias, and position bias. Popularity bias has dual roots in data collection and model training, but we classify it as a result bias because it is primarily observed and measured in the recommendation output [15].

Popularity bias is the tendency of trained recommenders to over-recommend already popular items at the expense of long-tail items that would be relevant to niche-interest users. Klimashevskaia et al. [15] showed that popularity bias is present across all major recommendation algorithm families; Abdollahpouri et al. [69] introduced personalized re-ranking to mitigate it; and Zhu et al. [70] demonstrated that the problem is amplified through dynamic feedback in deployed systems. Although popularity bias has roots in both data collection and model training, we classify it as a result bias because it is primarily observed and measured in the recommendation output [15].

Consumer / provider demographic bias refers to the symmetric pair of disparities that arise on the two stakeholder sides of the recommendation system. On the consumer side, it manifests as systematic disparities in recommendation quality across user groups defined by protected attributes such as gender, age, race, or geographic location; Ekstrand et al. [71] provided early empirical evidence in book recommendation, and Melchiorre et al. [50] documented substantial accuracy disparities between male and female users in music recommendation across multiple state-of-the-art collaborative-filtering models. On the provider side, the same mechanism manifests as unfair exposure of items defined by the demographic attributes of their creators: Ferraro et al. [51] found, on Last.fm, that the first female artist appears on average only at the sixth or seventh recommendation position while male artists routinely occupy the top rank, and Ferraro et al. [72] subsequently showed that this gender imbalance is amplified through feedback loops. The two manifestations share the same underlying cause, namely the encoding of demographic attributes into the learned representations, and the contributions developed in this thesis address them jointly.

Position bias reflects the substantially higher click-through rates that items at the top of a ranked list receive, regardless of their intrinsic quality. It is properly understood as a result-level bias because it is an artifact of the ranked output and the user's examination behavior rather than of the data-collection process. Joachims et al. [73] modelled position-dependent examination probabilities, and Agarwal et al. [74] derived production-grade unbiased learning-to-rank estimators from observational click logs.

Summary. Result biases are observable in the recommendation output and in the user's interaction with it, downstream of both data collection and model training. They are the biases that users

and providers experience directly: a user observing recommendations of consistently lower quality, an artist observing systematically lower exposure than peers of comparable popularity. Because they are the most visible manifestation, they are also the family for which mitigation methods are most diverse, spanning post-hoc re-ranking, in-processing constraints, and counterfactual generation. The contributions developed in this thesis target the consumer-side and provider-side demographic instances of this category jointly at the representation level.

1.6.4 Feedback-loop bias

Recommendation systems do not operate on a static dataset: their outputs influence the users and the data that will be used to retrain the model. Chaney et al. [75] demonstrated that even a small initial bias may be progressively amplified through repeated retraining cycles, the so-called Matthew effect [70]. Mitigating bias in the present is therefore not only a matter of present fairness but also of preventing the long-term degradation of system behavior.

1.7 Perspectives on Fairness

Before enumerating formal fairness criteria and mitigation methods, it is helpful to make explicit what is meant by fairness in the context of recommendation. Broadly speaking, fairness is the property that a system does not systematically advantage or disadvantage individuals or groups on the basis of characteristics that are morally irrelevant to the recommendation decision. In practical terms this translates to requirements such as the following: a female user should receive recommendations of comparable accuracy to a male user with similar tastes; a niche music artist should receive proportional exposure relative to an equivalent mainstream artist; a new learner from an under-represented demographic should not receive markedly lower-quality course suggestions than a learner from the majority demographic.

Following Wang et al. [1] and Deldjoo et al. [4], fairness in recommender systems can be analyzed along two complementary axes: a *level* perspective (granularity of the protected unit) and a *subject* perspective (which stakeholder is protected). Both axes are summarized in Figure 1.9.

1.7.1 Level perspective

The level perspective concerns the granularity at which fairness is demanded. Three levels are distinguished in the literature.

Individual fairness requires that similar individuals receive similar outcomes, regardless of any group membership. The unit of comparison is the pair of users (or pair of items), and the criterion is symmetric: if two users are similar with respect to a task-relevant notion of similarity, the recommender should produce recommendation lists for them that are similar with respect to a

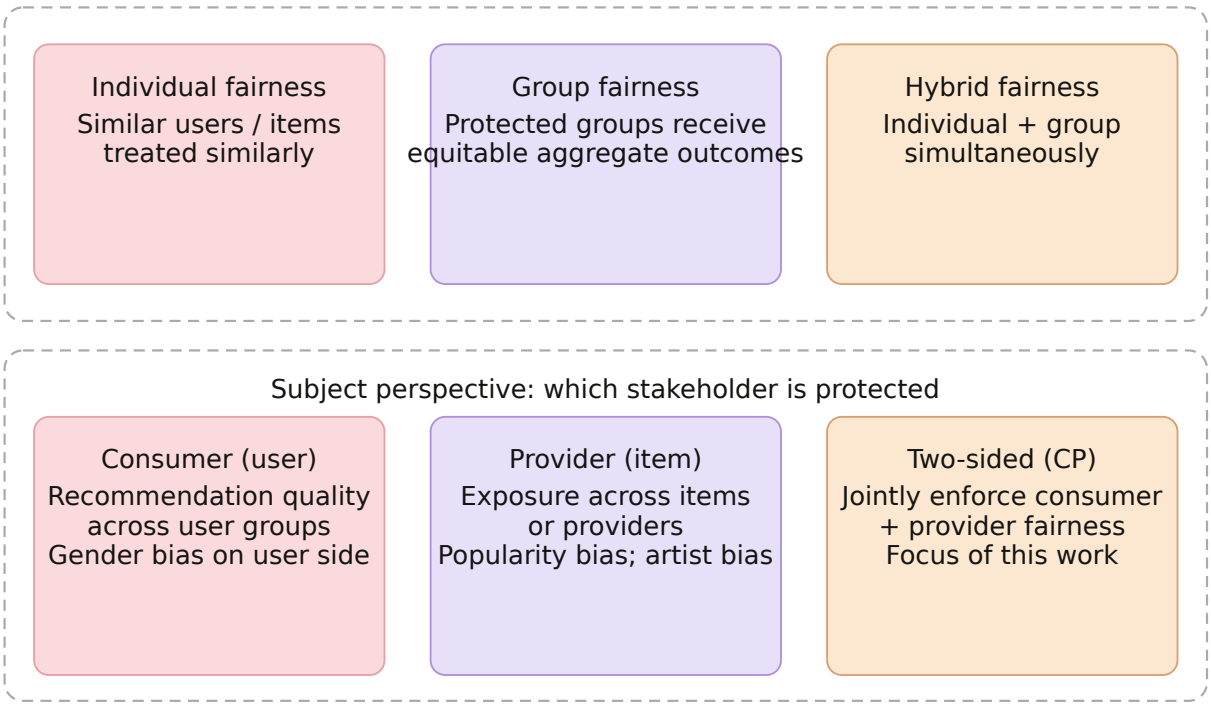


Figure 1.9: Fairness perspectives in recommender systems.

task-relevant notion of distance between lists. Consider, for illustration, two students enrolled in the same data-science master's program, with the same prior courses, the same listed interests, and comparable browsing histories on a course-recommendation platform. A C-fair recommender from the individual-fairness perspective should produce essentially the same top-K recommendations for both students, even if one is from a demographically over-represented background and the other is not. If the two students systematically receive different recommendations despite their preference profiles being indistinguishable, the recommender violates individual fairness.

The formalization by Dwork et al. [76] captures this through a Lipschitz condition: if $d(u, v)$ denotes a task-appropriate similarity metric over users, then a fair predictor f must satisfy $|f(u) - f(v)| \leq L \cdot d(u, v)$ for some constant L. In ranking, Biega et al. [77] introduced the equity-of-attention criterion, which requires that the cumulative attention an item receives across multiple recommendation rounds be proportional to its merit. Individual fairness is conceptually appealing but its operationalization hinges on the availability of a meaningful similarity metric, which is difficult to elicit for diverse user populations.

Group fairness requires that protected groups receive equitable aggregate outcomes. The unit of comparison is the demographic group rather than the individual, and the criterion is typically a parity statement: a statistic of interest (recommendation quality, exposure share, click-through rate) should be approximately equal across the protected groups. Returning to the course-

recommendation platform aggregated across all its users, a group-fair recommender should ensure that the average top-K relevance score, or the average enrollment rate from recommendations, is comparable between female and male users (the two groups defined by the chosen sensitive attribute). It is not required that any individual female user receive the same recommendations as any individual male user; only that the average outcome over each group be comparable. A recommender that systematically over-serves one gender group, even while satisfying individual fairness within each group, violates group fairness.

Adapting the criteria of Hardt et al. [78] to recommendation, three dominant notions are routinely instantiated. Let $A \in \{G_0, G_1\}$ denote the sensitive attribute, $Y \in \{0, 1\}$ the latent ground-truth preference, and $\hat{Y} \in \{0, 1\}$ the recommendation decision. Then:

- **Demographic Parity (DP)** requires that the probability of being recommended an item is independent of the sensitive attribute,

$$P(\hat{Y} = 1 \mid A = G_0) \;=\; P(\hat{Y} = 1 \mid A = G_1).$$

  DP is conceptually clean but ignores the underlying base rates of preference: if the two groups genuinely differ in their interests, DP can be enforced only by sacrificing accuracy.

- **Equal Opportunity (EO)** requires that, among users who *would* prefer the item, the true-positive rate be equal across groups,

$$P(\hat{Y} = 1 \mid Y = 1, A = G_0) \;=\; P(\hat{Y} = 1 \mid Y = 1, A = G_1).$$

  EO addresses the base-rate concern but is silent about false positives: a system that recommends irrelevant items at very different rates across groups can still be EO-fair.

- **Equalized Odds (EqOdds)** requires that both true-positive and false-positive rates be equal across groups simultaneously,

$$P(\hat{Y} = \hat{y} \mid Y = y, A = G_0) \;=\; P(\hat{Y} = \hat{y} \mid Y = y, A = G_1) \quad \forall \hat{y}, y \in \{0, 1\}.$$

  EqOdds is the strongest of the three, but Chouldechova [79] and Kleinberg et al. [80] independently proved that EqOdds is incompatible with calibration whenever base rates differ between groups, a result that any practical fairness pipeline must accommodate.

Group fairness is the dominant operationalization in recommendation research, accounting for roughly two-thirds of all studies in the systematic review of Wang et al. [1]. It is computationally tractable and admits well-defined gap metrics, but it operates at the aggregate level and may mask substantial within-group disparities.

Hybrid fairness combines criteria from both levels. A common instantiation enforces individual fairness within each demographic group while simultaneously enforcing group fairness across

groups, which closes the within-group blindspot of pure group fairness without abandoning its aggregate guarantees. On the course platform discussed above, the hybrid criterion would require both that two students with indistinguishable profiles receive comparable recommendations (individual fairness within groups) and that, on average, the quality of recommendations delivered to female and male students be equivalent (group fairness across groups). Either criterion alone leaves a loophole that the other closes.

Wang et al. [1] identify a growing body of such work, particularly suited to two-sided platforms where the consumer and provider sides have asymmetric notions of similarity and equity, and Zhao et al. [67] further proposed joint diversity-fairness regularization as a hybrid constraint applicable to embedding-based recommenders.

1.7.2 Subject perspective

The subject perspective concerns which stakeholder is the target of the fairness criterion. In a recommendation platform there are at least two distinct stakeholders, the consumer (the user receiving recommendations) and the provider (the entity whose item is being recommended), and a fairness criterion must specify which one it protects.

Consumer (user) fairness (C-fairness) concerns the quality of recommendations received by users. A system is C-fair if users belonging to different demographic groups receive recommendations of comparable relevance and accuracy. On a music streaming service, for instance, a C-fair recommender should ensure that a female listener whose taste profile emphasises hip-hop receives a top-K list of comparable quality (measured by NDCG, hit rate, or any other relevance metric) to a male listener whose taste profile is also hip-hop-leaning. If the recommendations systematically perform worse for the female listener because the recommender has implicitly learned gendered associations between listening histories and predicted preferences, the system is C-unfair, even if the female listener still finds the recommendations broadly acceptable in absolute terms.

Ekstrand et al. [71] provided early empirical evidence that recommendation accuracy varies significantly across demographic groups in book recommendation; Wu et al. [81] proposed graph-based debiased user representations to enforce C-fairness; and Melchiorre et al. [50] documented substantial accuracy disparities between male and female users on the LFM-2b dataset across several state-of-the-art collaborative-filtering models. The empirical consequences of C-unfairness include reduced engagement and trust among the disadvantaged groups, with downstream effects on platform retention and on the diversity of feedback signals collected in subsequent training rounds.

Provider (item) fairness (P-fairness) concerns the exposure that items or content creators receive. A system is P-fair if content from different creator groups (whether defined by popularity tier, demographic attribute, or any other relevant criterion) receives exposure commensurate with relevance rather than with the creator's group membership. On the same music streaming service, a P-fair recommender should ensure that a female artist whose musical quality is comparable to

that of a male artist receives a similar share of total recommendation exposure across the platform's user base. If male artists routinely occupy the top recommendation slots and the first female artist appears, on average, only at position six or seven, the platform is P-unfair to the female artist group, irrespective of the listening preferences of any individual user.

The most studied form of P-fairness is popularity-bias mitigation [69, 15]: long-tail items receive systematically less exposure than their relevance warrants. A second form is provider-demographic fairness: in the music domain, Ferraro et al. [51] documented severe under-exposure of female artists, and Ferraro et al. [72] showed that this imbalance is dynamically amplified by user feedback. The consequence of P-unfairness is structural: producers of high-quality niche content receive insufficient revenue and visibility, which over time depresses the platform's catalog diversity.

Two-sided (consumer-provider, CP) fairness frameworks consider C- and P-fairness jointly and seek to satisfy both simultaneously. Continuing the music platform scenario, a two-sided fair recommender should ensure both that female listeners receive recommendation lists of equivalent quality to male listeners (C-fairness) and that female artists receive a fair share of platform-wide exposure (P-fairness), all within a single recommendation model. The challenge is that these two criteria can interact: a naive intervention that increases female-artist exposure may inadvertently reduce the relevance of recommendations for users of any gender, and a naive intervention that equalizes recommendation quality across user genders may concentrate exposure on a small set of broadly appealing male artists.

Despite its conceptual appeal, two-sided fairness remains a comparatively young subfield: Deldjoo et al. [4] report that dedicated two-sided methods are substantially fewer than their one-sided counterparts, and that the existing two-sided literature is dominated by post-processing approaches that leave the underlying biased representations untouched. The detailed comparative analysis of two-sided frameworks, together with the in-processing alternatives developed in this thesis, is the subject of Chapter 2.

In this thesis we adopt the **two-sided fairness perspective**, combining group fairness on the consumer side (with user gender as the sensitive attribute) and group fairness on the item side (with popularity tier as the sensitive attribute).

1.8 Fairness-Aware Methods: A Three-Stage View

Following Mehrabi et al. [57], Li et al. [3], and Wang et al. [1], fairness-aware methods are organized into three stages depending on the point in the pipeline at which they intervene. Pre-processing methods modify the training data before model fitting; in-processing methods incorporate fairness directly into the learning objective or the model architecture; post-processing methods re-rank the trained model's outputs. Figure 1.10 illustrates the three stages and their characteristic trade-offs.

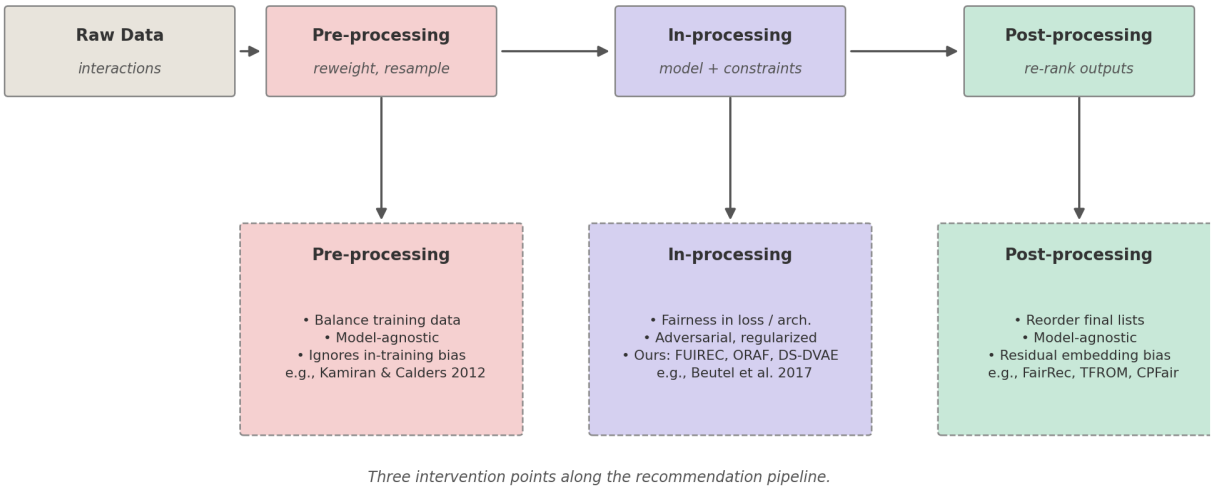


Figure 1.10: The three-stage view of fairness-aware methods.

1.8.1 Pre-Processing Methods

Pre-processing methods aim to neutralize bias in the training data before any recommendation model is fit. Because they intervene on the data alone they are model-agnostic: the same corrected dataset can feed any downstream architecture without modification. They cannot, however, address biases introduced during model training, such as loss bias or inductive bias. Wang et al. [1] distinguish three sub-families: data re-sampling, data re-labeling, and data modification or augmentation. Figure 1.11 summarizes the sub-families.

Data re-sampling

Data re-sampling addresses fairness by correcting the representation imbalance in the training set before any model sees the data. The underlying idea is that a model trained on a balanced dataset, in which protected and non-protected groups contribute equally to the learning signal, will learn representations and decision boundaries that are less correlated with the sensitive attribute. Re-sampling achieves this balance through two complementary mechanisms. *Over-sampling* duplicates or synthesizes new instances from under-represented groups until their contribution to the training loss matches that of majority groups; *under-sampling* removes instances from over-represented groups to produce the same effect. In the recommendation context, re-sampling operates on user-item interactions: interactions involving minority-group users or niche-provider items are replicated or up-weighted so that the model receives equally strong gradients from all demographic segments. The fairness guarantee is statistical: by construction, the empirical distribution of the corrected training set satisfies a target balance criterion, and any model that fits this distribution inherits the

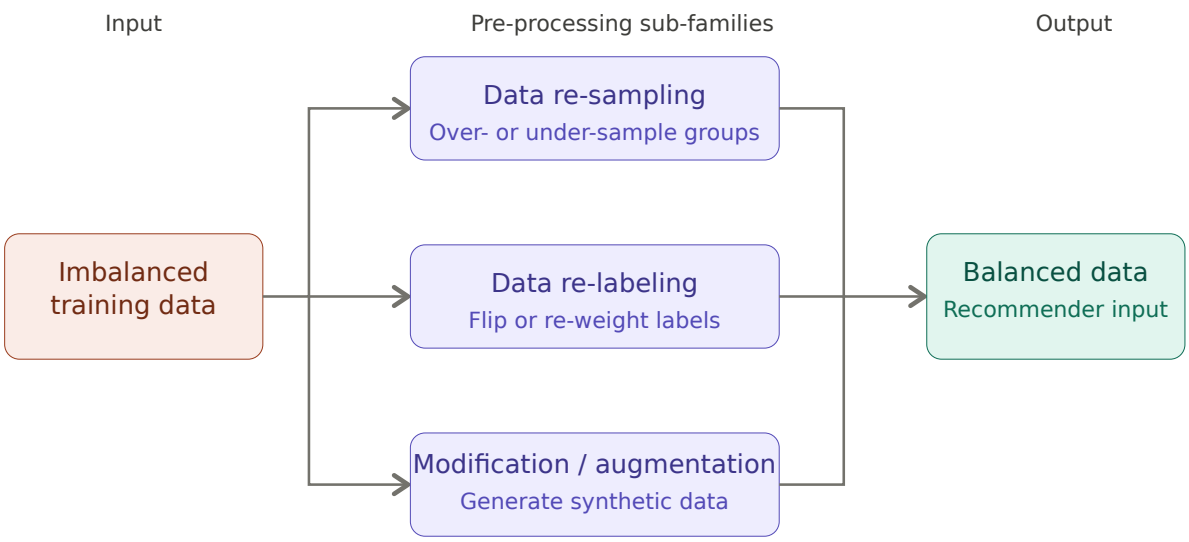


Figure 1.11: Pre-processing fairness-aware methods.

balance. The mechanism does not require modifying the recommendation objective or the model architecture, making it straightforwardly composable with any downstream learning method. Its principal limitation is that duplicated interactions do not add new information — they only redistribute weight — so re-sampling alone cannot compensate for interactions that are entirely absent from the data, a regime better addressed by augmentation.

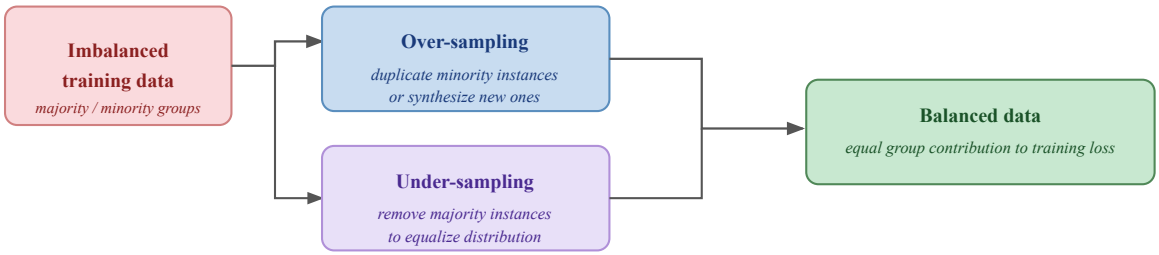


Figure 1.12: Data re-sampling.

Data re-labeling

Data re-labeling ensures fairness by modifying the ground-truth interaction labels themselves rather than their frequency. The intuition is that observed ratings or click signals reflect not only genuine preferences but also historical inequities: a female author whose book receives systematically lower

ratings due to gender bias, or a niche artist whose streams are suppressed because the platform surfaces mainstream content by default. Re-labeling corrects these distortions by flipping, re-weighting, or smoothing individual labels so that the resulting training set satisfies a target fairness criterion, most commonly demographic parity or equalized opportunity across protected groups. Concretely, a re-labeling procedure solves an optimization problem: find the minimum-cost perturbation of the label matrix such that, after perturbation, the distribution of labels is statistically independent of (or equalized across) the sensitive attribute. The fairness guarantee is baked into the label distribution itself: any model that correctly fits the re-labeled data will produce predictions consistent with the target fairness criterion. Re-labeling is particularly attractive when bias is concentrated in a small number of interactions, because the perturbation budget is low. Its risk is label corruption: if the corrected labels deviate too far from genuine user preferences, the model learns a distorted preference signal and recommendation utility degrades.

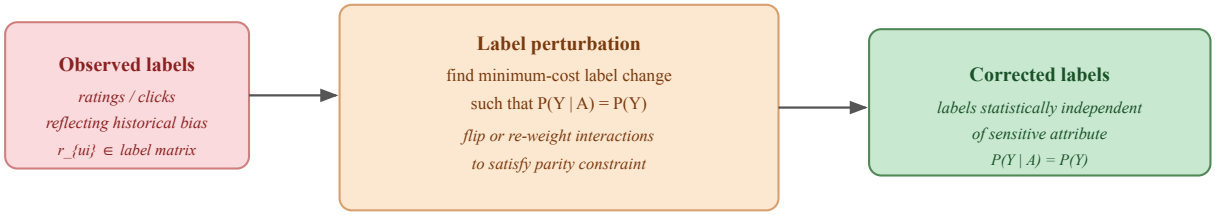


Figure 1.13: Data re-labeling.

Data modification and augmentation

Data modification and augmentation ensure fairness by generating new training instances rather than reweighting or correcting existing ones. Unlike re-sampling, which redistributes weight among observed interactions, augmentation expands the training set with synthetic data that fills the demographic gaps not present in the original corpus. The core mechanism is as follows: a generative or rule-based procedure produces new user-item interaction records for under-represented groups or item categories, crafted so that their statistical properties are consistent with the target fairness criterion. These synthetic records are then merged with the original data and the combined dataset is used to train the recommender. The fairness guarantee is distributional: the augmented training set has, by construction, a more balanced interaction density across demographic segments, which means that gradient updates from minority groups receive adequate weight throughout training. Because augmentation adds genuinely new instances rather than duplicating existing ones, it can compensate for structurally missing data — interactions that were never observed because certain users or items were systematically excluded from exposure. Its limitation is the fidelity of the synthetic data: generated interactions must be realistic enough that the model does not overfit

to artifacts of the generation process rather than to true preferences. Antidote-data methods are a specialized variant in which the synthetic dataset is adversarially crafted to cancel a specific observed bias without altering the deployed model, providing a post-hoc correction mechanism that requires no retraining.

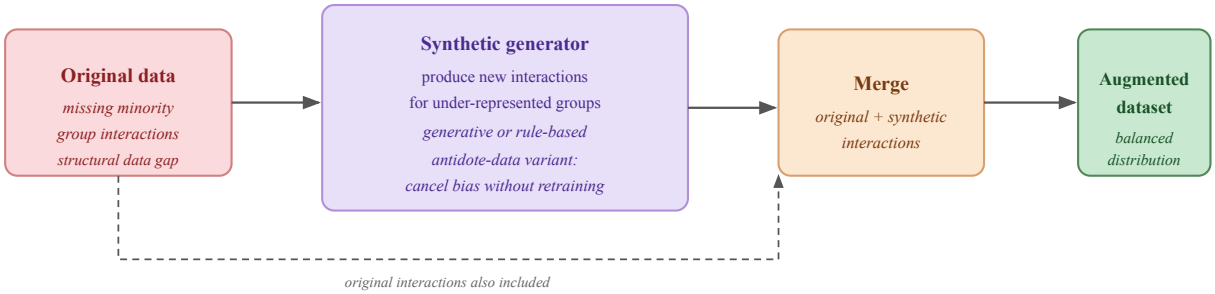


Figure 1.14: Data augmentation.

1.8.2 In-Processing Methods

In-processing methods incorporate fairness directly into the learning objective or the model architecture. They typically achieve a more favorable accuracy-fairness trade-off than post-processing methods, since they intervene at the level of the learned representations themselves. Wang et al. [1] identified five canonical sub-families: regularization-based, causal-inference-based, adversarial-learning-based, reinforcement-learning-based, and ranking-optimization-based methods. We additionally treat disentanglement-based and orthogonal-regularization methods as two further peer sub-families, whose architectural and geometric guarantees are logically independent of adversarial training. Figure 1.15 summarizes all seven sub-families.

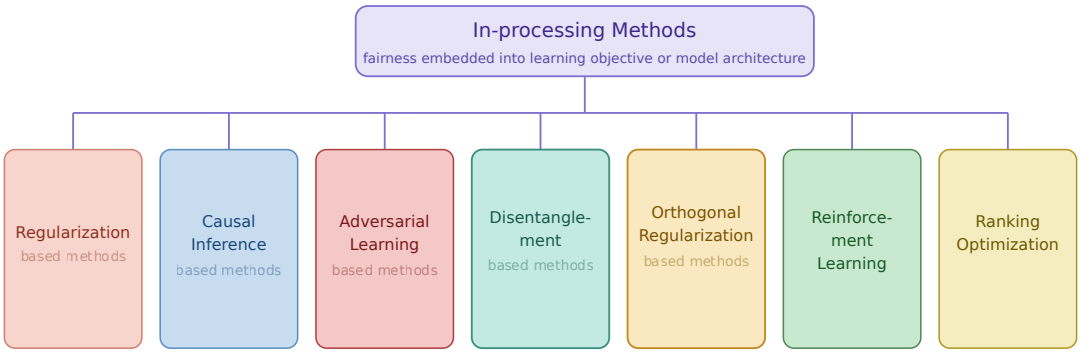


Figure 1.15: In-processing fairness-aware methods.

Regularization-based methods

Regularization is an approach that adds an extra penalty term to the standard training loss of the recommender. The penalty measures how unfair the model's current predictions are with respect to a chosen fairness criterion, and the model is trained to minimize both the recommendation error and this fairness penalty simultaneously. The practitioner controls the balance between the two objectives through a single trade-off hyperparameter.

The training objective takes the form

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{rec}}(\theta) + \lambda\, \mathcal{R}_{\text{fair}}(\theta),$$

where $\mathcal{L}_{\text{rec}}$ is the standard recommendation loss, $\mathcal{R}_{\text{fair}}$ is a fairness regularizer that measures the degree of demographic disparity in the model's predictions, and λ is a trade-off hyperparameter that controls the relative importance of the two objectives. The regularizer can take many forms depending on the fairness criterion: it may be the squared difference in average predicted scores between protected and non-protected user groups (targeting demographic parity), the difference in true positive rates across groups (targeting equal opportunity), or a mutual-information estimate between predicted scores and the sensitive attribute. Regardless of its specific form, the regularizer creates a gradient signal that pushes the model parameters away from configurations that produce demographic disparity, while the recommendation loss simultaneously pulls them toward configurations that fit the observed interaction data. At convergence, the model parameters represent a balance between these two competing objectives governed by λ. The fairness guarantee is statistical: the learned model minimizes disparity as measured by the chosen regularizer, but no geometric or structural separation of sensitive information is enforced. Figure 1.16 illustrates the three representative formulations.

Causal-inference-based methods

Causal inference is an approach that models the data-generating process of the recommendation system using a structural causal model, explicitly separating which statistical associations between sensitive attributes and outcomes are genuine causal effects from which are spurious correlations introduced by historical confounders. The model is then trained on data from which the spurious paths have been blocked, either through backdoor adjustment or counterfactual data augmentation. The key conceptual distinction is between *correlation* and *causation*: a recommender trained on observational data absorbs not only the genuine causal effect of preferences on interactions but also spurious correlations introduced by unobserved confounders. For example, a user's gender (sensitive attribute) may be correlated with their interaction history (input features) not because gender causally drives preferences but because historical platform policies exposed certain genders to certain content. A model that learns from this confounded data will encode the spurious correlation and perpetuate the bias. Causal-inference methods break this mechanism by explicitly

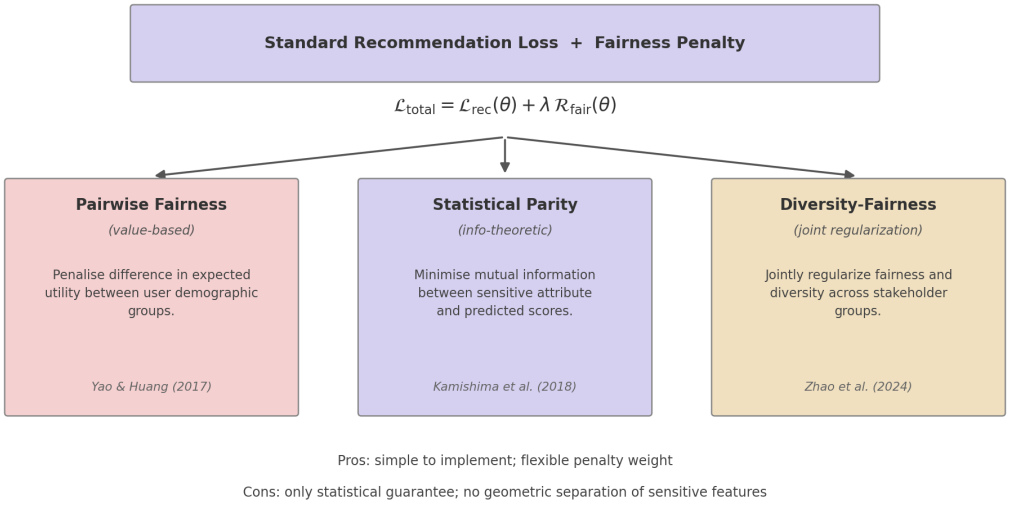


Figure 1.16: Regularization-based fairness.

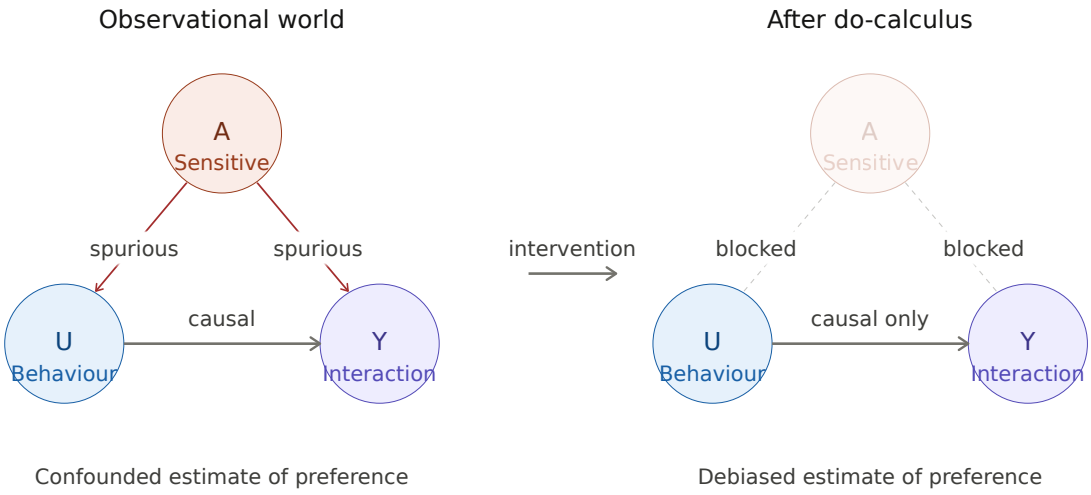


Figure 1.17: Causal inference for debiasing.

modeling the data-generating process as a structural causal model, identifying which associations are causal and which are spurious, and intervening to remove the spurious paths. The standard tool is backdoor adjustment: given a set of confounders Z blocking all back-door paths from the sensitive attribute A to the outcome Y, the causal effect of the recommendation action on Y is estimated by conditioning on Z and marginalizing it out. Counterfactual augmentation is a complementary strategy: synthetic interactions are generated by asking what a user's interaction would

have been had their sensitive attribute been different while holding all causally relevant factors constant. Training on these counterfactual examples teaches the model to produce predictions that are invariant to the sensitive attribute along non-causal paths. The fairness guarantee is causal: the model's predictions reflect genuine preferences rather than demographic correlations induced by historical inequity. The limitation is that a credible causal model must be specified in advance, which requires domain knowledge and is not always possible in practice. Figure 1.17 illustrates the confounding mechanism and the effect of do-calculus intervention.

Adversarial-learning-based methods

Adversarial learning is an approach borrowed from generative adversarial networks in which two neural networks, an encoder and a discriminator, are trained simultaneously against each other. The encoder maps users and items into a latent embedding space and tries to hide any sensitive information in those embeddings, while the discriminator tries to recover the sensitive attribute from the same embeddings. The competition between the two forces the encoder to learn representations from which the sensitive attribute cannot be inferred. Adversarial methods reframe

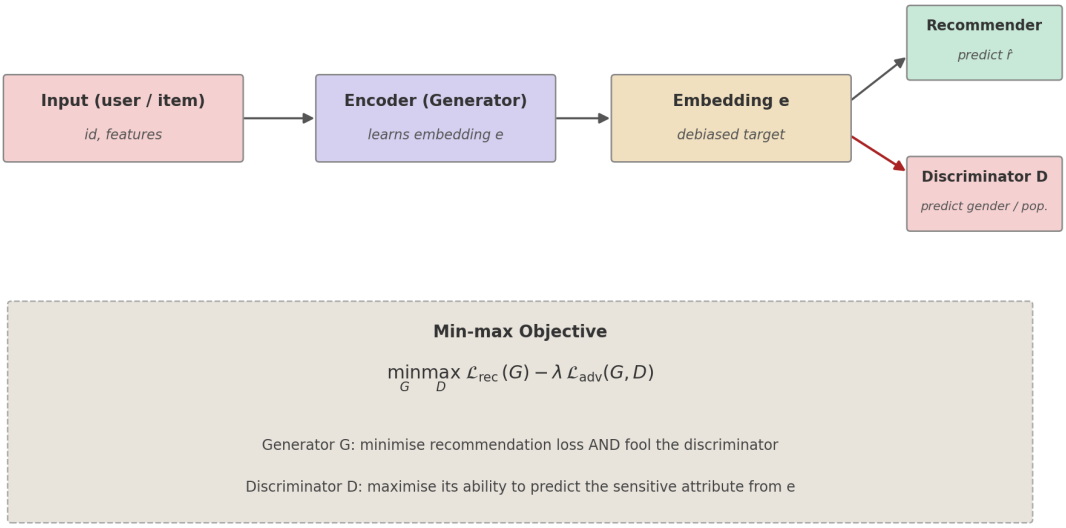


Figure 1.18: Adversarial learning architecture.

fair representation learning as a competitive two-player game. The first player is an *encoder* (or generator) G, whose role is to map raw user and item features into a latent embedding space. The second player is a *discriminator* D, whose role is to predict the sensitive attribute (for example, user gender or item popularity tier) from the embeddings produced by G. The two players are trained simultaneously with opposing objectives: the encoder seeks to produce embeddings that are highly informative for recommendation while making it impossible for the discriminator to infer the

sensitive attribute, whereas the discriminator seeks to recover the sensitive attribute as accurately as possible. Formally, the joint objective is

$$\min_{G} \max_{D} \quad \mathcal{L}_{\text{rec}}(G) - \lambda \, \mathcal{L}_{\text{adv}}(G, D),$$

where $\mathcal{L}_{\text{rec}}$ measures recommendation quality and $\mathcal{L}_{\text{adv}}$ measures the discriminator's ability to predict the sensitive attribute. At convergence, the encoder has learned a representation from which the sensitive attribute is distributionally indistinguishable. Figure 1.18 illustrates the canonical architecture.

A well-known limitation is that adversarial training enforces only statistical independence: residual sensitive information typically remains linearly recoverable from the trained embedding by a separately fitted classifier [22]. This motivates the two peer sub-families below.

Disentanglement-based methods

Disentanglement is an approach in which the encoder is explicitly designed to produce two separate representation subspaces: one that contains only preference-relevant information used for recommendation, and one that absorbs all sensitive attribute information and is discarded at inference time. Because the prediction is made exclusively from the preference subspace, the sensitive attribute has no architectural path to influence the output, making fairness a structural property of the model rather than a learned statistical behavior. Disentanglement-based methods ensure

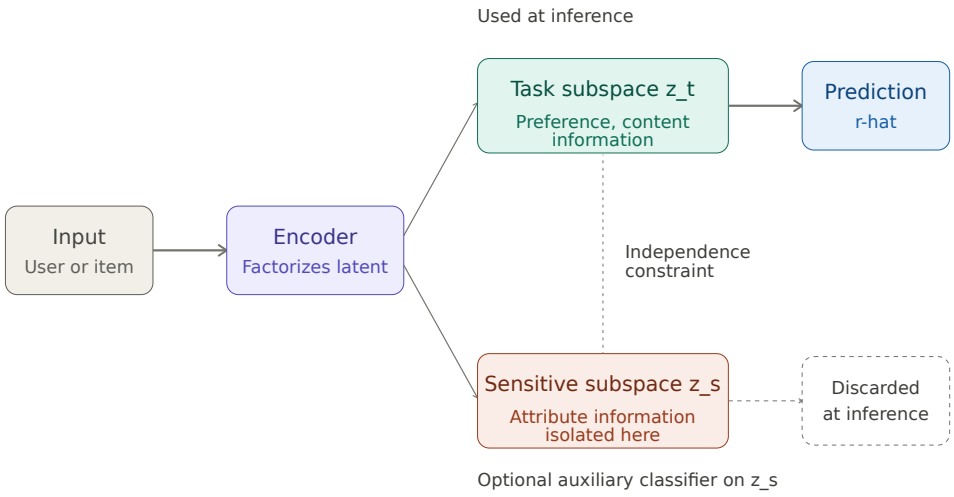


Figure 1.19: Disentanglement-based debiasing approach.

fairness through structural independence: the encoder explicitly partitions the latent space into a *task subspace* $\mathbf{z}^t$, which captures preference-relevant information, and a *sensitive subspace* $\mathbf{z}^s$, which absorbs information about the sensitive attribute. At inference time, the recommendation score is

computed exclusively from $\mathbf{z}^t$, so the sensitive attribute has no architectural path to influence the prediction output. Fairness is thus guaranteed by the architecture itself rather than by a post-hoc penalty. The factorization is enforced during training by an independence constraint, typically a Kullback-Leibler divergence regularizer that pressures $\mathbf{z}^t$ to be statistically independent of the sensitive attribute, and optionally by an auxiliary classifier trained to predict the sensitive attribute from $\mathbf{z}^s$ alone, which forces the encoder to deposit sensitive information into the designated subspace rather than allow it to leak into $\mathbf{z}^t$. This sub-family does not require an adversarial discriminator: the factorization can be achieved entirely through a generative variational objective. Figure 1.19 illustrates the architecture.

Orthogonal-regularization methods

Orthogonal regularization is an approach that imposes a geometric constraint on the relationship between two embedding vectors in the latent space: the task embedding and the sensitive embedding are forced to be perpendicular to each other. When two vectors are orthogonal, no linear operation can project one onto the other, which means no linear classifier can use the task embedding to predict the sensitive attribute, providing a deterministic geometric guarantee that adversarial training alone cannot deliver. Orthogonal-regularization methods ensure fairness through geometric independence:

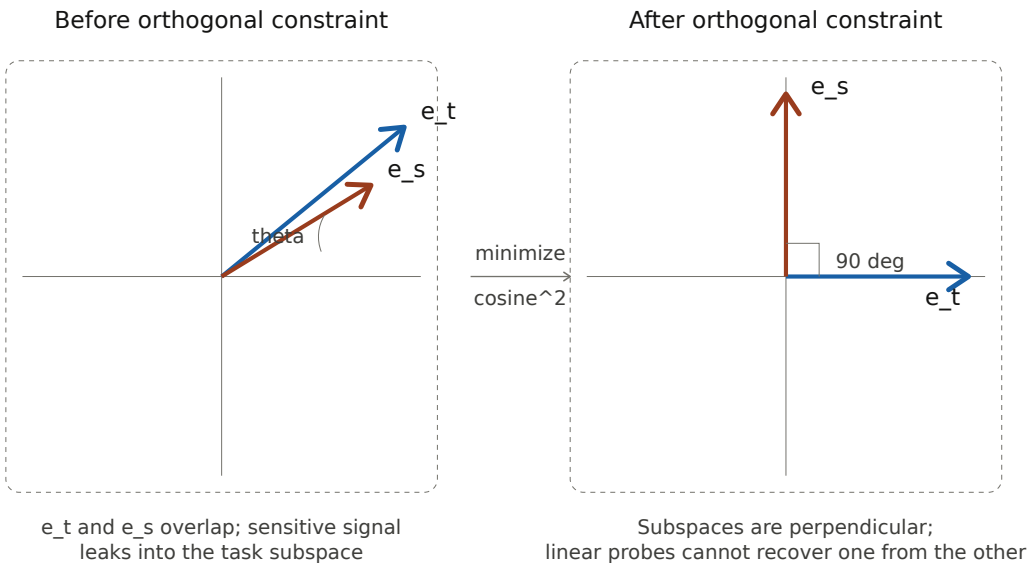


Figure 1.20: Orthogonal regularization approach

the task embedding $\mathbf{e}^t$ and the sensitive embedding $\mathbf{e}^s$ are constrained to be perpendicular in the shared vector space. When two vectors are orthogonal, no linear operation can project one onto the other, meaning the task representation cannot be exploited by any linear classifier to predict

the sensitive attribute. The orthogonality penalty is the squared cosine similarity between the two embedding vectors,

$$\mathcal{L}_{\text{ortho}} = \left(\frac{(\mathbf{e}^t)^\top \mathbf{e}^s}{\|\mathbf{e}^t\| \, \|\mathbf{e}^s\|} \right)^2,$$

which equals zero if and only if $\mathbf{e}^t \perp \mathbf{e}^s$. This geometric constraint is logically independent of the statistical guarantee of adversarial training and can be applied on top of any base objective — adversarial, disentanglement, or plain reconstruction. In practice, combining it with adversarial training yields stronger and more complementary fairness guarantees than either alone. Figure 1.20 illustrates the before-and-after effect of the constraint.

Reinforcement-learning-based methods

Reinforcement learning is an approach that models the recommender as an autonomous agent interacting with users over time. Rather than optimizing a fixed loss on a static dataset, the agent learns a policy by receiving reward signals after each recommendation action. Fairness is built into the reward function so that the agent is rewarded not only for relevant recommendations but also for demographically equitable ones, learning through repeated interaction to balance both objectives across the entire recommendation horizon.

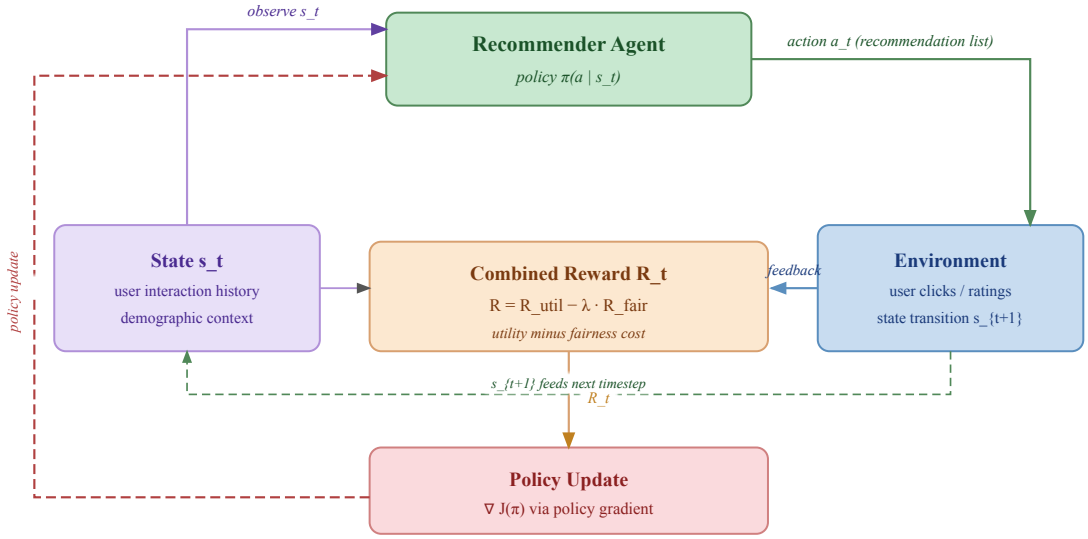


Figure 1.21: Reinforcement-learning-based fairness-aware recommendation.

Reinforcement-learning methods ensure fairness by embedding a fairness criterion directly into the reward signal that drives the recommender's sequential decision-making. The recommender is modeled as an agent that operates in a dynamic environment: at each timestep t it observes a user state s_t (encoding the user's interaction history and context), selects a recommendation list as action a_t according to its policy $\pi(a \mid s_t)$, and receives a scalar reward R_t from the environment.

What distinguishes fairness-aware RL from standard RL is the composition of the reward:

$$R_t = R_{\text{util}}(s_t, a_t) - \lambda \cdot R_{\text{fair}}(s_t, a_t),$$

where R_{util} captures recommendation quality (click-through rate, rating accuracy, or a surrogate thereof) and R_{fair} measures the demographic disparity induced by action a_t — for example, the difference in exposure between majority and minority item groups, or the difference in recommendation quality between user groups. The policy is updated via policy gradient to maximize the expected cumulative reward $\mathbb{E}[\sum_t \gamma^t R_t]$, where γ is a discount factor. Because the reward is evaluated at each interaction step, the agent learns to balance utility and fairness *dynamically*, adapting to feedback-loop effects that static in-processing methods cannot capture. A more principled formulation casts the problem as a constrained Markov decision process in which the long-term cumulative fairness cost is bounded by an explicit constraint, converting the trade-off hyperparameter λ into a hard guarantee rather than a soft penalty. The fairness guarantee is therefore longitudinal: the agent's policy satisfies the fairness criterion in expectation over the entire interaction horizon, not just at a single recommendation step. Figure 1.21 illustrates the agent-environment loop.

Ranking-optimization-based methods

Ranking optimization is an approach that encodes the fairness criterion directly into the ranking loss used to train the recommender, so that the model learns to produce fair ranked lists as part of its normal training procedure without any post-processing step. The key insight is that a ranked list is a structured allocation of attention across items, and fairness must be defined and enforced at the level of this ranked output rather than at the level of individual item scores.

Ranking-optimization methods ensure fairness by encoding the fairness criterion directly into the ranking loss used to train the recommender, rather than adding it as a post-hoc regularizer or reward. The central observation is that a ranked list is not merely an ordered set of items but a structured allocation of attention: items at higher positions receive disproportionately more user exposure and clicks regardless of their intrinsic quality (position bias). A ranking-optimization method treats the ranked list as the primary object of fairness concern and defines a loss that simultaneously penalizes irrelevance and demographic imbalance in the ranked output. One canonical formulation models exposure as a probability distribution over positions and imposes that the expected exposure received by items from different demographic groups be proportional to their average relevance — a criterion called *fairness of exposure*. This is formalized as a linear programming problem over the space of doubly stochastic position-assignment matrices, whose solution yields a fair stochastic ranking policy. Another formulation embeds a demographic-balance constraint directly into the pairwise or listwise ranking objective, penalizing configurations of the ranked list in which protected items consistently appear below non-protected items of equal or lower relevance. Both formulations ensure that the model, after training on the fairness-aware ranking loss, produces ranked lists that

satisfy the target criterion at inference time without any re-ranking step. The fairness guarantee is structural: it applies to the shape of the entire ranked list rather than to individual item scores, making it particularly effective for provider-side exposure fairness. The limitation is that constrained ranking objectives often require solving a linear program or differentiating through a sorting operation at each training step, which increases computational cost and can limit scalability to large item catalogs.

1.8.3 Post-Processing Methods

Post-processing methods take the trained recommender as a black box and modify only its outputs, treating the recommendation list as a ranked set to be reordered to satisfy fairness constraints. They are computationally light and require no retraining, which has made them the most widely adopted family in industrial deployments. Wang et al. [1] divide post-processing methods into *non-parametric* and *parametric* re-ranking, depending on whether the re-ranker has tunable parameters. Figure 1.22 illustrates the post-processing architecture.

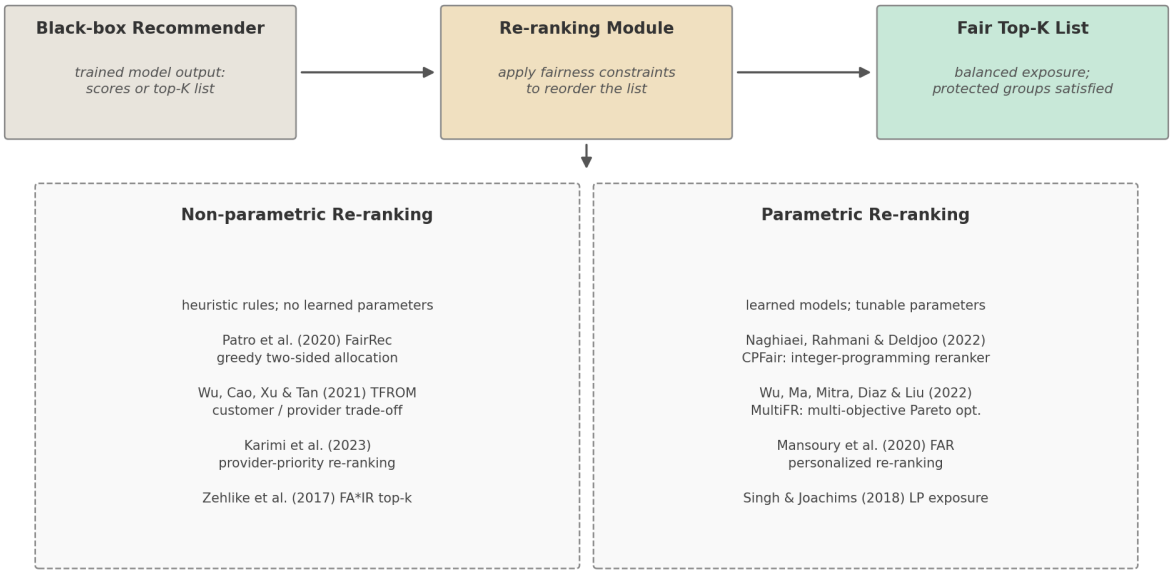


Figure 1.22: Post-processing fairness-aware methods.

Non-parametric re-ranking

Non-parametric re-ranking enforces fairness by applying a fixed, parameter-free procedure to the scored candidate list produced by the trained recommender. The procedure takes as input the ranked list of items with their associated relevance scores and applies a deterministic or greedy

reordering rule that respects a demographic balance constraint. The fairness constraint is stated explicitly and independently of any learned parameters: for example, at every prefix position k of the final list, the fraction of items from a protected group must be at least p (a minimum proportional representation requirement), or the allocation of exposure across item groups must achieve a Maximin Share guarantee. The algorithm then constructs the reordered list greedily: it iterates through candidate positions and at each step selects the highest-scoring item that does not violate the constraint. Because the procedure is deterministic and parameter-free, the fairness guarantee is exact: the output list satisfies the constraint by construction, regardless of the original recommender's behavior. The limitation is rigidity: the fixed rule cannot adapt to context, and a greedy construction may sacrifice recommendation quality more than necessary when the constraint is tight.

Parametric re-ranking

Parametric re-ranking enforces fairness through a learned, tunable reordering model that takes the candidate list and its scores as input and produces a reordered list as output. Unlike non-parametric re-ranking, the parameters of the re-ranker are optimized during training to minimize a joint objective that combines a relevance loss and a fairness loss, allowing the re-ranker to discover reordering strategies that satisfy the fairness constraint at lower cost in recommendation quality than a fixed rule would achieve. The fairness guarantee is learned rather than exact: the re-ranker minimizes demographic disparity in expectation over the training distribution, but does not offer a worst-case constraint guarantee for every possible input. More sophisticated formulations cast re-ranking as an integer programming problem in which the fairness constraint is imposed as a hard linear constraint on the space of permutation matrices; the solution is the permutation of maximum utility subject to the constraint, giving both a fairness guarantee and an optimal utility-fairness trade-off for each input. A key limitation shared by both parametric and non-parametric post-processing is that they operate on the recommender's output scores, not on its internal representations: the underlying embeddings remain biased, and a sufficiently adversarial audit of the representations would still reveal demographic correlations. This residual representation-level bias is the fundamental limitation that motivates the in-processing, representation-level approach adopted in this thesis.

1.9 Domains Affected by Bias and Unfairness

We close this chapter with a brief overview of the application domains in which bias and unfairness have been most extensively documented. For each domain we identify the bias types most prominently reported and cite the works that established the empirical evidence.

1.9.1 Music recommendation

Music recommendation exhibits both provider-demographic bias and user-side demographic bias. Ferraro et al. [51] analyzed a widely used collaborative-filtering algorithm on Last.fm and showed that male artists routinely occupy the top rank in recommendation lists while the first female artist appears, on average, only at position six or seven. Ferraro et al. [72] subsequently demonstrated that this gender imbalance is amplified through feedback loops. On the user side, Melchiorre et al. [50] documented significant accuracy disparities between male and female users on the LFM-2b dataset across several state-of-the-art collaborative-filtering models, and Bauer and Ferraro [82] proposed targeted strategies for mitigating artist gender bias.

1.9.2 Movie and book recommendation

Ekstrand et al. [71] found, in a study of book recommendation, that recommendation accuracy varied significantly across user demographic groups a clear case of user-side demographic bias. The MovieLens datasets exhibit pronounced popularity bias [69, 15], with independent and foreign films consistently under-exposed relative to mainstream blockbusters. The combination of popularity bias with demographic bias has been studied on book and movie data by Gómez et al. [83], who showed that factorization-based recommenders systematically under-recommend the content of providers based outside of dominant geographic regions.

1.9.3 Employment and job recommendation

In employment recommendation, the consequences of bias are particularly severe and have attracted regulatory attention. Lambrecht and Tucker [84] demonstrated that an apparently gender-neutral STEM job advertisement on online platforms was disproportionately shown to men, primarily for economic optimization reasons a form of provider-demographic bias propagated through ad-targeting algorithms. Imana et al. [85] confirmed in a Web Conference 2021 audit that job ads continue to be disproportionately shown along gender lines aligned with the gender concentration in specific professions, even in the absence of explicit gender targeting. Lahoti et al. [86] found, on the Xing platform, that more qualified women were often ranked lower than less qualified men, manifesting both user-side demographic bias and popularity bias toward visible candidates. The fairness-aware ranking framework of Geyik et al. [87], deployed at LinkedIn Talent Search, represents the first large-scale production countermeasure in this domain.

1.9.4 News and content recommendation

News and content recommendation is the principal vehicle through which filter bubbles and echo chambers form [88]. The interplay of three biases is particularly visible in this domain: exposure bias (users see only a subset of available news), conformity bias (users engage preferentially with

content aligned with prior views), and feedback-loop bias (the algorithm amplifies these tendencies through repeated training cycles). Bakshy et al. [65] provided an early large-scale empirical analysis on Facebook; Wang et al. [89] note in their recent survey of trustworthy recommendation that news platforms increasingly need to balance personalization with diversity to avoid representational harms.

1.9.5 Educational (learning) recommendation

Educational recommendation systems suggest courses, learning resources, or study paths to learners, and exhibit two principal forms of bias. First, cold-start bias disproportionately affects new learners who have not yet accumulated interaction history, leading to generic suggestions that fail to account for their background knowledge, learning objectives, or pace [41, 43]. Second, popularity bias and demographic bias intersect: Marras et al. [90] showed that course recommendations on a real-world platform fail to deliver equal learning opportunities across learners with different backgrounds and proposed an individual-fairness-aware re-ranking strategy. At the provider side, Gómez et al. [91] documented severe geographic imbalance in MOOC recommendation courses produced outside the United States are systematically under-recommended and Gómez et al. [92] proposed cross-continent provider-fairness re-rankers that mitigate this disparity. The stakes in educational recommendation are high: biased suggestions affect learning outcomes and career trajectories, making this domain an important target for the fairness methods developed in this thesis.

1.9.6 Other high-stakes domains

Bias and fairness concerns have additionally been documented in healthcare recommendation [93] and in financial-services and e-commerce settings. As emphasized by recent surveys [4, 1, 67], the ethical, economic, and regulatory significance of fairness in these high-stakes domains is rapidly growing the European Union AI Act of 2024 classifies several of them, including employment, as high-risk applications providing additional motivation for the methodological contributions developed in the present thesis.

1.10 Summary

In the first part, the survey of recommendation algorithms revealed that the latent embedding space, shared across virtually all modern architectures, is not a neutral mathematical object but a fairness-critical substrate in which sensitive attributes and popularity signals become structurally entangled with task-relevant preference information, regardless of the underlying model family or architectural complexity. In the second part, the expanded taxonomy of bias types, the articulation of fairness perspectives, and the survey of mitigation methods across the three intervention stages all converge on

a single observation: pre-processing methods cannot reach representation-level bias, post-processing methods leave the underlying embeddings structurally intact, and in-processing methods offer the most favorable accuracy-fairness trade-offs precisely because they intervene on the representations themselves during training. These observations collectively motivate the representation-level approach adopted throughout this thesis. In the next chapter, the discussion will narrow to the specific problem of two-sided fairness, reviewing the state of the art in joint consumer-provider fairness frameworks and identifying the precise methodological gaps that the proposed contributions are designed to address.

Chapter 2

Related Works on Two-Sided Fairness in Recommender Systems

This chapter presents a focused review of the state of the art in fairness-aware recommendation, organized around the four dimensions that frame the contributions of this thesis. Whereas the preceding chapter formalized the general taxonomy of bias and outlined the three intervention stages of fairness-aware methods, the present chapter narrows the discussion to the precise problems on which this dissertation operates. Section 2.1 reviews the literature on mitigating consumer demographic bias and content provider demographic bias. Section 2.2 reviews the literature on item popularity bias, organized by methodological approach rather than by intervention stage. Section 2.3 then turns to the central object of the chapter, namely the family of two-sided stakeholder fairness frameworks that address consumer-side and provider-side concerns jointly within a single recommendation model. Each framework is examined in terms of its methodological positioning, technical approach, evaluation protocol, and reported empirical results, and the section is supported by a series of comparative tables that systematize the landscape. Section 2.4 distills the structural limitations that emerge from this comparative analysis and articulates the research gaps that motivate the methodological direction adopted in the remainder of this dissertation.

2.1 Related Works on Mitigating User and Item Provider Demographic Bias

Demographic bias in recommender systems manifests symmetrically on the two stakeholder sides. On the consumer side, users belonging to historically under-served demographic groups (typically defined by gender, age, race, or socio-economic status) receive recommendations of systematically lower utility than users in advantaged groups, in contradiction with the principle that recommendation quality should be invariant to demographic identity. On the content provider side, creators

belonging to under-represented demographic groups (typically female artists in music, or female authors in literature) receive systematically reduced exposure in top-K recommendation lists, in contradiction with the principle that exposure should be commensurate with relevance rather than with the creator's demographic identity. The literature has investigated these two manifestations through partially overlapping but methodologically distinct lines of work, which we review in turn.

2.1.1 Mitigating User Demographic Bias

Approaches to consumer-side demographic bias have evolved along the three intervention stages introduced in Chapter 1. Pre-processing techniques operate on the training data prior to learning. Kamiran and Calders [94] introduced the canonical pre-processing toolkit, comprising re-labeling, re-weighting, and re-sampling, which has subsequently been adapted to recommendation by several authors. Boratto et al. [95] combined upsampling of under-represented user-item interactions with a pair-wise regularizer and demonstrated reductions in disparity at modest accuracy cost. More recently, Chen et al. [96] proposed FDA, a fairness-oriented data augmentation framework that constructs additional training instances to rebalance the demographic distribution and that is applicable to any embedding-based recommender; their evaluation reported improvements in group fairness across MovieLens and Last.fm benchmarks but conceded that representation-level leakage remained measurable, since pre-processing cannot reach the embeddings learned during training.

In-processing methods constitute the methodologically richest family. Regularization-based approaches add a fairness penalty to the standard recommendation loss. Yao et al. [53], in their seminal NeurIPS work, introduced four pairwise and value-based fairness penalties for collaborative filtering, defined respectively over the disparity in average prediction, in average absolute disparity, in under-estimation, and in over-estimation across demographic groups. Their MovieLens evaluation reported a reduction of inter-group rating disparity by approximately 30 percent at a less than 3 percent accuracy cost; however, the penalties operate only on the predicted scores and leave the underlying user embeddings free to encode the protected attribute. Kamishima et al. [54] extended probabilistic matrix factorization with an information-theoretic independence regularizer that minimizes the mutual information between predictions and the sensitive attribute, with comparable accuracy-fairness trade-offs on the SUSHI and Flixster datasets.

Adversarial approaches recast debiasing as a minimax game between the recommender (the generator) and a sensitive-attribute discriminator that attempts to recover the protected attribute from the learned embeddings. Beutel et al. [23] introduced this paradigm to the fairness-in-machine-learning literature, and Bose and Hamilton [97] extended it to graph-based recommendation through compositional fairness constraints, reporting reductions of attribute recoverability from above 0.85 to approximately 0.60 AUC on MovieLens and Reddit. Wu et al. [81] introduced a graph-based learning of fair representations for recommendation that combines adversarial filtering with a graph autoencoder; their evaluation on MovieLens-1M reported AUC values for gender inference in the

range of 0.55 to 0.60, an improvement over earlier regularization approaches but still distant from the theoretical chance level of 0.5. Wei and He [98] proposed CLOVER, a comprehensive fair meta-learned recommender that combines a multi-task adversarial mechanism with meta-learning to simultaneously target individual, counterfactual, and group fairness on the consumer side, illustrating that combining adversarial learning with auxiliary objectives can outperform pure adversarial filtering, while still leaving a residual leakage gap observable under the best reported configurations.

A methodologically distinctive entry in this family is the framework of Liu et al. [99], referred to as DALFRec (Dual-side Adversarial Learning based Fair Recommendation). DALFRec is labelled *dual-side* not because it targets two different fairness objects, but because it applies the adversarial mechanism on two embedding spaces in order to remove a single sensitive attribute. The authors first conduct a statistical analysis demonstrating that items' content attributes carry a latent association with users' sensitive attributes, so that debiasing the user embedding alone leaves a residual inference channel through the item embedding. They then deploy two gender discriminators, one on user embeddings and one on item embeddings, both trained to predict the same user-side sensitive attribute, while the recommender is jointly trained to fool both through gradient reversal. DALFRec is therefore a user-side demographic debiasing framework with a dual-encoder architecture; the item-side branch exists to close the leakage channel of user gender that propagates through user-item interactions, not to address an independent item-side fairness object such as popularity or provider exposure. As a representation-level adversarial method targeting user gender, DALFRec is the closest methodological cousin to the contributions developed in this thesis, and it serves as the principal in-processing baseline in the empirical chapters that follow.

Causal-inference approaches view consumer bias as the product of confounding between user behavior and the sensitive attribute, and aim to disentangle the two through structural causal models or counterfactual reasoning. Boratto et al. [100] proposed a counterfactual graph-augmentation framework that generates counterfactual user-item interactions consistent with an explicit causal graph, achieving reductions in consumer unfairness on MovieLens and LastFM-1B comparable to those of CLOVER, but at the cost of requiring a credible causal model whose validity is itself an open question.

Post-processing methods modify the trained recommender's output through re-ranking. Most consumer-side post-processing approaches re-order the top-K list to equalize average recommendation quality across demographic groups. Although operationally simple, these approaches share a common structural limitation: they treat the recommender as a black box and do not intervene in the embeddings, which means that the underlying representations continue to encode the sensitive attribute and may leak it through any downstream pipeline that re-uses the embedding.

2.1.2 Mitigating Item Provider Demographic Bias

Whereas consumer demographic bias has attracted a sizeable literature, the analogous problem on the content provider side has received markedly less attention, and the body of dedicated methods is correspondingly thinner. Ferraro et al. [51], in their study, conducted the most influential audit of artist gender bias in music recommendation. Using a widely adopted collaborative-filtering algorithm on the Last.fm-360K dataset, they showed that male artists routinely occupy the top position in recommendation lists while the first female artist appears, on average, only at position six or seven, despite female artists representing approximately 25 percent of the catalog. They additionally demonstrated that a simple post-hoc re-ranker increases female artist exposure but only at a significant accuracy cost, and that the re-ranker leaves the underlying biased representations unchanged. Their later simulation study [72] showed that this imbalance is amplified over multiple feedback cycles, reinforcing the case for representation-level interventions.

Melchiorre et al. [50] reported, on the LFM-2b dataset, that several state-of-the-art collaborative-filtering models systematically under-represent female artists in the top-K, with the most severe imbalance observed for matrix factorization with implicit feedback. Bauer and Ferraro [82] surveyed strategies for mitigating artist gender bias and grouped existing approaches into re-ranking, regularization, and post-hoc exposure-balancing methods, noting that none of the surveyed approaches operates at the level of the learned item embeddings.

On the book side, Ekstrand et al. [71] found in their study of book recommendation, that the proportion of female authors recommended is consistently below their proportion in the catalog across several algorithm families, and that recommendation accuracy itself varied significantly across user demographic groups, providing one of the earliest empirical demonstrations that consumer-side and provider-side demographic biases co-occur within the same system.

A common observation across these studies is that the existing methodological repertoire for provider demographic bias is dominated by post-hoc re-ranking. Singh and Joachims [101] formalized the principle that an item's exposure should be proportional to its relevance and proposed a linear-programming-based re-ranker that enforces this constraint; the framework can be applied to provider demographic fairness through a suitable definition of groups but operates entirely on the ranked output. Biega et al. [77] proposed an amortized fairness re-ranker that distributes attention equitably across providers over a sequence of ranked lists. Zehlike et al. [102], introduced FA*IR, a top-K re-ranker that enforces a target proportion of protected items. While each of these methods is effective at its target metric, none modifies the embedding space, and their efficacy is therefore conditional on the trained model not being heavily contaminated by the protected attribute. As Karimi et al. [103] noted in their recent survey of provider-fair re-rankers, the dominance of post-hoc methods has left the representation-level treatment of provider demographic bias as a clear methodological gap.

2.2 Related Works on Mitigating Item Popularity Bias

Item popularity bias refers to the tendency of a recommender to allocate a disproportionate share of recommendation slots to a small fraction of head items, while the long tail of niche items receives systematically lower exposure than their actual relevance would warrant. Klimashevskaia et al. [15] reported that popularity bias has been observed in virtually every recommendation algorithm family, and characterized the phenomenon along three dimensions: its causal origin in the data generation process, its propagation through the learning algorithm, and its amplification through feedback loops in deployed systems. Mitigation methods are reviewed below by methodological approach rather than by intervention stage, since the most informative groupings cut across the pre-processing, in-processing, and post-processing partition.

Regularization-based methods add a popularity-related penalty term to the standard recommendation objective. Abdollahpouri et al. [104] introduced a regularized matrix-factorization model that penalizes the learning of popularity-correlated factors and reported a measurable shift of exposure toward the long tail, although at a substantial accuracy cost on MovieLens-1M and the MovieLens 20M benchmarks. Subsequent regularization variants have explored different parametrizations of the penalty, with broadly similar conclusions: regularization can dampen exposure concentration, but at a non-trivial utility cost that grows with the strength of the regularizer.

Re-weighting and sampling approaches modify the contribution of head and tail items during learning without altering the model class. The classical approach in implicit-feedback collaborative filtering samples negatives uniformly from the items the user has not interacted with, which structurally biases learning toward popular items. Popularity-aware negative sampling re-weights candidates so that the model is encouraged to discriminate among items of comparable popularity rather than to always learn that popular items are likely positives. Inverse propensity scoring, originally proposed in the causal inference literature and adapted to recommendation by several authors, re-weights observed interactions by the inverse of the propensity of being exposed to the item, thereby producing an estimator of user preferences that is invariant to the exposure mechanism. Although effective in the linear regime, propensity-based methods are sensitive to estimation error in the propensity model and tend to introduce high variance in deep recommenders.

Causal and counterfactual approaches have become prominent. Wei et al. [105] proposed MACR (Model-Agnostic Counterfactual Reasoning for Popularity Bias), which separates the direct effect of item popularity on the predicted score from the indirect effect mediated by the user-item match and removes the former through counterfactual inference. On MovieLens-10M and the Adressa news dataset, MACR reported improvements in recall and NDCG over the standard backbone while substantially reducing popularity skew, and it has become a standard baseline for popularity debiasing. Zheng et al. [106] proposed DICE (Disentangling User Interest and Conformity for Recommendation with Causal Embedding), which factorizes the user embedding into an interest

component and a conformity component, the latter capturing the user's tendency to follow popular items, and computes predictions only from the interest component. DICE reported improvements in recall on MovieLens-10M and Netflix when popularity bias is severe, but the method requires a causal graph that may not transfer across domains and its effectiveness depends on the accuracy of the conformity supervision.

Adversarial approaches, by analogy with the consumer-side line of work, treat popularity as a sensitive attribute and train a popularity discriminator that attempts to predict the popularity tier of an item from its embedding; the recommender is then trained to confuse the discriminator. Zhu et al. [70] proposed DPR (Debiasing Pairwise Ranking) and demonstrated, on MovieLens, Yelp, and Pinterest data, that adversarial debiasing reduces popularity-driven exposure disparity while preserving the bulk of recommendation utility, although the residual leakage that characterizes vanilla adversarial training in the consumer-side literature is also observed in this setting.

Meta-learning and reinforcement-learning approaches form a more recent strand. ADB, the end-to-end meta-learning debiasing framework, learns to identify and downweight biased interactions during training without explicit specification of the bias mechanism; on MovieLens-1M and Amazon datasets, ADB reported improvements in both accuracy and exposure fairness over fixed-policy baselines, but the meta-learning loop adds substantial computational overhead. UOF tackles popularity bias from the user side by ensuring balanced exposure of recommended items across user groups, using reinforcement learning to optimize fairness; UOF improves cross-group exposure equity on MovieLens-1M but at a notable accuracy cost.

Calibration and re-ranking approaches re-order the output of a trained model to enforce a target exposure distribution. Steck [107] introduced calibrated recommendations, which re-rank the top-K so that the genre or popularity distribution of the recommended set matches the user's historical distribution; calibration is conceptually elegant and computationally cheap but does not by itself address representation-level bias. Abdollahpouri et al. [69], at the FLAIRS conference, proposed a personalized re-ranking approach in which the strength of popularity correction is adapted to each user's tolerance for popular content, reporting improvements in long-tail exposure with limited utility loss on MovieLens. Mansoury et al. [108], at RecSys 2020, introduced FAR, a personalized fairness-aware re-ranking algorithm that targets long-tail exposure across multiple recommendation algorithms. xQuAD-based re-rankers extend the diversification framework originally proposed for search to the popularity-bias setting, balancing relevance with coverage of unpopular items.

As a synthesis, the most favorable trade-offs between accuracy and popularity fairness are currently produced by methods that intervene during training, especially causal and adversarial formulations, but the literature is dominated by methods that target popularity in isolation, without simultaneously addressing consumer demographic bias. The next section turns to the body of work that has attempted to overcome this isolation by treating both stakeholder sides jointly.

2.3 Two-Sided Stakeholder Fairness Frameworks

Two-sided stakeholder fairness frameworks address consumer-side and provider-side concerns as *distinct fairness objects* within a single recommendation model. Burke [109] first articulated the conceptual case for multi-sided fairness in recommendation, arguing that a platform that improves consumer fairness in isolation may inadvertently exacerbate provider unfairness, and vice versa, and that the two dimensions must therefore be optimized jointly. The years that followed produced a small but methodologically diverse body of frameworks. We review them below grouped by methodological family rather than chronologically, since the families exhibit substantially different trade-off structures, and we close the section with a series of comparative tables that systematize the landscape.

Allocation-Based and Game-Theoretic Frameworks

Patro et al. [110] introduced FairRec, the first formal two-sided fairness framework in the recommendation literature. FairRec recasts recommendation as a constrained fair allocation problem in the spirit of cooperative game theory: items are treated as indivisible goods to be distributed among consumers under two fairness constraints. For producers, the framework imposes a Maximin Share (MMS) guarantee, ensuring that every producer receives at least a target minimum level of exposure across the platform; for consumers, the framework imposes an Envy-Free up to One item (EF1) guarantee, meaning that no consumer envies the bundle assigned to another by more than the contribution of a single item. The allocation is computed by a greedy round-robin procedure that, at each step, gives the next user her most preferred remaining item subject to the producer-side cap. The method is purely post-hoc, in the sense that it operates on top of any precomputed relevance scores.

In their evaluation on MovieLens-1M, Last.fm, and Google Local, the authors reported that FairRec attains its targeted MMS and EF1 guarantees, with producer exposure inequality reduced by approximately 30 percent on MovieLens-1M and 40 percent on Google Local compared with a relevance-only baseline. The framework's principal structural limitation is that it does not exploit any sensitive attribute on the consumer side and does not address consumer demographic bias as understood in the recommendation fairness literature; it equalizes the quality experienced by consumers under EF1 but does not equalize quality across demographic groups, and it leaves the underlying relevance scores, and the embeddings that produced them, entirely untouched.

A complementary line of work treats the two-sided platform as a strategic ecosystem in which consumers and providers act as utility-maximizing agents whose behavior adapts to the recommender's policy. Chen et al. [111] proposed a game-theoretic two-sided fairness framework that models the recommendation problem as an equilibrium-seeking process between user satisfaction and provider exposure, and showed that an equilibrium can be reached in which both sides achieve at

least their reservation utility. Chen and colleagues introduced an incentive-aware framework based on constrained optimization with adaptive weighting, in which the weight of each fairness objective is determined by the marginal incentive each stakeholder has to remain on the platform. These frameworks are conceptually elegant and produce intuitive policy descriptions, but their empirical evaluation has so far been limited to synthetic or small-scale benchmarks, and they do not address representation-level debiasing: the equilibrium they reach concerns the allocation of exposure, not the geometry of the embeddings.

Post-Processing Re-Ranking Frameworks

Post-processing re-rankers form the most populous family of two-sided fairness frameworks. Wu et al. [112] proposed TFROM (Two-sided Fairness-aware Recommendation Model), a re-ranker that adjusts the top-K list to enforce two constraints simultaneously: a group-level exposure constraint for providers, configurable either as uniform exposure across provider groups or as exposure proportional to a quality weight, and a quality-distribution constraint for consumers, which spreads the recommendation quality loss caused by the re-ranking equally across all users. The re-ranking is computed by a position-aware exposure model that accumulates the discounted attention each provider has received so far and re-injects under-exposed providers into subsequent slots. On MovieLens-1M, Amazon-Beauty and Google Local, the authors reported that TFROM reduces provider exposure inequality by 40 to 60 percent relative to the unconstrained baseline while keeping the per-user NDCG@10 drop below 8 percent.

Naghiaei et al. [113] proposed CPFair (Consumer and Producer Fairness re-ranking), an optimization-based re-ranking approach that integrates fairness constraints from both the consumer and the producer side into a single joint objective formulated as a mixed-integer linear program. Consumer-side fairness in CPFair is operationalized through user groups; in the published experiments, these are defined by activity level (active versus inactive users), while producer-side fairness is defined through exposure across item groups. The authors reported that on eight benchmark datasets spanning movies, books and points of interest, CPFair improves both consumer-side and producer-side fairness without reducing overall recommendation quality, establishing CPFair as one of the strongest recent re-ranking baselines for two-sided fairness.

Wu et al. [114] proposed FairSort, a re-ranking model that re-frames the two-sided fairness trade-off as a runway-and-velocity allocation problem rather than a knapsack. Each item in a candidate list is treated as an entity traveling on a runway with an associated velocity, and the re-ranking is obtained by integrating these velocities over a fixed time horizon. The authors complement this perspective with a Minimum Utility Guarantee that bounds the personalized recommendation utility loss through a binary-search procedure, and provide online and offline variants of the algorithm. The reported experiments on three real-world datasets indicate that FairSort secures fairness for both providers and users with more reliable personalization than prior knapsack-based formulations.

The structural property shared by all three frameworks above is that they intervene on the ranked output rather than on the learned embeddings. The trained recommender continues to encode the sensitive attribute, so the re-ranker is required at every inference, and the embeddings themselves remain exposed to attribute-inference probes irrespective of the rebalancing applied to the final list.

Multi-Objective Optimization and Intersectional Frameworks

Wu et al. [11] proposed Multi-FR, a multi-objective optimization framework that jointly optimizes recommendation accuracy, consumer group fairness (defined over a demographic attribute such as gender or age), and producer group fairness (defined over an item-side attribute such as popularity or genre). The framework approximates the non-smooth ranking metrics through smooth surrogates and applies a Frank-Wolfe solver to identify Pareto-optimal trade-off points along the three-dimensional objective space; the user of the framework can then select a point on the Pareto front according to platform-specific priorities. On MovieLens-1M, Amazon-Beauty and Yelp, Multi-FR reported a measurable improvement in consumer group fairness compared with single-objective baselines, with a typical reduction of recommendation quality disparity across gender groups by 25 to 40 percent depending on the Pareto point, and a comparable improvement in producer exposure equity. Its principal limitation is that the framework, although operating during training through smooth surrogates, does not intervene in the geometry of the embedding space and does not impose any constraint that the sensitive attribute be removed from the user or item representations; the gender-inference AUC on the produced embeddings remains substantially above the chance baseline.

Wang et al. [115] proposed ITFR (Intersectional Two-sided Fairness Recommendation), which extends the two-sided fairness paradigm to intersectional groups defined by the overlap of user-side and item-side attributes. The authors show empirically that two-sided fairness can be achieved at the aggregate level while still failing on specific intersectional subgroups. ITFR addresses this through a sharpness-aware loss that identifies disadvantaged intersectional groups, a collaborative loss-balancing scheme that maintains consistent discrimination ability across them, and a predicted-score normalization step that aligns positive predicted scores for fair treatment of positives across intersectional groups. Experiments on three public datasets show that ITFR consistently reduces intersectional unfairness compared with non-intersectional baselines. The intersectional constraints are nevertheless enforced at the level of the ranked list rather than the embedding space.

Calibration Frameworks

Wang and colleagues proposed PCT, a calibration framework that aligns the recommendation list of each user with two reference distributions: the user's own interest distribution, estimated from her historical interactions, and a target item-side distribution that encodes the desired exposure across

item groups. The alignment is computed through a distribution-matching procedure that re-orders the top-K so that the discrepancy between the recommended distribution and the two references is minimized. PCT is a generalization of the calibration principle introduced by Steck [107] to the two-sided setting. On MovieLens-1M and Last.fm-1B, PCT reported a measurable reduction in genre-level exposure imbalance, in the range of 20 to 35 percent, with a modest accuracy cost. The framework does not require explicit access to consumer sensitive attributes, which makes it attractive for deployments in which demographic data are unavailable, but the same property limits its ability to certify that consumer demographic disparities are removed.

Constrained Optimization and Theoretical Frameworks

A more recent strand reframes two-sided fairness as a constrained optimization problem and provides either online algorithms or theoretical characterizations of the achievable trade-offs. Chen et al. [116] proposed FAIR with its accompanying online algorithm FORM, formulating multi-sided fairness over platform, items, and users through a constrained optimization that flexibly balances stakeholder interests, and providing a low-regret algorithm that performs real-time learning concurrently with fair recommendation. The authors provide theoretical guarantees on regret and a numerical case study on real-world data indicating that FORM maintains platform utility while delivering target levels of fairness on the user and item sides.

Greenwood et al. [117] provided a theoretical analysis of user-item fairness trade-offs in recommendation, identifying the regimes in which user-side and item-side fairness can be jointly satisfied at little or no cost to platform utility, and the regimes in which an unavoidable trade-off emerges. Their analysis is grounded in a theoretical model with user and item fairness constraints expressed through unfairness penalties and slack parameters, and yields existence conditions for the so-called free-fairness regime; the contribution is methodological rather than algorithmic and is best read as a complement to applied frameworks such as FAIR-FORM, CPFair, and FairSort, since it clarifies the regime in which these frameworks operate without loss of platform utility.

A common property of the frameworks reviewed in this subsection is that they target the level of the recommendation list or the allocation policy. None operates on the user or item embeddings during training, and none provides representation-level guarantees that survive an attribute-inference probe.

2.3.1 Comparative Synthesis

To consolidate the picture, this section summarizes the surveyed two-sided stakeholder fairness frameworks along three axes: methodological positioning (Table 2.1), the datasets and evaluation protocol adopted (Table 2.2), and the principal evaluation metrics (Table 2.3). The proposed contributions of this thesis are deferred to the chapters that introduce them and are therefore not

included in the comparative tables below.

Table 2.1: Comparison of two-sided stakeholder fairness frameworks.

Model Name	User Fairness	Item Fairness	Use of Sens. Attr.	Proposed Approach	Sensitive Attr. (User)	Sensitive Attr. (Item)
FairRec [110]	✓	✓		Fair allocation (MMS, EF1) via greedy round-robin	-	Provider groups
TFROM [112]	✓	✓	✓	Post-processing re-ranking with position-aware exposure	-	Provider groups
CPFair [113]	✓	✓	✓	Mixed-integer optimization for re-ranking	User activity	Item groups
FairSort [114]	✓	✓	✓	Runway-and-velocity re-ranking with minimum utility guarantee	User groups	Provider groups
Multi-FR [11]	✓	✓	✓	Multi-objective optimization (MOO) with smooth ranking	Gender, Age	Popularity, Genre
ITFR [115]	✓	✓	✓	Sharpness-aware loss with intersectional groups	Gender, Age	Genre, Popularity
PCT [118]	✓	✓		Distribution matching for calibration	-	Genre
Incentive-Aware [111]	✓	✓		Game-theoretic equilibrium analysis	-	Provider groups
FAIR-FORM [116]	✓	✓	✓	Constrained optimization with online low-regret algorithm	User groups	Item groups
Greenwood et al. [117]	✓	✓	✓	Theoretical analysis of fairness trade-offs	User groups	Item groups

Table 2.2: Datasets used in the evaluation of two-sided stakeholder fairness frameworks.

Framework	Datasets	Domain	Users	Items	Interactions	Sensitive Attr.
FairRec	ML-1M, Last.fm, Google Local	Movies, Music, POI	6040/1892/5073	3952/17632/6742	∼1M/∼92K/∼71K	Provider groups
TFROM	ML-1M, Amazon-Beauty, Google Local	Movies, E-comm., POI	6040/22363/5073	3952/12101/6742	∼1M/∼198K/∼71K	Provider groups
CPFair	ML-1M, Epinions, BookCrossing, others (8 in total)	Movies, Reviews, Books, POI	Varies	Varies	Varies	User activity, item groups
FairSort	Three public datasets	Mixed	Varies	Varies	Varies	User and provider groups
Multi-FR	ML-1M, Amazon-Beauty, Yelp	Movies, E-comm., Reviews	6040/22363/30887	3952/12101/18995	∼1M/∼198K/∼860K	Gender, age, popularity, genre
PCT	ML-1M, Last.fm-1B	Movies, Music	6040/120322	3952/585095	∼1M/∼1B	Genre
Incentive-Aware	Synthetic + ML-1M	Synthetic, Movies	Varies	Varies	Varies	Provider groups
Game-Theoretic	ML-100K, ML-1M	Movies	943/6040	1682/3952	100K/1M	Provider groups
ITFR	Three public datasets including Tenrec	Mixed	Varies	Varies	Varies	Intersectional user-item groups
FAIR-FORM	Real-world case study	Mixed	Varies	Varies	Varies	User and item groups
Greenwood et al.	Theoretical	N/A	N/A	N/A	N/A	User and item groups

Three patterns emerge from Tables 2.1, 2.3, and 2.4, and frame the comparative argument of the chapter. First, the dominant intervention stage across the surveyed two-sided stakeholder frameworks is post-processing or output-level optimization (FairRec, TFROM, CPFair, FairSort,

Table 2.3: Evaluation metrics employed by two-sided stakeholder fairness frameworks.

Framework	**Utility Metrics**	**Consumer Fairness Metrics**	**Provider Fairness Metrics**
FairRec	NDCG@K	Envy-Freeness up to 1 item (EF1)	Maximin Share (MMS) ratio, exposure distribution
TFROM	Precision@K, NDCG@K	Average quality loss per user	Group exposure ratio, Gini index
CPFair	NDCG@K, Precision@K	Quality disparity across user activity groups	Item-group exposure equality
FairSort	NDCG@K, utility loss	User-side fairness disparity	Provider exposure ratio
Multi-FR	NDCG@K, MRR	Inter-group quality disparity (gender/age)	Group exposure ratio (popularity/genre)
PCT	NDCG@K	Calibration KL divergence (user distribution)	Calibration KL divergence (item-group distribution)
Incentive-Aware	NDCG@K	Group utility floor	Group exposure floor
Game-Theoretic	NDCG@K	Equilibrium consumer utility	Equilibrium provider exposure
ITFR	NDCG@K, Recall@K	Worst-case intersectional disparity	Worst-case intersectional exposure
FAIR-FORM	Regret bound, platform utility	User-side fairness gap	Item-side fairness gap
Greenwood et al.	Theoretical bounds	User fairness criterion	Item fairness criterion

PCT, Game-Theoretic, FAIR-FORM) and multi-objective formulations that, although applied during training, operate on smooth surrogates of the ranking metrics rather than on the embeddings themselves (Multi-FR, ITFR). No surveyed two-sided stakeholder framework intervenes at the level of the learned user and item embeddings.

Second, the fairness criteria adopted across frameworks are heterogeneous on both sides, and the choice of criterion on the user side is a major source of variation. Some frameworks adopt demographic-group fairness, in which the user side is partitioned by a sensitive attribute such as gender or age and equality of recommendation quality is enforced across these groups; this is the case for Multi-FR, ITFR, and partially for FAIR-FORM. Other frameworks adopt attribute-agnostic notions: envy-freeness up to one item in FairRec, quality-distribution equalization in TFROM, calibration to the user's own historical distribution in PCT, equilibrium or reservation utility in the game-theoretic and incentive-aware frameworks. The attribute-agnostic notions equalize a property of the recommendation experience across users without reference to demographic identity and therefore do not, on their own, certify that consumer demographic disparities have been removed.

Third, the representation-level treatment of consumer demographic bias is not entirely empty, but the only methodologically mature framework in this space is DALFRec (Section 2.1.1), which is user-side in fairness target despite its dual-encoder architecture. The space of frameworks that simultaneously enforce consumer demographic fairness and item-side fairness at the representation level is, to the best of our knowledge, empty. This methodological vacuum is the principal gap identified by the present chapter and the principal motivation for the contributions developed in the remainder of the dissertation.

2.4 Gaps in the State of the Art and Thesis Motivations

The comparative analysis of the preceding section makes it possible to articulate, in a single catalog, the principal methodological deficits of the existing state of the art and the corresponding directions

Table 2.4: Fairness criteria adopted by two-sided stakeholder fairness frameworks, decomposed by stakeholder side.

Framework	**User-side fairness criterion**	**Item-side fairness criterion**
FairRec	Envy-freeness up to one item (EF1) over the assigned bundles; no sensitive attribute used	Maximin share (MMS) guarantee over provider exposure
TFROM	Equal distribution of quality loss across all users (quality-distribution equalization); no sensitive attribute used	Provider group exposure (uniform or proportional to a quality weight)
CPFair	User activity groups (active vs. inactive users); equality of recommendation quality across these groups	Item groups (typically popular vs. unpopular); equality of exposure across these groups
FairSort	User group disparity in personalized utility, bounded below by a minimum utility guarantee	Provider group exposure equity through runway-and-velocity allocation
Multi-FR	Demographic group fairness (gender, age); equality of recommendation quality across groups	Item-group fairness over popularity or genre; equality of exposure across groups
PCT	Calibration to the user's own historical interest distribution; no sensitive attribute used	Calibration to a target item-group exposure distribution (genre)
Incentive-Aware	Group utility floor under strategic-agent behavior; no demographic sensitive attribute	Group exposure floor under strategic-agent behavior
Game-Theoretic	Equilibrium consumer utility (reservation utility threshold); no demographic sensitive attribute	Equilibrium provider exposure (reservation exposure threshold)
ITFR	Intersectional user groups (e.g. gender age); worst-case quality across intersectional subgroups	Intersectional item groups (e.g. genre popularity); worst-case exposure across intersectional subgroups
FAIR-FORM	User-side fairness gap across user groups, enforced through a constrained optimization with online low-regret learning	Item-side fairness gap across item groups, enforced symmetrically with the user-side constraint
Greenwood et al.	User fairness criterion expressed as an unfairness penalty or slack constraint over predicted relevance	Item fairness criterion expressed symmetrically as an unfairness penalty or slack constraint over exposure

pursued by this thesis. We group these deficits under three headings. The first two concern the structural design of two-sided fairness frameworks, namely the isolation of the two stakeholder sides at the representation level and the dominance of output-level corrections. The third concerns the intrinsic limits of adversarial learning when used as the sole representation-level mechanism, and motivates the hybrid paradigm in which the adversarial objective is paired with a complementary, deterministic constraint. The two contributions developed in the remainder of this dissertation, FUIREC and ORAF, both belong to this hybrid paradigm and are positioned precisely at the

intersection of the unmet needs identified below.

2.4.1 Absence of Representation-Level Two-Sided Stakeholder Fairness

The dominant line of work in the recommendation fairness literature has investigated consumer-side and item-side biases as separate optimization problems. The methods reviewed in Sections 2.1 and 2.2 are predominantly one-sided: regularization, adversarial filtering, and causal disentanglement on the consumer side, and re-ranking, calibration, and counterfactual reasoning on the item side. The frameworks that do treat both stakeholder sides jointly, surveyed in Section 2.3, operate without exception at the level of the ranked list or the allocation policy, never at the level of the learned embeddings. As Burke [109] anticipated and Boratto et al. [95] subsequently confirmed through empirical study, optimizing for one stakeholder in isolation can amplify unfairness on the other. A consumer-only intervention that equalizes recommendation quality across demographic groups may concentrate exposure on a narrow subset of items that maximize cross-group satisfaction, thereby worsening provider exposure inequality. Conversely, a provider-only intervention that redistributes exposure across the catalog may inadvertently steer demographic groups toward stereotyped content if the catalog itself is demographically clustered. The closest representation-level precursor in this space is DALFRec [99], but DALFRec is a user-side framework with a dual-encoder architecture; both of its adversarial branches target the same user-side sensitive attribute, and the framework therefore does not enforce a separate fairness object on the item side.

The first gap addressed by this thesis is therefore the absence of a framework that handles consumer demographic bias and item-side bias as *distinct fairness objects* within a single unified objective enforced at the representation level. The first motivation of the thesis is to design in-processing frameworks in which fairness for both stakeholders is enforced directly on the user and item representations, rather than as an external corrective layer. The objective is not to append a fairness term to a pre-trained recommender, but to make fairness an architectural property of the representation space itself, so that the recommendations generated from these representations are fair by design rather than fair by correction.

2.4.2 Dominance of Output-Level Intervention

Among the two-sided stakeholder fairness frameworks surveyed, the overwhelming majority intervene at the level of the recommendation list rather than at the level of the learned embeddings. This is true of FairRec, TFROM, CPFair, FairSort, PCT, the game-theoretic and incentive-aware frameworks, and FAIR-FORM. The structural consequence of operating exclusively at the output is twofold. First, the trained recommender continues to learn embeddings in which the sensitive attribute remains correlated with the predicted scores, so the re-ranker is obliged to apply a corrective operation at every inference, which is computationally costly at deployment scale. Sec-

ond, and more fundamentally, the post-hoc paradigm leaves the embeddings themselves exposed to attribute-inference attacks, since a probe trained on the representations can still recover the protected attribute even when the recommendation list has been rebalanced. As Karimi et al. [103] observed in their survey of provider-fair re-rankers, output-level interventions cannot certify representation-level fairness; they can only patch the symptoms at the surface.

The second motivation of this thesis is therefore to operate at the representation level on both stakeholder sides, in such a way that the sensitive signal is removed from the geometric directions used to generate recommendations.

2.4.3 The Insufficiency of Adversarial Learning Alone and the Case for Hybrid Frameworks

Once the choice has been made to intervene at the representation level, the candidate mechanism most often adopted in the literature is adversarial learning, instantiated as a minimax game between an embedding generator and one or more discriminators tasked with recovering the sensitive attribute. The closest precursor in this space is DALFRec [99], which deploys two adversarial discriminators on user and item embeddings to remove the user sensitive attribute, and which is the strongest user-side adversarial representation-level baseline currently available. Despite the conceptual elegance of the adversarial formulation, the empirical and theoretical evidence accumulated over the past five years shows that adversarial training, when used as the sole debiasing mechanism, is structurally insufficient. We organize the critique under three converging arguments.

Residual representational leakage. The adversarial objective drives the generator toward distributions of embeddings that are indistinguishable to the current discriminator across the protected groups. This is a distributional, not a geometric, criterion. As Elazar and Goldberg [22] demonstrated in the broader fairness-in-machine-learning literature, a generator can fool a discriminator of a given capacity while still encoding the sensitive attribute in a form that a stronger probe, trained post-hoc on the same embeddings, can recover. The recent re-examination by Wang and Singh [119] confirms that the magnitude of this residual leakage is itself a function of the probing capacity and of the accuracy with which the protected attribute is observed.

Training instability and absence of equilibrium guarantees. The minimax optimization underlying adversarial debiasing is well known to be unstable and hyperparameter-sensitive, as documented in the foundational work of Ganin et al. [120] and subsequently observed across the fairness-aware recommendation literature. The training trajectory often fails to converge to a Nash equilibrium in which the embeddings are genuinely invariant to the sensitive attribute; instead, it oscillates between regimes in which either the generator or the discriminator dominates. In sparse-data regimes characteristic of real-world recommendation, such as the Book-Crossing benchmark,

the discriminator receives noisy gradients and the generator faces a fluctuating adversarial signal, which further degrades the reliability of the debiasing effect. The literature has therefore increasingly recognized that a deterministic, closed-form constraint, evaluated at every training step, is needed to stabilize the optimization and to prevent the utility collapse that purely adversarial models tend to exhibit when the adversarial signal is weak.

Latent invariance does not entail output-level fairness. A more subtle limitation, articulated by Beutel et al. [23] and confirmed in several subsequent empirical studies, is that fairness in the latent space does not strictly entail fairness in the recommendation output. Even when the embeddings have been rendered indistinguishable to a discriminator, the non-linear interactions of the downstream prediction layers can reintroduce disparate impacts at the ranking level. Adversarial-only frameworks, including DALFRec and CLOVER [98], consequently tend to exhibit elevated group fairness disparities at the output, even when their representational fairness scores are improved relative to unconstrained baselines.

The hybrid adversarial paradigm. These three limitations have motivated, across several adjacent areas of machine learning, the emergence of a hybrid adversarial paradigm in which the minimax objective is synergized with an auxiliary mechanism of a different mathematical character. The recurring insight, confirmed across this cross-domain literature, is that an additional constraint is consistently required to close the disentanglement gap that adversarial training alone cannot reliably close. The advantages of the hybrid paradigm, as documented across this body of work and confirmed by the empirical chapters of this thesis, can be summarized as follows. First, the auxiliary constraint provides a geometric or distributional guarantee that the adversarial objective alone cannot reach, and that survives stronger post-hoc probes. Second, when the auxiliary constraint is deterministic and closed-form, as in the case of orthogonal regularization, it remains informative even when the adversarial signal is noisy, which protects the model against the utility collapse observed in purely adversarial frameworks under sparse data. Third, when the auxiliary constraint operates at the recommendation output, as in the case of explicit individual fairness and popularity regularization, it neutralizes the residual disparate impacts that non-linear prediction layers can reintroduce, thereby providing the output-level guarantee that latent-only debiasing structurally lacks.

The third and central motivation of this thesis is therefore to combine the adversarial mechanism with a complementary constraint that closes the disentanglement gap in a verifiable way. Two such constraints are explored in the contributions that follow. The first, developed in Chapter 4 as part of the FUIREC framework, augments dual-discriminator adversarial learning with two explicit fairness constraints that operate directly at the recommendation output, namely an individual user fairness term that enforces consistent treatment for users with similar preferences and a popular-

ity regularizer that balances exposure across item groups. The second, developed in Chapter 5 as part of the ORAF framework, augments the same dual-discriminator backbone with an orthogonal regularization that enforces a strict geometric perpendicularity between bias-explicit and debiased embedding subspaces, semantically anchored by dedicated attribute predictors. In both cases, the auxiliary mechanism provides the verifiable guarantee that the distributional adversarial objective alone cannot deliver, and in both cases the empirical chapters will show that the hybrid framework strictly dominates its adversarial-only counterpart on every fairness metric considered while preserving competitive recommendation utility.

2.5 Summary

In conclusion, our review of the literature highlights a critical architectural limitation in current algorithmic fairness frameworks: while multi-stakeholder recommendation is increasingly recognized as a two-sided optimization problem, existing methods rely heavily on output-level re-ranking or soft regularizers. These approaches treat the symptoms of bias rather than its root cause, leaving learned user and item embeddings structurally entangled with sensitive or dominant historical interaction signals.

Achieving sustainable, robust, and verifiable two-sided fairness requires a fundamental shift toward purging the latent embedding space itself before recommendations are generated. By exposing these boundaries, this chapter establishes the rigorous justification for the technical contributions presented in Part II of this dissertation.

To directly address these limitations, the next chapter 3 introduces our first major technical contribution: DS-DVAE (Dual-Sided Disentangled Variational Autoencoder). Moving away from superficial output corrections, Chapter 3 establishes a "fairness by architectural design" paradigm, presenting a generative variational framework engineered to geometrically factorize user and item representations into independent task-relevant and sensitive subspaces.

Part II

Contributions

Chapter 3

Dual-Sided Fair Recommendation via Variational Disentanglement

This chapter presents DS-DVAE (Dual-Sided Disentangled Variational Autoencoder), a disentangled framework that aims to ensure two-sided fairness for both consumer and item provider by eliminating the influence of sensitive attributes from the recommendation process. To achieve this goal, DS-DVAE explicitly factorizes user and item embeddings into two orthogonal subspaces: a task subspace that captures preference-relevant information (e.g., genre preferences, listening habits), and a sensitive subspace that isolates demographic information (user gender and provider gender). Recommendations are generated solely from the task subspace, so sensitive attributes from both consumers and providers have no path to influence the output. The framework is instantiated as a variational autoencoder with factorized posteriors, auxiliary attribute decoders that supervise the sensitive subspace, and an orthogonality constraint that reinforces separation. Extensive experiments on the Last.fm-360K and MovieLens-1M datasets demonstrate that DS-DVAE substantially reduces user group fairness disparities, mitigates systematic underexposure of female artists, and achieves near-chance attribute inference resistance for both sides, all while maintaining competitive recommendation accuracy.

3.1 Introduction

The domain of user-oriented fairness primarily addresses the disparity in recommendation quality across different demographic groups. Early approaches focused on pre-processing techniques to balance datasets, yet modern research has gravitated toward in-processing methods that incorporate fairness constraints directly into the learning objective. A dominant paradigm involves the use of adversarial learning, where a discriminator attempts to predict sensitive attributes (e.g., gender or age) from user embeddings while the recommendation model aims to deceive it [121]. Similarly, regularization-based methods have been proposed to minimize the statistical dependence between

model predictions and sensitive attributes [53]. While these methods effectively reduce the performance gap between advantaged and disadvantaged user groups, often referred to as ensuring demographic parity or equal opportunity, they typically focus exclusively on the consumer side, neglecting the impact of these adjustments on the item providers. Parallel to user fairness, a significant body of literature addresses provider-side fairness, specifically concerning the "rich-get-richer" dynamic where popular items dominate exposure at the expense of the long-tail. Similar work by Singh and Joachims [101] formalized the concept of fair exposure, arguing that the visibility of an item should be proportional to its relevance. Subsequent studies have explored the mitigation of bias through post-processing re-ranking algorithms that enforce constraints on the representation of minority providers in top-k lists [77, 102]. However, the majority of existing provider-fairness literature focuses on popularity bias (promoting niche items) rather than demographic bias (e.g., gender or race of the content creator). Consequently, systems may satisfy popularity constraints while still systematically under-exposing specific demographic groups of artists or creators.

While significant advances have been made in studying fairness for either users or items in isolation, research addressing fairness for both parties together remains limited. Existing frameworks that attempt this dual optimization, such as FairRec [110] or TFROM [112], typically rely on post-hoc re-ranking strategies or greedy optimization algorithms applied to the final recommendation lists. Crucially, these methods operate on the output scores rather than the embedding representations. Because they do not intervene in the latent space, the underlying model continues to learn biased representations. This limitation has two major consequences: first, the model's internal reasoning remains biased, requiring computationally expensive corrections at every inference step; second, because the embedding representations are left unaltered, they retain significant leakage of sensitive attributes, rendering the system vulnerable to privacy inference attacks. The disentanglement of sensitive attributes from latent embeddings, within a two-sided fairness framework remains largely unexplored in current literature.

3.2 Problem Formulation

Let $U = \{u_1, u_2, \ldots, u_n\}$ denote the set of n users and $I = \{i_1, i_2, \ldots, i_m\}$ denote the set of m items. Each user u has a sensitive attribute $s_u \in \{0, 1\}$ (e.g., gender), and each item i has a provider attribute $g_i \in \{0, 1\}$ representing the provider's gender. The interaction matrix $R \in \mathbb{R}^{nm}$ contains observed ratings r_{ui}.

Our goal is to learn a recommendation function $f : UI \to \mathbb{R}$ that predicts ratings while satisfying two fairness objectives:

1. **User-side fairness:** Recommendation quality should be similar across user groups defined by s_u.

2. **Provider-side fairness:** Relevant item exposure should be balanced across provider groups defined by g_i.

Notation. Throughout this chapter, $s_u \in \{0, 1\}$ denotes the user's sensitive attribute (gender: 0 = male, 1 = female), $g_i \in \{0, 1\}$ denotes the provider's gender attribute, $\mathbf{z}_u^t$ and $\mathbf{z}_u^s$ denote the task and sensitive latent subspaces for users, and $\mathbf{z}_i^t$, $\mathbf{z}_i^s$ their item counterparts.

3.3 Fairness Definitions

Definition 1 (Consumer Fairness): A recommendation system satisfies consumer fairness if users with similar preferences but different sensitive attributes receive similar recommendations:

$$|f(u, i) - f(u', i)| \approx 0 \quad \text{when} \quad \text{Pref}(u) \approx \text{Pref}(u')$$

$$\text{and} \quad s_u \neq s_{u'}$$

Definition 2 (Item Provider Fairness): A recommendation system satisfies provider fairness if the item's exposure is determined by item relevance $\mathcal{R}(i, u)$ rather than provider gender g_i. Formally, for items i, i' with providers of different gender, where $\mathcal{R}(i, u) \approx \mathcal{R}(i', u)$ and $g_i \neq g_{i'}$:

$$\frac{P(\text{recommend } i \mid \mathcal{R}(i, u), g_i)}{P(\text{recommend } i' \mid \mathcal{R}(i', u), g_{i'})} \approx 1 \tag{3.1}$$

where $\mathcal{R}(i, u)$ captures genre compatibility, stylistic similarity, acoustic features, and user listening history. This ensures that female and male providers with equivalent item relevance receive equal recommendation exposure.

Definition 3 (Two-Sided Fairness): A system satisfies two-sided fairness if it simultaneously satisfies both consumer and provider fairness.

3.4 DS-DVAE Proposed Architecture

Our framework addresses gender bias through explicit representation disentanglement applied symmetrically to both users and items. Gender bias in recommendation systems operates on two sides: user-side bias, where users of different genders receive differential recommendation quality, and content provider-side bias, where content creators (artists) of different genders receive unequal exposure and visibility. The architecture consists of four main components: user disentangled encoder, item disentangled encoder, task-specific recommendation decoder, and auxiliary sensitive attribute decoders. By construction, the recommendation pathway uses only task-relevant representations, ensuring that gender information from both the user side and the content provider side is architecturally isolated from prediction, thereby mitigating bias along both dimensions simultaneously.

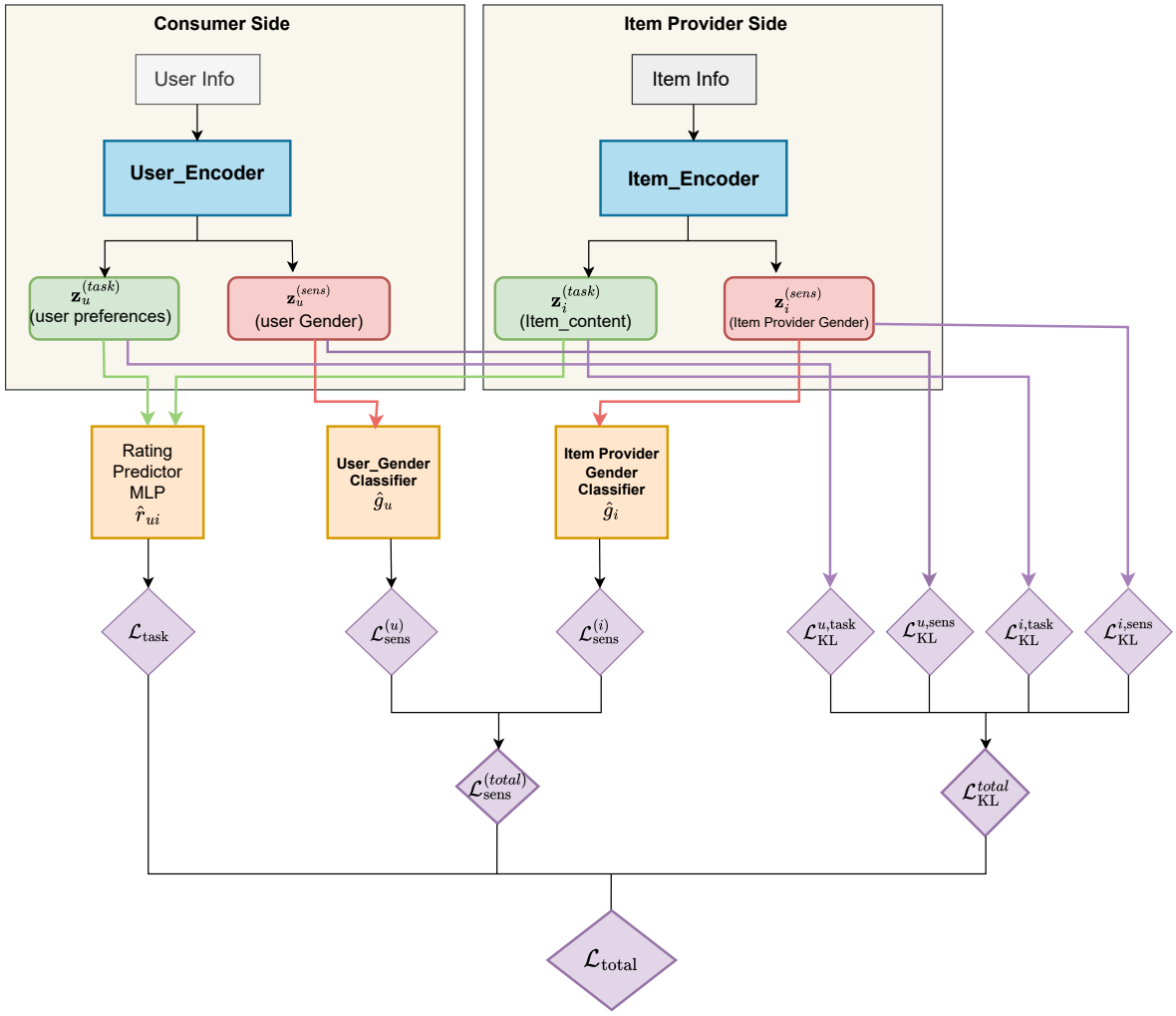


Figure 3.1: Proposed Dual-Sided Disentangled VAE (DS-DVAE) architecture.

3.4.1 User Disentangled Encoder

For each user $u \in \mathcal{U}$, we obtain an initial embedding $\mathbf{e}_u \in \mathbb{R}^d$ through a trainable embedding layer. This embedding is then encoded into two probabilistically independent latent subspaces using separate encoder networks to address user-side gender bias. The task latent subspace $\mathbf{z}_u^t \in \mathbb{R}^{d_t}$ captures user preferences, listening patterns, and behavioral characteristics relevant for recommendation quality, independent of the user's gender. The sensitive latent subspace $\mathbf{z}_u^s \in \mathbb{R}^{d_s}$ isolates demographic information, specifically user gender attributes. By separating these representations, we ensure that recommendation quality does not vary based on whether the user is male or female, thereby mitigating user-side gender bias. This decomposition is formalized through a factorized

variational posterior:

$$q_{\phi_u}(\mathbf{z}_u^t, \mathbf{z}_u^s \mid \mathbf{e}_u) = q_{\phi_u^t}(\mathbf{z}_u^t \mid \mathbf{e}_u) \cdot q_{\phi_u^s}(\mathbf{z}_u^s \mid \mathbf{e}_u) \tag{3.2}$$

where each component follows a Gaussian distribution with parameters computed by dedicated neural networks:

$$\mathbf{z}_u^t \sim \mathcal{N}(\boldsymbol{\mu}_u^t, \mathrm{diag}(\boldsymbol{\sigma}_u^t)^2) \tag{3.3}$$

$$\mathbf{z}_u^s \sim \mathcal{N}(\boldsymbol{\mu}_u^s, \mathrm{diag}(\boldsymbol{\sigma}_u^s)^2) \tag{3.4}$$

The means and log-variances are obtained through:

$$\boldsymbol{\mu}_u^t, \log \boldsymbol{\sigma}_u^t = \mathrm{MLP}_{\phi_u}^t(\mathbf{e}_u) \tag{3.5}$$

$$\boldsymbol{\mu}_u^s, \log \boldsymbol{\sigma}_u^s = \mathrm{MLP}_{\phi_u}^s(\mathbf{e}_u) \tag{3.6}$$

where $\mathrm{MLP}_{\phi_u}^t$ and $\mathrm{MLP}_{\phi_u}^s$ denote multi-layer perceptrons with parameters ϕ_u for task and sensitive encoders respectively.

3.4.2 Item Disentangled Encoder

Symmetrically to the user-side approach, each item $i \in \mathcal{I}$ is processed through an embedding layer to obtain an initial item embedding $\mathbf{e}_i \in \mathbb{R}^d$. This embedding is subsequently encoded into two independent latent subspaces using separate encoder networks to address provider-side gender bias. The task latent subspace $\mathbf{z}_i^t \in \mathbb{R}^{d_t}$ captures content-relevant features such as genre, acoustic characteristics, lyrical themes, and other properties that influence recommendation quality, independent of the provider's demographic attributes. The sensitive latent subspace $\mathbf{z}_i^s \in \mathbb{R}^{d_s}$ isolates provider demographic information, specifically provider gender. By separating these representations, we ensure that item exposure in recommendations is determined by content quality and relevance rather than by whether the provider is male or female, thereby mitigating provider-side gender bias.

This decomposition is formalized through a factorized variational posterior mirroring the user encoder:

$$q_{\phi_i}(\mathbf{z}_i^t, \mathbf{z}_i^s \mid \mathbf{e}_i) = q_{\phi_i^t}(\mathbf{z}_i^t \mid \mathbf{e}_i) \cdot q_{\phi_i^s}(\mathbf{z}_i^s \mid \mathbf{e}_i) \tag{3.7}$$

where each component follows a Gaussian distribution:

$$\mathbf{z}_i^t \sim \mathcal{N}(\boldsymbol{\mu}_i^t, \mathrm{diag}(\boldsymbol{\sigma}_i^t)^2) \tag{3.8}$$

$$\mathbf{z}_i^s \sim \mathcal{N}(\boldsymbol{\mu}_i^s, \mathrm{diag}(\boldsymbol{\sigma}_i^s)^2) \tag{3.9}$$

The task latent representation $\mathbf{z}_i^t$ encodes content properties relevant for matching user preferences (e.g., musical genre, acoustic features, artist style), while the sensitive representation $\mathbf{z}_i^s$ isolates provider gender information. The symmetry between user and item encoders ensures that fairness interventions are applied consistently across both sides of the recommendation interaction.

3.4.3 Task-Specific Recommendation Decoder

The recommendation network constitutes the critical fairness mechanism in our architecture for achieving dual-sided gender fairness. By design, it operates exclusively on the task latent subspaces, ensuring that predictions are constructed without access to gender information from either users or content providers. This architectural constraint simultaneously addresses both user-side bias (ensuring equal recommendation quality across male and female users) and provider-side bias (ensuring fair exposure for male and female artists). For a user-item pair (u, i), the predicted interaction score is computed as:

$$\hat{r}_{ui} = f_\theta(\mathbf{z}_u^t \oplus \mathbf{z}_i^t) \tag{3.10}$$

where $\oplus$ denotes concatenation and f_θ is a multi-layer perceptron with sigmoid activation:

$$f_\theta(\mathbf{z}) = \sigma(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) \tag{3.11}$$

with learnable parameters $\theta = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$. The exclusion of sensitive latent variables $\mathbf{z}_u^s$ and $\mathbf{z}_i^s$ from Equation 3.10 ensures that gender attributes from both users and content providers cannot influence recommendation scores, achieving dual-sided fairness through architectural constraint rather than regularization. This guarantees that users of different genders receive comparable recommendation quality, and that artists of different genders receive exposure proportional to their relevance to user preferences rather than their gender.

3.4.4 Auxiliary Sensitive Attribute Decoders

To ensure effective disentanglement, we employ auxiliary classifiers that predict gender attributes from the sensitive subspaces. These decoders serve two purposes: verifying that gender information is successfully captured in the sensitive subspaces, and providing supervision signals that guide the separation between task and sensitive information. For users, we define:

$$\hat{s}_u = g_{\psi_u}(\mathbf{z}_u^s) \tag{3.12}$$

where g_{ψ_u} is a classification network that predicts user gender from the sensitive latent representation. Similarly, for items:

$$\hat{g}_i = g_{\psi_i}(\mathbf{z}_i^s) \tag{3.13}$$

predicts artist gender. Both networks are implemented as multi-layer perceptrons with sigmoid activations for binary gender classification.

3.4.5 Training Objective

The complete training objective integrates multiple loss components that jointly optimize for recommendation accuracy, representation quality, and effective disentanglement. The objective comprises three main terms: recommendation loss, variational regularization, and sensitive attribute prediction loss.

The recommendation loss ensures accurate prediction of user-item interactions using mean squared error:

$$\mathcal{L}_{\text{rec}} = \sum_{(u,i)\in\mathcal{D}} (r_{ui} - \hat{r}_{ui})^2 \tag{3.14}$$

where $\mathcal{D}$ denotes the observed interaction dataset and r_{ui} represents the normalized interaction strength between user u and item i.

The variational regularization enforces proper probabilistic structure in the latent spaces through KL divergence terms. For users, the KL divergence includes both task and sensitive subspaces:

$$\mathcal{L}^u_{\text{KL}} = \text{KL}(q_{\phi_u}(\mathbf{z}^t_u)\|p(\mathbf{z}^t_u)) + \text{KL}(q_{\phi_u}(\mathbf{z}^s_u)\|p(\mathbf{z}^s_u)) \tag{3.15}$$

For items, the KL divergence is computed analogously:

$$\mathcal{L}^i_{\text{KL}} = \text{KL}(q_{\phi_i}(\mathbf{z}^t_i)\|p(\mathbf{z}^t_i)) + \text{KL}(q_{\phi_i}(\mathbf{z}^s_i)\|p(\mathbf{z}^s_i)) \tag{3.16}$$

where $p(\mathbf{z})$ is the standard Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The KL divergence has closed form:

$$\text{KL}(q_\phi(\mathbf{z})\|\mathcal{N}(\mathbf{0}, \mathbf{I})) = -\frac{1}{2}\sum_{k=1}^{d}[1 + \log((\sigma_k)^2) - (\mu_k)^2 - (\sigma_k)^2] \tag{3.17}$$

The total KL divergence loss combines both user and item components:

$$\mathcal{L}_{\text{KL}} = \mathcal{L}^u_{\text{KL}} + \mathcal{L}^i_{\text{KL}} \tag{3.18}$$

The sensitive attribute prediction loss provides supervision for the disentanglement process by ensuring that gender information is captured in the sensitive subspaces. For users, we define the user-specific sensitive attribute loss using binary cross-entropy:

$$\mathcal{L}^u_{\text{attr}} = \sum_{u\in\mathcal{U}} \text{BCE}(s_u, \hat{s}_u) \tag{3.19}$$

For items, the item-specific sensitive attribute loss is:

$$\mathcal{L}_{\text{attr}}^{i} = \sum_{i \in \mathcal{I}} \text{BCE}(g_i, \hat{g}_i) \tag{3.20}$$

where s_u and g_i denote the true gender labels for users and item providers respectively, and $\text{BCE}(a, \hat{a}) = -[a \log \hat{a} + (1 - a) \log(1 - \hat{a})]$. The total sensitive attribute loss combines both components:

$$\mathcal{L}_{\text{attr}} = \mathcal{L}_{\text{attr}}^{u} + \mathcal{L}_{\text{attr}}^{i} \tag{3.21}$$

To further enforce separation between task and sensitive subspaces, we introduce an orthogonality constraint that penalizes correlation between the two representations. For users, the orthogonality loss is:

$$\mathcal{L}_{\text{ortho}}^{u} = \text{Ortho}(\boldsymbol{\mu}_u^t, \boldsymbol{\mu}_u^s) \tag{3.22}$$

For items, the orthogonality loss is:

$$\mathcal{L}_{\text{ortho}}^{i} = \text{Ortho}(\boldsymbol{\mu}_i^t, \boldsymbol{\mu}_i^s) \tag{3.23}$$

where the orthogonality function for two vectors is computed as:

$$\text{Ortho}(\mathbf{v}_1, \mathbf{v}_2) = \left(\frac{\mathbf{v}_1^\top \mathbf{v}_2}{\|\mathbf{v}_1\|_2 \|\mathbf{v}_2\|_2} \right)^2 \tag{3.24}$$

The total orthogonality loss is:

$$\mathcal{L}_{\text{ortho}} = \mathcal{L}_{\text{ortho}}^{u} + \mathcal{L}_{\text{ortho}}^{i} \tag{3.25}$$

This loss encourages the cosine similarity between task and sensitive representations to approach zero, promoting independence between the subspaces. When the representations are orthogonal, their dot product is zero, minimizing the loss and ensuring that task-relevant features contain minimal gender information.

The complete training objective combines these components with balancing hyperparameters:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}} + \alpha \mathcal{L}_{\text{attr}} + \gamma \mathcal{L}_{\text{ortho}} \tag{3.26}$$

where the hyperparameter β controls the strength of the variational regularization, α weights the attribute prediction objective, and γ controls the orthogonality constraint strength. This multi-objective formulation ensures that the model learns representations that are simultaneously predictive, probabilistically well-behaved, effectively disentangled with respect to gender information, and structurally orthogonal between task and sensitive subspaces.

3.4.6 Inference and Fair Recommendation Generation

During inference, fair recommendations are generated through a streamlined process that excludes sensitive information. For a given user u and candidate item i, we first extract the mean vectors from the task subspaces: $\boldsymbol{\mu}_u^t$ from the user encoder and $\boldsymbol{\mu}_i^t$ from the item encoder. The predicted interaction score is then computed using only these task representations:

$$\hat{r}_{ui} = f_\theta(\boldsymbol{\mu}_u^t \oplus \boldsymbol{\mu}_i^t) \tag{3.27}$$

Note that we use the mean vectors rather than sampling from the posterior distributions, following standard practice in variational inference for prediction tasks. Importantly, the sensitive latent variables $\mathbf{z}_u^s$ and $\mathbf{z}_i^s$ are never computed or accessed during inference, ensuring that gender information cannot influence recommendation scores. For a target user, items are ranked by their predicted scores, and the top-K items with highest $\hat{r}_{ui}$ values constitute the final recommendation list. This inference procedure guarantees that recommendations are generated based solely on preference-relevant features, achieving fairness through architectural design.

3.5 Experimental Setup

3.5.1 Dataset Description

We used a dataset provided by Ekstrand et al. [122], which is based on the Last.fm 360K dataset [123], and enhanced with artist gender information (male/female) extracted from MusicBrainz.org, as applied by Ekstrand et al. [122]. The dataset includes 46,469 artists, with 10,535 women and 35,922 men [123]. Following Ekstrand et al. [122] methodology, we only considered users and artists with more than 30 interactions to ensure enough data for training and evaluation. This left us with 220,444 users and 12,900 artists, about a third of whom are women.

In addition to the Last.fm-360K music dataset, we evaluate DS-DVAE on the MovieLens-1M (ML-1M) dataset, one of the most widely used benchmarks in recommender systems research, in order to validate the generalizability of the framework beyond the music domain.

The MovieLens-1M dataset was released by GroupLens Research and contains 1,000,209 explicit ratings provided by 6,040 users on 3,952 movies. Ratings are integer values on a five-point scale (1-5). The dataset includes demographic information for each user, specifically gender (male/female), age group, and occupation, making it particularly well suited for fairness-aware evaluation. Among the 6,040 users, approximately 71% are male and 29% are female, introducing a demographic imbalance that is representative of real-world platforms and that allows direct measurement of user-side gender bias.

On the provider side, the original ML-1M release does not contain any information about content creators. We therefore enriched the dataset with director gender, which serves as the provider

Algorithm 1 DS-DVAE Training and Fair Inference.

Require: Interaction data $\mathcal{D} = \{(u, i, r_{ui})\}$, Gender labels $\{s_u, g_i\}$
Require: Hyperparameters: $\alpha_u, \alpha_i, \beta_u, \beta_i, \gamma, \eta$ (learning rate)
Ensure: Optimized parameters $\Theta = \{\phi_u, \phi_i, \theta, \psi_u, \psi_i\}$

Training Phase:
while not converged **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    // Encode to disentangled subspaces
    $\boldsymbol{\mu}_u^t, \boldsymbol{\sigma}_u^t \leftarrow \text{MLP}_{\phi_u}^t(\mathbf{e}_u)$; $\boldsymbol{\mu}_u^s, \boldsymbol{\sigma}_u^s \leftarrow \text{MLP}_{\phi_u}^s(\mathbf{e}_u)$
    $\boldsymbol{\mu}_i^t, \boldsymbol{\sigma}_i^t \leftarrow \text{MLP}_{\phi_i}^t(\mathbf{e}_i)$; $\boldsymbol{\mu}_i^s, \boldsymbol{\sigma}_i^s \leftarrow \text{MLP}_{\phi_i}^s(\mathbf{e}_i)$
    // Sample latent representations via reparameterization
    $\mathbf{z}_u^t \leftarrow \boldsymbol{\mu}_u^t + \boldsymbol{\sigma}_u^t \odot \boldsymbol{\epsilon}$; $\mathbf{z}_u^s \leftarrow \boldsymbol{\mu}_u^s + \boldsymbol{\sigma}_u^s \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{z}_i^t \leftarrow \boldsymbol{\mu}_i^t + \boldsymbol{\sigma}_i^t \odot \boldsymbol{\epsilon}$; $\mathbf{z}_i^s \leftarrow \boldsymbol{\mu}_i^s + \boldsymbol{\sigma}_i^s \odot \boldsymbol{\epsilon}$
    // Forward pass through decoders
    $\hat{r}_{ui} \leftarrow f_\theta(\mathbf{z}_u^t \oplus \mathbf{z}_i^t)$ ▷ Task decoder
    $\hat{s}_u \leftarrow g_{\psi_u}(\mathbf{z}_u^s)$; $\hat{g}_i \leftarrow g_{\psi_i}(\mathbf{z}_i^s)$ ▷ Sensitive decoders
    // Compute loss components
    $\mathcal{L}_{\text{rec}} \leftarrow \sum (r_{ui} - \hat{r}_{ui})^2$
    $\mathcal{L}_{\text{KL}}^u \leftarrow \text{KL}(q_{\phi_u}(\mathbf{z}_u^t)) + \text{KL}(q_{\phi_u}(\mathbf{z}_u^s))$
    $\mathcal{L}_{\text{KL}}^i \leftarrow \text{KL}(q_{\phi_i}(\mathbf{z}_i^t)) + \text{KL}(q_{\phi_i}(\mathbf{z}_i^s))$
    $\mathcal{L}_{\text{attr}}^u \leftarrow \text{BCE}(s_u, \hat{s}_u)$; $\mathcal{L}_{\text{attr}}^i \leftarrow \text{BCE}(g_i, \hat{g}_i)$
    // Orthogonality constraint
    $\mathcal{L}_{\text{ortho}} \leftarrow \text{Ortho}(\boldsymbol{\mu}_u^t, \boldsymbol{\mu}_u^s) + \text{Ortho}(\boldsymbol{\mu}_i^t, \boldsymbol{\mu}_i^s)$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{rec}} + \beta_u \mathcal{L}_{\text{KL}}^u + \beta_i \mathcal{L}_{\text{KL}}^i + \alpha_u \mathcal{L}_{\text{attr}}^u + \alpha_i \mathcal{L}_{\text{attr}}^i + \gamma \mathcal{L}_{\text{ortho}}$
    // Backpropagation and update
    $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{\text{total}}$
  end for
end while
Inference Phase:
Generate Recommendations(u, K)
$\boldsymbol{\mu}_u^t \leftarrow \text{MLP}_{\phi_u}^t(\mathbf{e}_u)$ ▷ Extract user task representation
for each candidate item $i \in \mathcal{I}$ **do**
  $\boldsymbol{\mu}_i^t \leftarrow \text{MLP}_{\phi_i}^t(\mathbf{e}_i)$ ▷ Extract item task representation
  $\hat{r}_{ui} \leftarrow f_\theta(\boldsymbol{\mu}_u^t \oplus \boldsymbol{\mu}_i^t)$ ▷ Sensitive subspaces not used
end for
return Top-K items ranked by $\hat{r}_{ui}$

attribute g_i, using the official MovieLens website. Each movie identifier in ML-1M was matched to its corresponding entry on the MovieLens website, from which the director's name was retrieved. The gender of each director (male/female) was then annotated based on the publicly available biographical information linked from the same movie page. Movies for which the director could not be reliably identified were excluded from the provider-side fairness evaluation. This procedure mirrors the artist gender enrichment applied to Last.fm-360K, ensuring that provider fairness is

measured consistently across both domains.

Implementation details. DS-DVAE is implemented in PyTorch and trained using the Adam optimizer with a learning rate of $\eta = 0.001$. As summarized in Table 3.1, the embedding dimension is set to $d = 64$, which is factorized into a task subspace of dimension $d_t = 32$ and a sensitive subspace of dimension $d_s = 32$. The attribute decoder loss weights are set to $\alpha_u = \alpha_i = 1.0$ for user and item sensitive attribute supervision respectively. The KL divergence regularization weights are set to $\beta_u = \beta_i = 0.01$ for both user and item subspaces. The model is trained for 40 epochs with a mini-batch size of 4,096. For dataset preprocessing, a k-core filter with $k = 30$ minimum interactions is applied iteratively to both users and items until convergence. The data is split per user into 80% for training, and 20% for testing.

Table 3.1: DS-DVAE hyperparameter configuration.

Hyperparameter	**Symbol**	**Value**
Embedding dimension	d	64
Task subspace dimension	d_t	32
Sensitive subspace dim.	d_s	32
User attr. loss weight	α_u	1.0
Item attr. loss weight	α_i	1.0
User KL weight	β_u	0.01
Item KL weight	β_i	0.01
Orthogonality weight	γ	0.1
Learning rate	η	0.001
Batch size	B	4,096
Training epochs	-	40
k-core filter threshold	k	30
Train / test split	-	80% / 20%

3.5.2 VAE Backbone Models

Our evaluation is organized in two stages. In the first stage (Table 3.2), DS-DVAE is compared against the *original vanilla VAE* recommender, i.e., a standard variational autoencoder trained on the interaction matrix without any disentanglement or fairness mechanism, which serves as the reference baseline. In the second stage (Table 3.3), we assess the generality of the approach by plugging DS-DVAE into ten enhanced VAE-based recommendation models from the literature, briefly described below.

- **MultVAE** [124] extends variational autoencoders to collaborative filtering with implicit feedback by modeling each user's interaction vector with a multinomial likelihood and training

with a partially regularized (β-annealed) ELBO.

- **RecVAE** [125] improves upon MultVAE through a redesigned encoder architecture, a composite prior that mixes the standard Gaussian with the previous variational posterior, user-dependent rescaling of the KL term, and alternating training, yielding more stable and accurate models.
- **MacridVAE** [126] performs macro-micro disentangled representation learning: it separates the high-level user intentions associated with different concepts (macro level) and further encourages statistical independence between individual latent dimensions (micro level) using prototype-based latent clustering.
- **SVAE** [127] is a sequential VAE that combines a recurrent (GRU-based) encoder with variational inference, modeling the temporal order of user interactions and predicting the distribution of future items.
- **BiVAE** [128] is a bilateral VAE designed for dyadic data: it introduces two coupled inference networks that learn user and item latent variables symmetrically, making it naturally suited to two-sided user-item modeling.
- **CVGA** [129] is a collaborative variational graph autoencoder that encodes the user-item bipartite graph with graph convolutional layers, aggregating neighborhood information into the variational latent representations.
- **ACVAE** [130] augments the sequential VAE with adversarial variational Bayes training and a contrastive learning objective, producing more personalized and discriminative latent representations.
- **M2VAE** [131] is a multi-modal multi-view VAE that learns type-specific latent variables for item IDs, categorical attributes, and content features, and combines them through a Product-of-Experts mechanism into a common representation, with a contrastive objective separating shared from modality-specific views.
- **HCVAE** [132] is a hybrid conditional VAE that integrates both user side information (e.g., demographics) and item side information (e.g., categories) as conditions within the variational framework, constraining the learned latent factors to produce personalized top-n recommendations.
- **CVAE** [133] is a collaborative variational autoencoder that jointly models user-item interactions together with auxiliary item content information within a unified Bayesian generative framework.

These backbones were selected to cover a broad architectural spectrum - flat (MultVAE, RecVAE), disentangled (MacridVAE), sequential (SVAE, ACVAE), dyadic (BiVAE), graph-based (CVGA), conditional side-information-based (HCVAE, CVAE), and multi-modal (M2VAE) - so that the backbone-agnostic behavior of DS-DVAE can be assessed across structurally diverse latent spaces.

3.5.3 Evaluation Metrics

We evaluate our DS-DVAE framework across three critical dimensions: recommendation accuracy, user-side fairness, and provider-side fairness.

For recommendation accuracy. We evaluated the top-N recommendation performance with normalized discounted cumulative gain (NDCG@N), precision@k, and recall@k where N is the number of recommended items, we set N = 10.

To evaluate fairness across user demographic groups, we employ User Group Fairness (UGF) metrics that measure disparities in recommendation quality between male and female users. For each accuracy metric M (Precision, Recall, or NDCG), we compute the absolute difference between the average metric values for the two gender groups:

$$\text{UGF}_M = |M^{male} - M^{female}| \tag{3.28}$$

where $M^{male} = \frac{1}{|\mathcal{U}^{male}|} \sum_{u \in \mathcal{U}^{male}} M_u$ and M^{female} is computed analogously. Lower UGF values indicate more equitable recommendation quality across gender groups. We report three UGF metrics: UGF-Precision@10, UGF-Recall@10, and UGF-NDCG@10.

For provider-side fairness, we adopt the metrics:

Ranking-based Statistical Parity (RSP@K): Measures overall exposure equality:

$$\text{RSP@}K = \frac{\text{std}(p_0(K), p_1(K))}{\text{mean}(p_0(K), p_1(K))} \tag{3.29}$$

where $p_j(K)$ is the probability of items in group j appearing in top-K recommendations.

Ranking-based Equal Opportunity (REO@K): measures relevant exposure equality:

$$\text{REO@}K = \frac{\text{std}(p_0(K \mid y = 1), p_1(K \mid y = 1))}{\text{mean}(p_0(K \mid y = 1), p_1(K \mid y = 1))} \tag{3.30}$$

Finally, to quantify privacy leakage, we measure the **attribute inference AUC**. An external MLP classifier is trained to predict the gender attribute from the learned user (respectively item) representations, and we report the area under the ROC curve, denoted AUC-User and AUC-Provider. An AUC close to 0.5 indicates that the embeddings are uninformative with respect to gender (the attack performs no better than random guessing), whereas values approaching 1.0 indicate severe demographic leakage.

3.6 Results

3.6.1 Comparative Performance Evaluation

Table 3.2 provides a comprehensive comparison between the original vanilla VAE model - a standard variational autoencoder recommender trained without any disentanglement or fairness mechanism - and the proposed DS-DVAE framework across Accuracy, User Fairness, Item Fairness, and Privacy on the LFM-360K and ML-1M datasets. Note that this vanilla baseline is distinct from the enhanced VAE backbones evaluated later in Table 3.3; the absolute values in the two tables are therefore not directly comparable, as they correspond to different base architectures.

Table 3.2: Comprehensive Performance Comparison: Baseline VAE vs. DS-DVAE at K=10.

		LFM-360K		ML-1M	
Category	Metric	Baseline	DS-DVAE	Baseline	DS-DVAE
Accuracy	Precision@10 ↑	0.134	0.133	0.369	0.394
	Recall@10 ↑	0.177	0.176	0.180	0.192
	NDCG@10 ↑	0.361	0.314	0.364	0.365
User Fairness	UGF-Precision ↓	0.030	0.006	0.025	0.001
	UGF-Recall ↓	0.038	0.008	0.011	0.003
	UGF-NDCG ↓	0.045	0.015	0.011	0.009
Item Fairness	RSP ↓	0.365	0.136	0.7395	0.6455
	REO ↓	0.176	0.044	0.3639	0.1578
Privacy	AUC-User ↓	0.905	0.502	0.843	0.526
	AUC-provider ↓	0.852	0.516	0.849	0.525

In terms of Accuracy, DS-DVAE exhibits distinct behaviors depending on the dataset. For LFM-360K, the model maintains stable Precision@10 (dropping marginally from 0.134 to 0.133) and Recall@10 scores (from 0.177 to 0.176) relative to the baseline, with a trade-off in ranking quality as NDCG@10 drops from 0.361 to 0.314. Conversely, on ML-1M, DS-DVAE yields an accuracy improvement on the precision-oriented metrics: Precision@10 increases from 0.369 to 0.394 and Recall@10 from 0.180 to 0.192, while NDCG@10 is fully preserved (0.364 vs. 0.365).

This divergent performance across datasets can be attributed to the underlying structural differences in user interaction matrices and catalog distributions. In the LFM-360K dataset, interactions are characterized by extreme data sparsity, a vast long-tail item catalog, and niche listening habits. Because the baseline model relies on subtle collaborative filtering associations across this massive artist space, isolating demographic attributes shifts latent coordinates enough to cause a slight

ranking trade-off (the fairness cost). Part of this drop is also a direct consequence of the hyperparameter configuration: the fixed 50/50 split of the embedding into task and sensitive subspaces ($d_t = d_s = 32$) halves the representational capacity available to the recommendation decoder, while the attribute supervision weight ($\alpha = 1.0$) and the orthogonality penalty inject competing gradients into the encoder. Since these hyperparameters were kept identical across both datasets rather than tuned for the sparse Last.fm-360K setting, a dataset-specific search over α, β, γ, and the subspace dimensions would be expected to recover much of the NDCG loss. Conversely, the ML-1M dataset exhibits a much denser interaction matrix that is heavily confounded by popular, gender-skewed movie genres. In this scenario, demographic dependencies introduce structural noise where gender-stereotypical choices obscure actual preference trends. By forcing the latent embeddings to remain strictly orthogonal to sensitive features, DS-DVAE functions as an effective regularizer. It strips away these misleading demographic correlations and purifies the task subspace to isolate genuine interaction signals, translating directly into an accuracy boost on ML-1M.

For User Fairness, the framework effectively minimizes demographic performance gaps. On LFM-360K, the UGF-Precision gap drops from 0.030 to 0.006, while UGF-Recall falls from 0.038 to 0.008 and UGF-NDCG decreases from 0.045 to 0.015. On ML-1M, the baseline disparities are heavily compressed, leading to a UGF-Precision gap reduction from 0.025 to 0.001, a UGF-Recall drop from 0.011 to 0.003, and a UGF-NDCG gap reduction from 0.011 to 0.009, ensuring equitable service quality for all users.

Regarding Item Fairness, DS-DVAE mitigates systematic provider exposure bias. On LFM-360K, RSP falls from 0.365 to 0.136, and REO drops from 0.176 to 0.044. On the heavily skewed ML-1M dataset, which presents a high baseline imbalance (RSP of 0.7395 and REO of 0.3639), the framework reduces RSP to 0.6455 and more than halves the REO to 0.1578, confirming fairer visibility for minority-group providers based on item relevance.

Finally, Privacy metrics validate the representation disentanglement. While baseline embeddings leak significant gender information with AUC-User scores between 0.843 and 0.905 and AUC-provider scores between 0.849 and 0.852, the inference attack success rate under DS-DVAE collapses to near-random guessing, dropping to 0.502 for users and 0.516 for artists on LFM-360K, and 0.526 for users and 0.525 for artists on ML-1M.

3.6.2 Backbone-Agnostic Evaluation

Having established the behavior of DS-DVAE against the original vanilla VAE baseline, we next assess the generality of the framework when it is applied as a plug-in module on top of the enhanced, state-of-the-art VAE recommenders described in Section 3.5.2.

Table 3.3 reports the performance of ten VAE-based recommendation backbones with and without the proposed DS-DVAE disentanglement, on both Last.fm-360K and MovieLens-1M. The most striking pattern is the consistency of the fairness gains: for every backbone and on both datasets, the

user-side UGF metrics and the provider-side RSP and REO scores improve when DS-DVAE is applied. The provider-side improvements are particularly pronounced, with RSP and REO frequently reduced by an order of magnitude (e.g., MultVAE on Last.fm-360K drops from RSP = 0.4180 to 0.0435 and from REO = 0.3056 to 0.0206, while ACVAE on ML-1M reduces RSP from 0.4762 to 0.0016). This confirms that exposure inequality is largely driven by demographic information embedded in the item representations themselves; once provider gender is confined to the sensitive subspace and excluded from the scoring pathway, exposure realigns with relevance. User-side disparities, by contrast, are roughly halved on average rather than eliminated, since part of the residual quality gap stems from group differences in interaction density and catalog coverage that no representation-level intervention can fully remove. On the accuracy side, the cost of disentanglement is generally modest: for standard backbones such as MultVAE, RecVAE, BiVAE, and CVAE on Last.fm-360K, Precision@10 and Recall@10 decrease by only 1-3% relative to the unconstrained model, indicating that gender carries little unique predictive signal once genuine preference patterns are available to the task subspace. An important part of this residual accuracy decrease is attributable to the hyperparameter configuration rather than to the disentanglement principle itself. First, a single shared configuration (Table 3.1) is applied uniformly to all ten backbones without per-model tuning, so architectures whose optimal KL weight, learning rate, or latent dimensionality differ from this common setting are penalized. Second, the factorization of the $d = 64$ embedding into two equal halves restricts the effective task capacity to $d_t = 32$, meaning the recommendation decoder operates on half of the original representational budget; a sensitive subspace of $d_s = 32$ is also larger than what a binary attribute strictly requires, and reallocating capacity toward the task subspace would likely recover part of the loss. Third, the auxiliary attribute loss ($\alpha = 1.0$) and the orthogonality constraint introduce competing gradients that pull the encoder away from the purely accuracy-optimal solution, especially during early training. The largest degradations occur precisely where these factors interact with structured latent spaces: MacridVAE (Precision@10 from 0.0977 to 0.0782 on Last.fm-360K) already imposes its own macro-level disentanglement, CVGA (NDCG from 0.6353 to 0.4947 on ML-1M) relies on graph-based aggregation, and the dyadic BiVAE couples user and item posteriors tightly; superimposing an additional orthogonal task-sensitive factorization with untuned weights perturbs the latent geometries these architectures depend on. This suggests that a per-backbone hyperparameter search over α, β, γ, and the subspace split (d_t, d_s) would narrow the accuracy gap further without sacrificing fairness.

Interestingly, DS-DVAE does not uniformly trade accuracy for fairness; for several backbones it improves both simultaneously. On Last.fm-360K, M2VAE and HCVAE achieve higher Precision, Recall, and NDCG with DS-DVAE than without, and on ML-1M the same holds for RecVAE, ACVAE, and CVAE, with SVAE also gaining in Recall and NDCG. This supports a regularization interpretation: in datasets where consumption is confounded by gender-skewed popularity (most visibly the genre-stratified ML-1M), demographic correlations act as structural noise, and forcing

Table 3.3: Experimental results of different VAE backbones with (w) and without (w/o) the proposed DS-DVAE disentanglement applied.

Dataset	**Model**		**Accuracy**			**User Fairness**			**Provider Fairness**	
			Prec@10	Rec@10	NDCG	UGF_Prec	UGF_Rec	UGF_NDCG	RSP	REO
LastFM360K	MultVAE	w/o	**0.1049**	**0.2294**	**0.3998**	0.0417	0.0611	0.1038	0.4180	0.3056
		w	0.1026	0.2249	0.3906	**0.0195**	**0.0278**	**0.0584**	**0.0435**	**0.0206**
	RecVAE	w/o	**0.1007**	**0.2222**	**0.3845**	0.0418	0.0614	0.0694	0.3556	0.3206
		w	0.0993	0.2185	0.3836	**0.0197**	**0.0285**	**0.0315**	**0.0442**	**0.0109**
	MacridVAE	w/o	**0.0977**	**0.2143**	**0.3778**	0.0337	0.0497	0.0996	0.5077	0.3653
		w	0.0782	0.1701	0.3173	**0.0185**	**0.0268**	**0.0530**	**0.2734**	**0.0821**
	SVAE	w/o	**0.0835**	**0.1801**	**0.3203**	0.0290	0.0364	0.0038	0.4490	0.3691
		w	0.0792	0.1700	0.3034	**0.0244**	**0.0302**	**0.0001**	**0.1989**	**0.0129**
	BiVAE	w/o	**0.0992**	**0.2177**	**0.3783**	0.0289	0.0419	0.0811	0.3361	0.3479
		w	0.0969	0.2130	0.3691	**0.0170**	**0.0242**	**0.0471**	**0.1740**	**0.0384**
	CVGA	w/o	**0.0648**	**0.1431**	**0.2663**	0.0255	0.0377	0.0629	0.4176	0.3780
		w	0.0565	0.1246	0.2339	**0.0094**	**0.0120**	**0.0308**	**0.1483**	**0.044**
	ACVAE	w/o	**0.0832**	**0.1791**	**0.3186**	0.0220	0.0315	0.0604	0.3708	0.4160
		w	0.0781	0.1676	0.3008	**0.0216**	**0.0309**	**0.0594**	**0.0272**	**0.0998**
	M2VAE	w/o	0.0986	0.2147	0.3800	0.0134	0.0165	0.0117	0.5413	0.4758
		w	**0.1018**	**0.2229**	**0.3890**	**0.0012**	**0.0031**	**0.0077**	**0.073**	**0.0445**
	HCVAE	w/o	0.0996	0.2182	0.3816	0.0265	0.0381	0.0805	0.3610	0.3413
		w	**0.1010**	**0.2207**	**0.3884**	**0.0158**	**0.0225**	**0.0365**	**0.1043**	**0.079**
	CVAE	w/o	**0.1047**	**0.2286**	**0.3984**	0.0320	0.0468	0.0914	0.5307	0.3377
		w	0.1020	0.2237	0.3904	**0.0092**	**0.0120**	**0.0238**	**0.0982**	**0.0660**
ML-1M	MultVAE	w/o	**0.1286**	0.1917	**0.3377**	0.0380	0.0075	0.0692	0.4509	0.2868
		w	0.1231	**0.1935**	0.3084	**0.0247**	**0.0022**	**0.0463**	**0.0238**	**0.0117**
	RecVAE	w/o	0.1307	0.1886	**0.3358**	0.0383	0.0071	0.0485	0.4943	0.2772
		w	**0.1329**	**0.1902**	0.3320	**0.0309**	**0.0022**	**0.0306**	**0.0064**	**0.1157**
	MacridVAE	w/o	**0.1262**	0.1877	**0.3296**	0.0403	0.0057	0.0583	0.4684	0.2525
		w	0.1159	**0.1912**	0.3020	**0.0252**	**0.0023**	**0.0495**	**0.0399**	**0.0503**
	SVAE	w/o	**0.1169**	0.1820	0.2989	0.0385	0.0065	0.0746	0.4479	0.2278
		w	0.1168	**0.1948**	**0.3014**	**0.0268**	**0.0001**	**0.0738**	**0.0068**	**0.0556**
	BiVAE	w/o	**0.3571**	**0.1700**	**0.6487**	0.0385	0.0058	0.0334	0.4885	0.3026
		w	0.2319	0.0900	0.4991	**0.0248**	**0.0038**	**0.0217**	**0.0363**	**0.1774**
	CVGA	w/o	**0.3428**	**0.1581**	**0.6353**	0.0352	0.0075	0.0281	0.4837	0.3138
		w	0.3251	0.0871	0.4947	**0.0310**	**0.0017**	**0.0275**	**0.0665**	**0.0359**
	ACVAE	w/o	0.1174	0.1843	0.2996	0.0396	0.0054	0.0786	0.4762	0.2404
		w	**0.1175**	**0.2010**	**0.3013**	**0.0224**	**0.0047**	**0.0727**	**0.0016**	**0.0129**
	M2VAE	w/o	**0.1222**	**0.1930**	**0.3127**	0.0388	0.0066	0.0683	0.4745	0.2336
		w	0.1187	0.1896	0.3054	**0.0250**	**0.0047**	**0.0581**	**0.1506**	**0.2317**
	HCVAE	w/o	**0.1297**	0.1933	**0.3392**	0.0398	0.0056	0.0624	0.4850	0.2455
		w	0.1211	**0.1970**	0.3067	**0.0242**	**0.0031**	**0.0613**	**0.0957**	**0.0522**
	CVAE	w/o	0.1297	0.1893	0.3392	0.0396	0.0064	0.0328	0.4844	0.2573
		w	**0.1312**	**0.1996**	**0.3468**	**0.0232**	**0.0061**	**0.0221**	**0.0861**	**0.1445**

the task subspace to be orthogonal to the sensitive subspace strips away these spurious shortcuts, allowing the model to fit the genuine preference signal more cleanly.

Finally, the cross-dataset comparison highlights where residual difficulty remains. Improvements on the sparse, long-tailed Last.fm-360K are achieved at a small but systematic ranking cost, whereas the denser ML-1M more often rewards disentanglement with accuracy gains. At the same time, a few ML-1M cases show only partial provider-fairness correction (e.g., M2VAE's REO decreases marginally from 0.2336 to 0.2317, and RecVAE retains REO = 0.1157), suggesting that when relevant exposure is entangled with strong genre-gender confounds, representation-level debiasing alone narrows but does not fully close the gap. Overall, Table 3.3 demonstrates that DS-DVAE behaves as a backbone-agnostic, plug-in fairness module whose benefits generalize across architecturally diverse VAE families.

3.7 Group Fairness Analysis

Consumer Group Fairness Analysis

We evaluate recommendation quality equity across male and female users using Precision, Recall, and NDCG metrics.

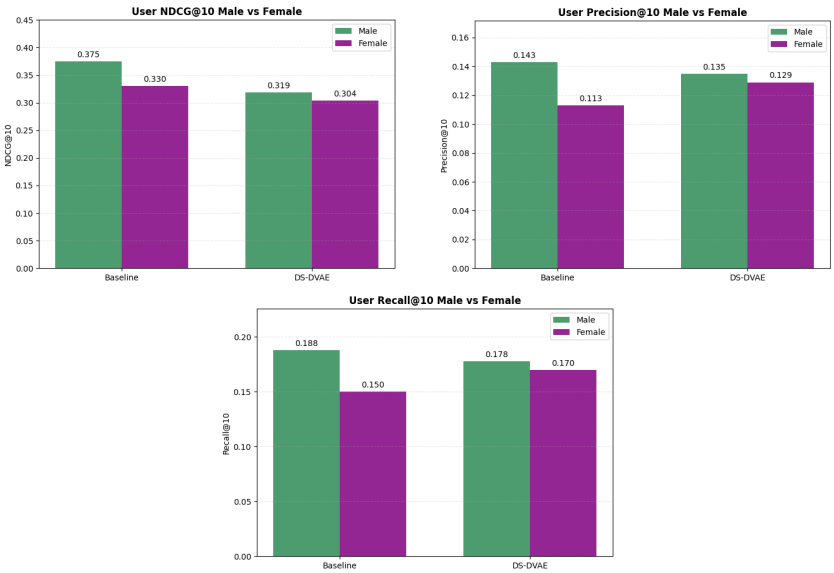
Figure 3.2 demonstrates substantial improvements in user fairness across both datasets. On Last.fm-360K, Precision, Recall, and NDCG gaps are significantly reduced, indicating that male and female users now receive substantially more equitable recommendation quality. Similarly, on MovieLens-1M, all fairness gaps are dramatically reduced, achieving near-parity in recommendation quality across gender groups.

The comparative analysis reveals important dataset-specific patterns. Last.fm-360K shows more pronounced baseline disparities, suggesting stronger gender bias in the baseline model. However, DS-DVAE achieves consistent reduction rates across all metrics, demonstrating robust debiasing across different datasets and interaction domains. MovieLens-1M achieves even tighter convergence across all metrics, indicating that the disentanglement mechanism effectively removes gender-dependent patterns regardless of underlying data characteristics. This consistency validates that DS-DVAE's fairness improvements are not dataset-specific artifacts but fundamental outcomes of the representation disentanglement approach.

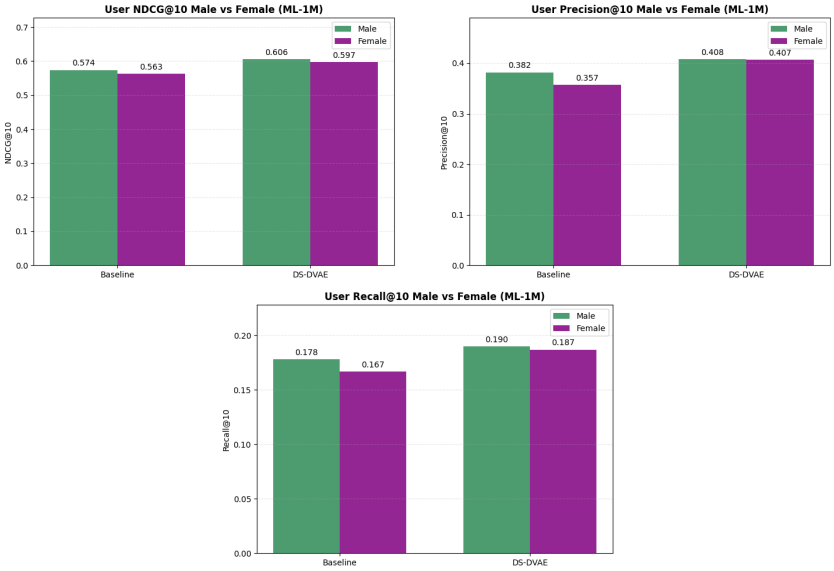
Item Provider Group Fairness

Figure 3.3 demonstrates dramatic improvements in provider fairness. On Last.fm-360K, the exposure disparity is significantly reduced, indicating that male and female artists now receive more balanced visibility in recommendations. On MovieLens-1M, the exposure gap is substantially reduced, representing a considerable increase in the visibility of underrepresented (female-directed) content.

The provider-side results reveal more severe baseline imbalances compared to user-side dispari-



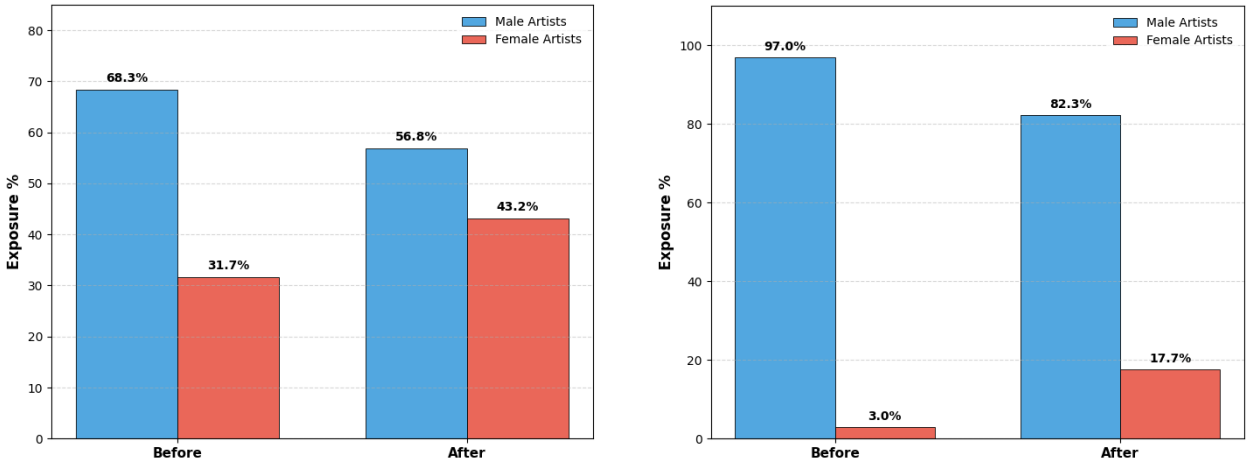
a- Last.fm-360K



b- MovieLens-1M

Figure 3.2: User group fairness before and after DS-DVAE.

ties. MovieLens-1M shows an extreme baseline gender gap in director exposure, reflecting systematic suppression of minority provider representation. DS-DVAE's ability to substantially reduce this gap is particularly significant, as it demonstrates that the disentanglement mechanism can address severe provider bias. The different levels of improvement between datasets suggest that provider bias patterns may be more deeply embedded in some datasets and require domain-specific tuning. Nevertheless, both datasets demonstrate that DS-DVAE successfully mitigates gender-based exposure bias, creating more equitable opportunities for providers to reach audiences based on relevance



Last.fm-360K MovieLens-1M

Figure 3.3: Provider exposure before and after DS-DVAE (artists for Last.fm-360K, directors for MovieLens-1M).

rather than demographic attributes.

3.8 Embedding Visualization

We applied t-SNE dimensionality reduction to visualize the learned embeddings in 2D space, providing intuitive evidence of our framework's effectiveness in creating dual-sided fair representations. Figures 3.4 and 3.5 present the user and item latent spaces before and after applying DS-DVAE on the MovieLens-1M and Last.fm-360K datasets respectively. In each panel, the left column shows the baseline biased representations and the right column shows the debiased task subspace learned by DS-DVAE; the Silhouette Score (SS) quantifies the degree of gender-based clustering, where a higher value indicates stronger demographic separation.

On the MovieLens-1M dataset (Figure 3.4), the baseline user embeddings show male users (blue) and female users (red) forming two clearly separated clusters, with a Silhouette Score of 0.4239 confirming strong gender-based structure in the representation space. This clustering indicates that gender information is heavily encoded in the baseline representations, directly enabling discriminatory treatment of users based on gender. The item space exhibits an even more pronounced separation, with male directors (green) and female directors (purple) forming two elongated, non-overlapping clusters (SS = 0.5010), indicating that the baseline model encodes director gender as a primary organizational axis of the item space, which systematically disadvantages female-directed content in exposure.

After applying DS-DVAE on MovieLens-1M, both subspaces undergo a dramatic transformation. The user task subspace shows male and female users thoroughly intermixed with no visible gender-

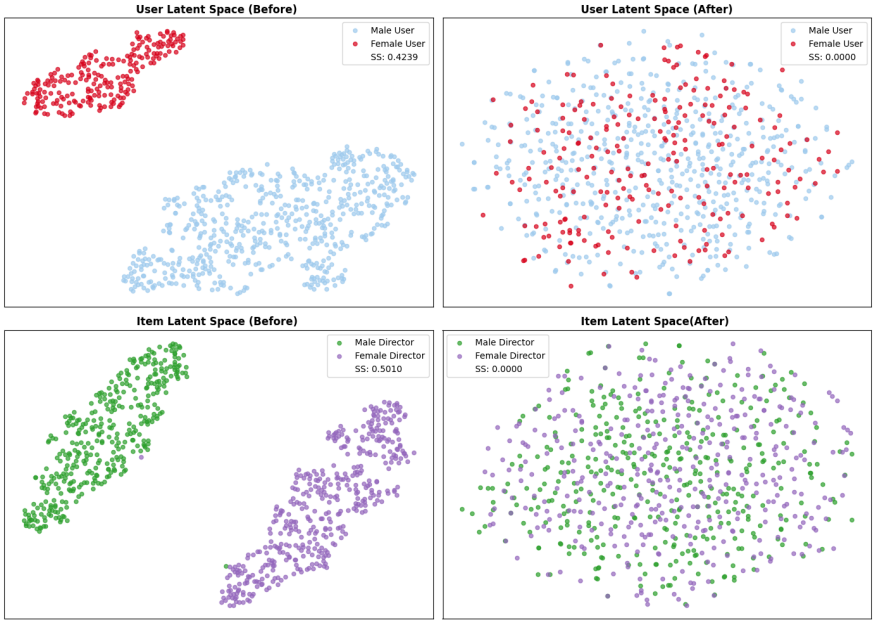


Figure 3.4: Embedding visualization before and after applying DS-DVAE on the MovieLens-1M dataset.

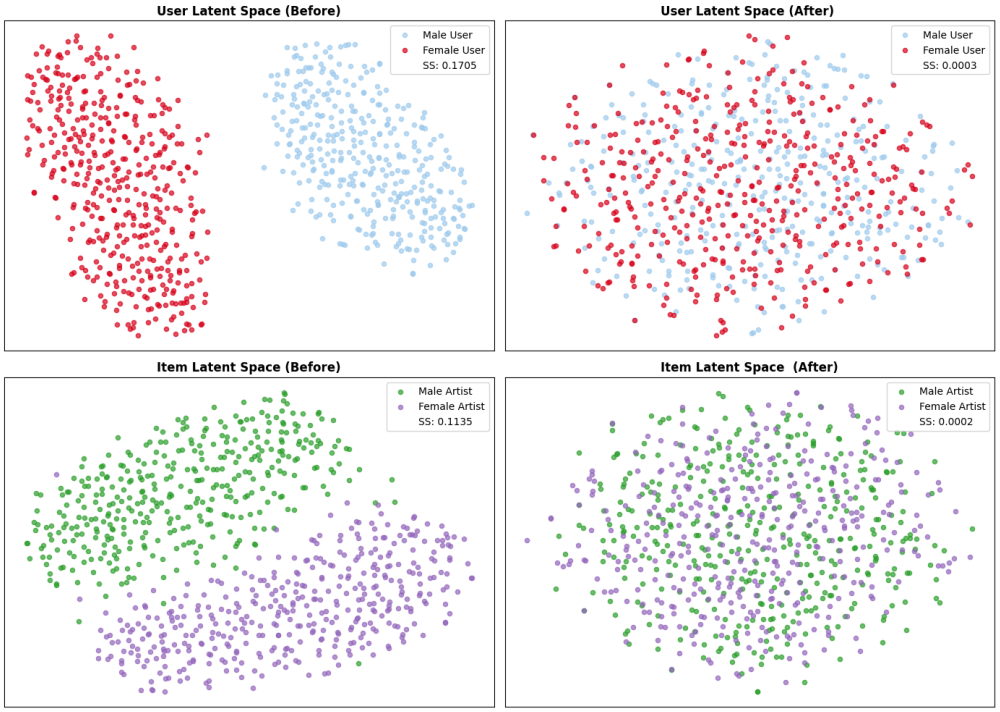


Figure 3.5: Embedding visualization before and after applying DS-DVAE on the Last.fm-360K dataset. Top row: user latent space colored by user gender (blue = male, red = female). Bottom row: item latent space colored by artist gender (green = male, purple = female).

based clustering (SS = 0.0000). The item task subspace similarly shows green and purple points uniformly distributed throughout the space (SS = 0.0000). When the model cannot distinguish users by gender in its task representations, it cannot systematically discriminate in recommendations; when items are organized by content features rather than director demographics, exposure is allocated by relevance rather than gender.

On the Last.fm-360K dataset (Figure 3.5), the baseline representations show a similar pattern at a lower baseline severity. Male and female users form partially separated clusters (SS = 0.1705), and male and female artists occupy distinct regions of the item space (SS = 0.1135). The weaker baseline separation on Last.fm-360K compared to ML-1M reflects the sparser interaction matrix and broader catalog diversity of the music dataset, where demographic correlations are less structurally dominant than in the genre-skewed movie domain.

After DS-DVAE, the Last.fm-360K task subspaces show near-complete removal of demographic structure: the user SS drops to 0.0003 and the item SS to 0.0002. Male and female users are thoroughly intermixed, and male and female artists are uniformly distributed throughout the item space. While local musical structure is preserved — reflecting genuine acoustic and genre similarities — the clear gender-based segregation has vanished, enabling equal recommendation opportunities regardless of artist gender. Together, these results across both datasets confirm that DS-DVAE successfully removes gender-specific patterns from the task subspace while preserving the preference-relevant structure required for accurate recommendation.

3.9 Summary

This work addressed the critical challenge of gender bias in recommendation systems through a novel dual-subspace disentanglement approach. We proposed DS-DVAE, one of the first variational frameworks that simultaneously addresses both user-side and provider-side fairness by explicitly separating representations into preference-relevant and gender-sensitive subspaces. By constructing recommendations solely from the preference subspaces, our approach achieves fairness by architectural design rather than through post-hoc corrections. Our empirical evaluation on the Last.fm-360K and MovieLens-1M datasets demonstrates that DS-DVAE effectively mitigates bias along multiple dimensions while maintaining competitive accuracy. The model achieves substantial reductions in user group fairness disparities, reduces systematic underexposure of female providers (artists and directors), and significantly improves resistance to gender inference attacks. These results validate that the fairness-accuracy trade-off can be favorably balanced through principled representation learning.

Chapter 4

A Unified Two-Sided Fairness Framework for Recommender Systems via Joint Adversarial Learning and Explicit Regularization

Recommender systems frequently exhibit two fundamental types of bias: user-side bias, where recommendations discriminate based on protected attributes such as gender, and item-side bias, where popular items dominate exposure irrespective of their actual relevance or quality. To address both simultaneously, we propose a dual-adversarial framework with two discriminators: a user-gender discriminator designed to mitigate gender bias by removing sensitive information from user embeddings, and an item-popularity discriminator aimed at reducing popularity bias among items. While adversarial learning encourages the model to generate bias-free embeddings by fooling discriminators, residual sensitive information may persist due to imperfect discriminator inference and distributional discrepancies. To further reduce this residual bias and enforce balanced, fair recommendations, we enhance the dual adversarial learning framework with two explicit fairness constraints that operate on the learned debiased user and item embeddings: an individual user fairness constraint that guarantees equitable recommendations for users with similar preferences, and a popularity regularization term that balances exposure across popular and less popular item groups. Experiments on three real-world datasets (MovieLens 1M, MovieLens 100K, and Book-Crossing) demonstrate significant gender predictability drops and popularity skew decreases, while maintaining competitive recommendation quality.

4.1 Introduction

Recommender systems (RSs) have become pivotal in shaping user experiences across a wide spectrum of digital platforms, including e-commerce, entertainment, and social media. By leveraging vast datasets to personalize content, RSs drive user engagement and satisfaction, facilitating the discovery of products, services, and content that align with individual preferences. However, as these systems grow in importance, concerns about fairness have emerged, particularly regarding the biases they inadvertently propagate [57]. These biases not only affect the personalization and utility of recommendations but also have significant social and economic consequences for both users and providers [7].

Recommender systems operate in a multi-sided market [134], where the platform serves two key stakeholders: users and providers (such as product manufacturers, content creators, or service providers). Providers rely heavily on the exposure their items receive on the platform, as visibility in recommendation lists directly impacts their success [69]. A higher ranking means more attention, translating into increased sales, greater user engagement, and ultimately higher revenue. However, when the system disproportionately favors a small subset of items, a winner-takes-all dynamic can emerge, where a few items dominate the platform's visibility, pushing others into obscurity. This creates a popularity bias [135], which leads to unequal exposure opportunities for items. While popular items naturally attract more attention, this skewed distribution leaves lesser-known or niche items marginalized, reducing their potential for engagement and exposure. As a result, items that consistently receive less visibility may struggle to succeed, causing providers to become dissatisfied and potentially leave the platform [136]. This not only diminishes the diversity of items available to users but also weakens the platform's ecosystem. Without a rich variety of offerings, the platform's ability to attract and retain diverse user groups is compromised. Item fairness thus becomes critical, as it addresses the unequal distribution of exposure caused by popularity bias, ensuring that all items, regardless of their popularity, are given fair opportunities to be recommended. This imbalance harms not just the providers but also the users, as it reduces the variety of content and products available, ultimately affecting the overall user experience and platform sustainability.

On the user side, user fairness is crucial in ensuring that individuals are not discriminated against based on sensitive attributes such as gender, race, socio-economic status, or other demographic characteristics [137]. In an ideal system, recommendations are made solely based on users' true preferences, irrespective of these attributes, thereby providing each user with an equitable opportunity to be exposed to relevant content [53]. However, in practice, recommender systems often face challenges when they inadvertently favor certain groups over others due to biases in the data or algorithms. Demographic bias is one of the most common forms of user fairness bias, where users from certain demographic groups may either receive fewer recommendations or be presented with biased content that does not align with their true preferences. For example, in job recom-

mendation systems, gender bias could result in women being underrepresented in recommendations for high-status technical roles, reinforcing harmful stereotypes about gender roles in the workplace [121].

To clearly explain the problem addressed in this chapter, consider Figure 4.1, which illustrates a simplified scenario of a movie recommendation system. This example can be readily extended to real-world scenarios, as explained later. The figure presents two female users (F1, F2) and two male users (M1, M2), each with different preferences for movie genres. F1 has 100% female-oriented preferences, meaning she prefers romance films exclusively, while F2 has 80% male-oriented preferences, favoring 80% horror and 20% romance. M1, on the other hand, has 100% male-oriented preferences, meaning he prefers horror films exclusively, and M2 has 80% romance and 20% horror in his preferences.

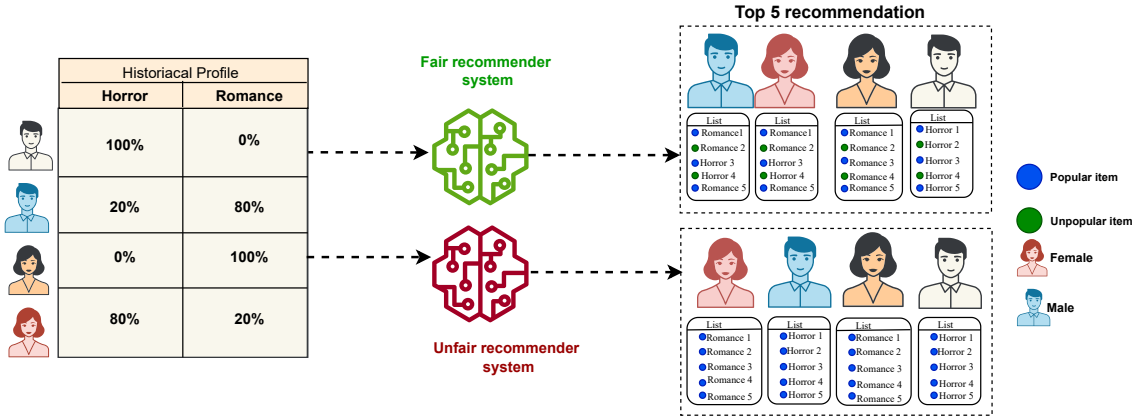


Figure 4.1: Simplified Movie Recommendation Scenario Illustrating the Two-Sided Fairness Concept in RSs

In the case of the unfair recommender system, the top recommendations for both F1 and F2 are fully dominated by romance films, even though F2 has 80% horror preference. This occurs because the system is trained on historical interaction data where romance films may be disproportionately associated with female users. The model learns to optimize overall predictive accuracy by capturing these majority patterns, causing it to rely on gender-based correlations rather than individual user preferences. Consequently, the embeddings for female users become clustered around stereotypical genre preferences [137], overriding F2's significant interest in horror films. Similarly, for males, M2, with 80% romance preference, would still receive top recommendations primarily for horror films, disregarding his true preference for a mix of genres. Additionally, the top recommendations for all users would be dominated by popular movies that are heavily exposed in the system, leaving out new or niche films that could better match the users' true preferences. In this case, the model is

unfair toward F2 and M2, as well as toward new or niche movies, which are ignored in favor of the most popular ones.

In contrast, a fair recommender system would take into account both user fairness and item fairness. For F2, the system would recommend a mix of romance and horror films, reflecting her 80% horror preference. Similarly, for M2, the system would recommend a combination of romance and horror, rather than focusing solely on horror. Moreover, the system would recommend a more diverse set of movies, including both popular and niche films that align with users' specific interests. This would ensure that less-exposed content, as well as popular content, is recommended, providing fair exposure to a wide range of items.

In this context, the challenge we are addressing is to design a system that achieves two-sided fairness for both users and items by giving equal opportunities to all items while still reflecting true user preferences, regardless of demographic characteristics such as gender. This approach ensures that no group of users is unfairly disadvantaged and that all items, whether popular or niche, have a fair chance of being recommended to users based on what they truly like, rather than being skewed by biased factors.

Although significant strides have been made in addressing both user and item fairness, the literature largely tends to treat these dimensions independently, addressing either user fairness or item fairness, but rarely both. This one-sided approach presents inherent limitations, as improving fairness for one group often inadvertently exacerbates bias in the other. For instance, prioritizing user fairness through demographic parity may lead to neglecting diversity in the items recommended, ultimately reinforcing popularity bias. Similarly, reducing item bias may inadvertently impact the fairness of recommendations to different user groups. These challenges underscore the importance of adopting a two-sided bias mitigation approach that accounts for both user fairness and item fairness in tandem, while preserving the utility and relevance of recommendations.

Various approaches have been proposed to mitigate biases in RSs, which can be broadly categorized into three paradigms: pre-processing, in-processing, and post-processing methods. Pre-processing approaches [121, 110] aim to debias training data by reweighting or resampling instances to balance representation across groups, but they often fail to address latent biases emerging during model training. In-processing methods [53, 138] modify learning algorithms directly by incorporating fairness constraints or adversarial learning, yet these typically focus on a single fairness dimension (e.g., either user or item fairness). Post-processing techniques [101, 139, 140] adjust recommendation outputs post-hoc through re-ranking or score calibration, but they risk compromising recommendation quality due to their decoupled nature from the core learning process. While adversarial learning has shown promise in removing sensitive attributes, its standalone application often leaves residual biases due to incomplete feature disentanglement. These limitations highlight a critical gap: current methods either address only one fairness side, creating trade-offs between user and item fairness, or use monolithic bias mitigation strategies that fail to fully disentangle

multi-dimensional biases while preserving recommendation utility.

To address these critical gaps, we introduce FUIREC (Fairness-aware Unified framework with adversarial and Explicit Regularization for two-sided Recommendation), a novel framework that simultaneously tackles user-side demographic biases and item-side popularity biases through an integrated hybrid architecture. Unlike existing approaches that address these biases separately or rely on partial solutions, FUIREC innovatively combines: (1) dual adversarial discriminators; a gender discriminator that adversarially removes gender-sensitive attributes from user embeddings, ensuring recommendations are invariant to the user's gender, and a popularity discriminator that eliminates popularity signals from item representations, decoupling relevance from historical popularity patterns; and (2) explicit fairness constraints that operate directly at the recommendation output, namely an individual user fairness term that enforces consistent treatment for users with similar preferences and a popularity regularizer that balances exposure across item groups. This dual-layer approach ensures that fairness is enforced both at the representation level (via adversarial debiasing) and at the output level (via explicit constraints).

To evaluate FUIREC's effectiveness, we conducted comprehensive experiments across three real-world recommendation scenarios. Our results demonstrate that FUIREC significantly outperforms existing methods in both user-side fairness (gender bias mitigation) and item-side fairness (popularity bias reduction), while maintaining competitive recommendation quality. Crucially, the model shows substantially reduced predictive accuracy when attempting to infer gender from debiased user embeddings or item popularity from debiased item embeddings, confirming effective bias removal. Further, FUIREC exhibits robust counterfactual fairness, where recommendations remain stable when synthetically altering user gender or item popularity. The framework also ensures strong group fairness, achieving near-identical performance between male and female users and between popular and unpopular items across all evaluated fairness metrics. Finally, FUIREC preserves high recommendation accuracy, with consistently strong NDCG and MAE scores comparable to unbiased state-of-the-art recommenders.

In summary, our key contributions are:

- We propose FUIREC, a novel unified framework that simultaneously addresses user gender bias and item popularity bias to ensure two-sided fairness through a hybrid adversarial-explicit approach.
- Our framework employs dual adversarial discriminators; a gender discriminator and a popularity discriminator, that work in tandem to remove sensitive attributes from both user and item embeddings.
- We introduce explicit fairness constraints that operate directly at the recommendation output: an individual user fairness term that enforces consistent treatment for users with similar

preferences, and a popularity regularizer that balances exposure across item groups. These constraints correct disparities that adversarial learning alone cannot eliminate.

- We conduct extensive experiments on three real-world datasets, demonstrating that FUIREC significantly outperforms state-of-the-art baselines in both user-side and item-side fairness while maintaining competitive recommendation quality.

4.2 Proposed Architecture

The proposed architecture, FUIREC, integrates four key components aimed at jointly optimizing recommendation accuracy and fairness. At its core is the recommendation generator, which functions as the primary model for learning informative user and item embeddings to produce high-quality recommendations. To mitigate systemic bias, we incorporate two adversarial discriminators: the user-gender discriminator aimed at mitigating gender bias and thereby promoting individual-level user fairness, while the item-popularity discriminator addresses popularity bias, ensuring more balanced exposure across the item space.

Recognizing that adversarial discriminators alone may not fully eliminate latent biases, we introduce two explicit fairness constraints to reinforce fairness objectives. The individual user fairness constraint encourages the model to treat users with similar preferences equitably, independent of their sensitive attributes. In parallel, the popularity regularizer penalizes exposure imbalances, promoting fair visibility for both popular and underrepresented items in a manner aligned with user interests.

The full architecture of FUIREC is illustrated in Figure 4.2, with a detailed description of each module presented in the subsequent sections.

4.2.1 Recommendation Model

The recommendation model serves a dual purpose in our framework: it functions as the primary prediction component for user-item interactions and simultaneously acts as the generator in the adversarial training process. Its primary objective is to learn user and item embeddings that are both predictive for accurate recommendations and free from sensitive information related to user demographics and item popularity.

Each user u is initially represented by an embedding vector $e_u \in \mathbb{R}^d$, and each item i by an embedding vector $e_i \in \mathbb{R}^d$, where d denotes the embedding dimensionality. These embeddings are learned parameters that encode latent user preferences and item characteristics. Critically, throughout training, these embeddings undergo continuous adversarial refinement through discriminators D_g (user-gender) and D_p (item-popularity), which penalize the generator whenever sensitive signals remain detectable. This adversarial pressure progressively purifies the embeddings, forcing

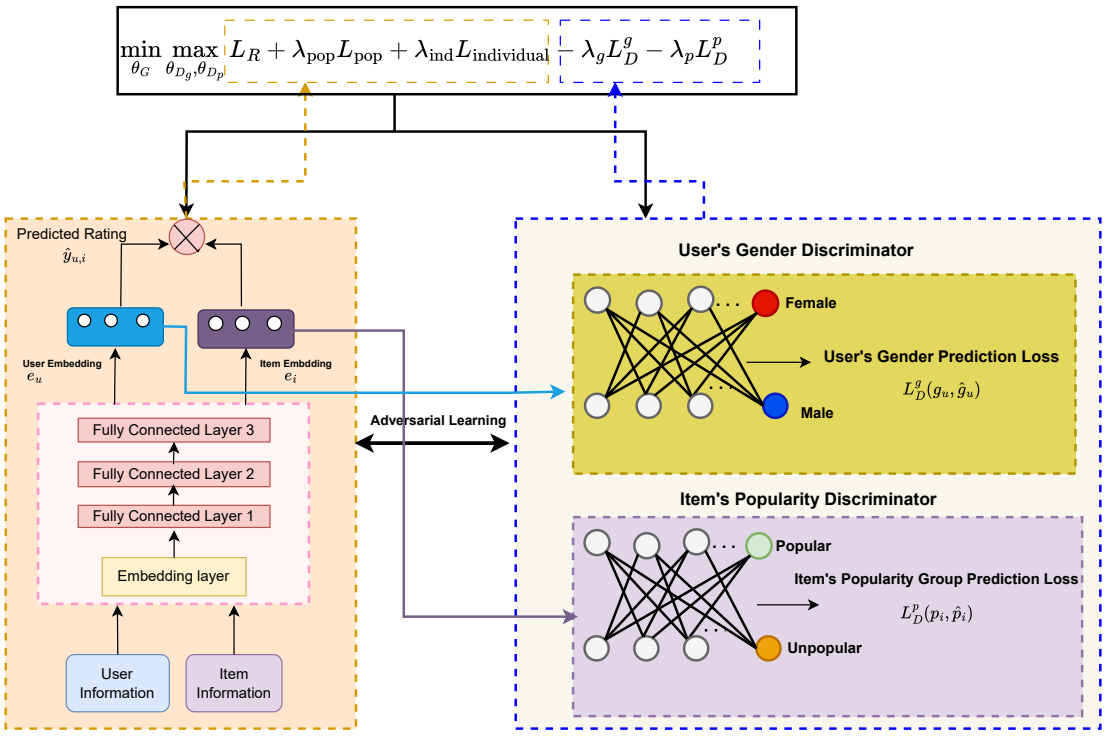


Figure 4.2: The overall architecture of FUIREC.

them to encode preference and relevance information while suppressing demographic and popularity cues. The resulting debiased user and item embeddings serve as the foundation for generating fair recommendations.

To generate recommendations, we utilize the debiased user embeddings e_u and debiased item embeddings e_i produced through the adversarial training process. These embeddings, having been purged of sensitive gender and popularity signals, ensure that the recommendation function operates on representations that prioritize user preferences and item relevance rather than demographic or popularity-driven patterns. The debiased embeddings are concatenated and passed through a multi-layer perceptron (MLP) to model complex user-item interactions:

$$h_{ui}^1 = \sigma(W_1[e_u \| e_i] + b_1) \tag{4.1}$$

$$h_{ui}^2 = \sigma(W_2 h_{ui}^1 + b_2) \tag{4.2}$$

$$\vdots \tag{4.3}$$

$$h_{ui}^n = \sigma(W_n h_{ui}^{n-1} + b_n) \tag{4.4}$$

where W_j and b_j denote the weight matrices and bias vectors for the j-th layer, $\sigma(\cdot)$ represents the activation function (ReLU), and $\|$ denotes concatenation. To enhance model expressiveness

and prevent overfitting, the MLP incorporates dropout layers, residual connections, and batch normalization.

Prediction Output. The output layer of the MLP generates the predicted interaction score $\hat{y}_{ui}$ using the debiased representations:

$$\hat{y}_{ui} = \sigma(W_o h_{ui}^n + b_o) \tag{4.5}$$

where W_o and b_o are the output layer parameters, and $\sigma(\cdot)$ is a sigmoid function that maps the output to the probability range $[0, 1]$. Because this prediction is computed from debiased embeddings e_u and e_i, the resulting scores $\hat{y}_{ui}$ are inherently less influenced by sensitive attributes and popularity bias, providing a fairer basis for ranking and recommendation.

Training Objective. The recommendation model is trained using binary cross-entropy loss, comparing the predicted interaction $\hat{y}_{ui}$ with the ground truth interaction y_{ui}:

$$\mathcal{L}_R = -\sum_{(u,i)\in\mathcal{D}} [y_{ui} \log(\hat{y}_{ui}) + (1 - y_{ui}) \log(1 - \hat{y}_{ui})] \tag{4.6}$$

where $\mathcal{D}$ represents the set of observed user-item interactions. This loss ensures that the model maintains high prediction accuracy while the adversarial components and fairness constraints work to ensure the predictions are fair and unbiased.

4.2.2 Adversarial Discriminators

To address these biases the user and item embeddings generated by the recommendation model, two adversarial discriminators are employed: the **user-gender discriminator** mitigates gender bias in the user's representation, while the **item-popularity discriminator** mitigates popularity bias in the item's representation.

User-Gender Discriminator

The user-gender discriminator D_g ensures that user embeddings e_u are free from sensitive information related to the gender of the user. This is crucial to ensure user fairness and prevent the recommendation model from making biased decisions based on gender. By employing adversarial learning, the system ensures that user embeddings are gender-neutral while maintaining prediction accuracy.

Thi discriminator D_g functions as a binary classifier, attempting to predict the sensitive attribute S_u (user gender) from the user embedding e_u. The prediction is given by:

$$\hat{g}_u = D_g(e_u) = \text{softmax}(W_g^2 \sigma(W_g^1 e_u + b_g^1) + b_g^2), \tag{4.7}$$

where W_g^1, W_g^2, b_g^1, and b_g^2 are the weights and biases of the discriminator layers and $\sigma(\cdot)$ is the activation function

The true gender label g_u for user u is defined as:

$$g_u = \begin{cases} 1 & \text{if user } u \text{ is male,} \\ 0 & \text{if user } u \text{ is female.} \end{cases} \tag{4.8}$$

The discriminator D_g is trained to maximize its ability to classify users as male or female based on their embeddings. The training objective is the cross-entropy loss:

$$L_D^g = -\sum_{u \in U} \left[g_u \log(\hat{g}_u) + (1 - g_u) \log(1 - \hat{g}_u) \right]. \tag{4.9}$$

Item-Popularity Discriminator

Similarly, the item-popularity discriminator D_p mitigates popularity bias in item embeddings e_i. It predicts whether an item is popular or unpopular based on its embedding:

$$\hat{p}_i = D_p(e_i) = \text{softmax}(W_p^2 \sigma(W_p^1 e_i + b_p^1) + b_p^2), \tag{4.10}$$

where W_p^1, W_p^2, b_p^1, and b_p^2 are the weights and biases of the discriminator layers. The ground truth popularity p_i for an item i is:

$$p_i = \begin{cases} 1 & \text{if item } i \text{ is popular,} \\ 0 & \text{if item } i \text{ is unpopular.} \end{cases} \tag{4.11}$$

The discriminator is trained using a cross-entropy loss:

$$L_D^p = -\sum_{i \in I} \left[p_i \log(\hat{p}_i) + (1 - p_i) \log(1 - \hat{p}_i) \right]. \tag{4.12}$$

4.2.3 Fairness Constraints Terms

User Fairness Constraint

To ensure equitable treatment of users regardless of their sensitive attributes, we introduce an individual fairness constraint. This constraint is grounded in the principle of individual fairness, which posits that similar individuals should be treated similarly. In the context of recommendation, this implies that users with similar preferences and behavioral histories should receive consistent prediction scores, irrespective of their gender.

A critical design consideration is the definition of "similarity." If similarity were calculated using raw user data or entangled representations, the metric might capture demographic correlations (e.g., grouping users by gender) rather than actual interests. To address this, our framework calculates user similarity using the adversarially debiased embeddings e_u. These embeddings are the output of the generator, which is simultaneously subjected to the adversarial game with the

User-Gender Discriminator (D_g). Because the discriminator actively penalizes the encoding of gender information, the resulting embeddings e_u are purged of demographic signals. Consequently, the proximity between two users in this specific latent space primarily captures preference alignment rather than sensitive attribute correlation.

The constraint operates by penalizing the model whenever it generates divergent predictions for users who are identified as behaviorally similar. The individual fairness loss is defined as:

$$\mathcal{L}_{individual} = \sum_{(u,u')\in\mathcal{B}} \text{sim}(e_u, e_{u'}) \cdot |\hat{y}_{ui} - \hat{y}_{u'i}| \tag{4.13}$$

where $\text{sim}(e_u, e_{u'})$ represents the cosine similarity between the debiased embeddings, acting as a dynamic weighting factor.

- **High Similarity:** If two users have highly aligned debiased embeddings (indicating similar tastes), the weight is large. This forces the model to minimize the difference in their predicted scores ($|\hat{y}_{ui} - \hat{y}_{u'i}|$) to zero.
- **Low Similarity:** If users have divergent preferences, the weight is small, allowing the model flexibility to output different scores without penalty.

Popularity Regularizer

The item-popularity discriminator mitigates bias in the latent representation space. To complement this, we introduce a popularity regularization term that operates directly on the recommendation output. This regularizer explicitly penalizes the disparity in predicted exposure between popular and unpopular items, ensuring that the model does not simply default to recommending trending items.

Items are categorized into two groups based on the Pareto Principle (20/80 rule):

- **Popular Items (I_{pop}):** The top 20% of items with the highest interaction frequency.
- **Unpopular Items (I_{unpop}):** The remaining 80% of items (tail items).

The regularizer minimizes the absolute difference in the average predicted probability (exposure) between these two groups. To ensure computational efficiency, this is computed dynamically within each training batch $\mathcal{B}$. Let $\mathcal{B}_{pop}$ and $\mathcal{B}_{unpop}$ denote the sets of user-item interactions in the current batch involving popular and unpopular items, respectively. The expected exposure for each group is calculated as:

$$E_{pop} = \frac{1}{|\mathcal{B}_{pop}| + \epsilon} \sum_{(u,i)\in\mathcal{B}_{pop}} \hat{y}_{ui}, \quad E_{unpop} = \frac{1}{|\mathcal{B}_{unpop}| + \epsilon} \sum_{(u,i)\in\mathcal{B}_{unpop}} \hat{y}_{ui} \tag{4.14}$$

where ϵ is a small constant to prevent division by zero in skewed batches. The popularity regularization loss is then defined as:

$$L_{pop} = |E_{pop} - E_{unpop}| \tag{4.15}$$

By minimizing L_{pop}, the framework forces the generator to uplift the predicted scores of relevant unpopular items, thereby promoting a more balanced catalog exposure.

4.2.4 Objective Function of FUIREC

Combining the various loss functions outlined above, the final objective function for the proposed FUIREC model L_G can be formulated as:

$$L_G = L_R - \lambda_g L_D^g - \lambda_p L_D^p + \lambda_{pop} L_{pop} + \lambda_{ind} L_{individual}$$

where L_R is the recommendation loss, L_D^g and L_D^p are the losses for the gender and popularity discriminators, respectively, and L_{pop} and $L_{individual}$ are the fairness constraints for item popularity and individual fairness. The hyperparameters λ_g, λ_p, λ_{pop}, and λ_{ind} control the contributions of the adversarial losses, and the user and item fairness constraints. These hyperparameters balance the trade-off between recommendation accuracy and fairness, and their influence will be discussed in Section 4.

4.2.5 Adversarial Training: The MinMax Game

The FUIREC framework employs adversarial training to ensure that user and item embeddings are free from sensitive information. In each training iteration, the generator (recommendation model) aims to produce embeddings that are both predictive for recommendations and indistinguishable with respect to sensitive attributes, while the discriminators D_g and D_p attempt to classify gender and popularity from these embeddings. This creates a minimax optimization:

$$\min_{G_\theta} \max_{D_g, D_p} \mathcal{L}_R - \lambda_g \mathcal{L}_D^g - \lambda_p \mathcal{L}_D^p + \lambda_{\text{pop}} \mathcal{L}_{\text{pop}} + \lambda_{\text{ind}} \mathcal{L}_{\text{individual}} \tag{4.16}$$

where the generator minimizes recommendation loss $\mathcal{L}_R$ while simultaneously minimizing the discriminators' ability to classify sensitive attributes, and the discriminators maximize their classification accuracy.

Training Dynamics and Stability. As detailed in Algorithm 2, we adopt an alternating update strategy. In each iteration, discriminators are updated for $k_d = 5$ gradient steps to effectively learn bias patterns, followed by a single generator update. This frequency ratio prevents generator-discriminator imbalance and maintains steady convergence. The discriminator updates strengthen their ability to detect residual bias, while the generator update produces increasingly debiased embeddings that fool the discriminators while satisfying fairness constraints.

Algorithm 2 Training Strategy of FUIREC

Require: Generator G, Discriminators D_g, D_p, Hyperparams $\lambda_g, \lambda_p, \lambda_{\text{ind}}, \lambda_{\text{pop}}$

Initialize parameters $\theta_G, \theta_{D_g}, \theta_{D_p}$
for each epoch **do**
  for each batch $\mathcal{B}$ **do**
    // Phase 1: Train Discriminators
    for $k = 1$ to K_d **do**
      Get embeddings: $e_u, e_i \leftarrow G(u, i)$
      Detach embeddings: $\tilde{e}_u \leftarrow \text{sg}(e_u), \tilde{e}_i \leftarrow \text{sg}(e_i)$
      Predict attributes: $\hat{g} \leftarrow D_g(\tilde{e}_u), \hat{p} \leftarrow D_p(\tilde{e}_i)$
      Update $\theta_{D_g}, \theta_{D_p}$ to **minimize** classification error
    end for
    // Phase 2: Train Generator (Recommendation & Fairness)
    Get embeddings: $e_u, e_i \leftarrow G(u, i)$
    Calculate Rec Loss: $\mathcal{L}_R$ using $\hat{y}_{ui}$
    Adversarial Loss: Pass e_u, e_i to D_g, D_p (keep gradients)
    $\mathcal{L}_{adv} = -(\lambda_g \log(D_g(e_u)) + \lambda_p \log(D_p(e_i)))$
    Fairness Constraints:
    Calculate $\mathcal{L}_{\text{pop}}$ using e_i
    Calculate $\mathcal{L}_{\text{ind}}$ using $\text{sg}(e_u)$ for weights (Eq. 2)
    Total Loss: $\mathcal{L}_G = \mathcal{L}_R + \mathcal{L}_{adv} + \lambda_{\text{pop}}\mathcal{L}_{\text{pop}} + \lambda_{\text{ind}}\mathcal{L}_{\text{ind}}$
    Update θ_G to **minimize** $\mathcal{L}_G$
  end for
end for

Empirical Training Visualization. Figure 4.3 illustrates the adversarial training dynamics of the FUIREC OVER epochs, demonstrating the min-max game between the Generator and Discriminators.

During the initial training phase (Epochs 1–10), the User and Item Discriminator losses are relatively low.This indicates that they are highly effective at distinguishing sensitive attributes; they can easily detect the user's gender and the item's popularity level from the latent embeddings. Conversely, the Generator Loss exhibits a significantly high magnitude during this period, which means that the initial embeddings are heavily entangled with sensitive attributes, and the generator fails to fool the discriminators.

After Epoch 10 and continuing to Epoch 35, a dramatic shift occurs. The Generator loss begins to decrease significantly, meaning it has started learning how to debias the user and item embeddings. As the generator improves, we simultaneously see the Discriminator losses start to increase

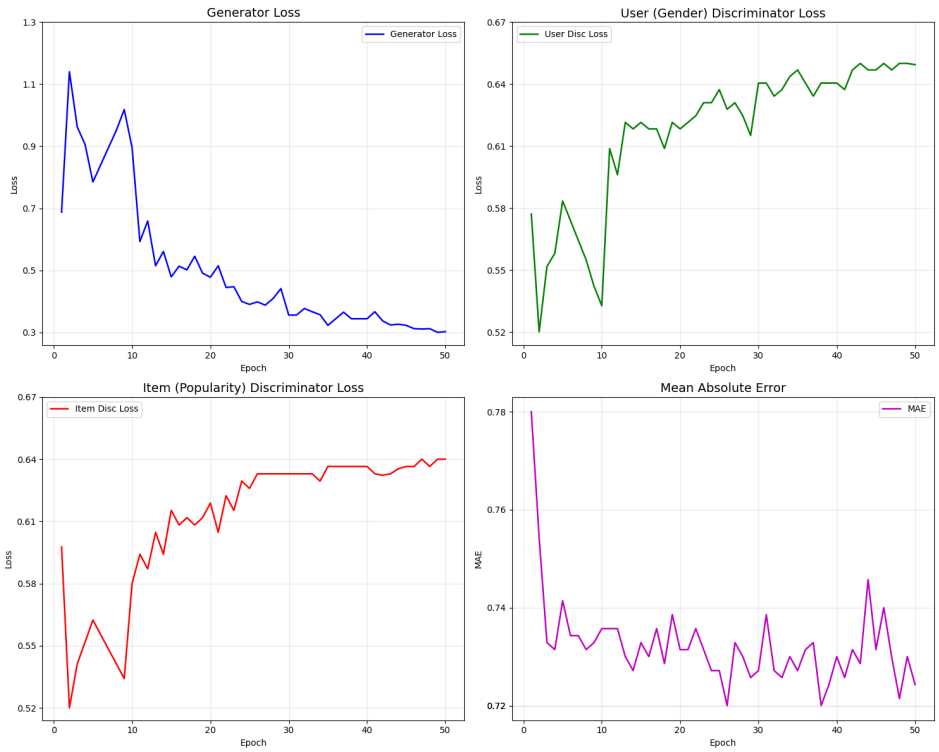


Figure 4.3: Training dynamics of FUIREC across epochs.

rapidly, meaning that it is successfully fooling the discriminators and optimizing the embeddings to be invariant to sensitive attributes, making it increasingly hard for them to identify gender or popularity signals within the representations.

Beyond Epoch 40, the system converges to a stable equilibrium. The Generator Loss plateaus at a low value, indicating the successful minimization of bias within the latent space. Concurrently, the Discriminator losses stabilize at a high level. This plateau implies that the adversaries have been forced into a state of random guessing, effectively demonstrating that the learned representations are invariant to specific gender and popularity biases while retaining the necessary information for accurate recommendations.

4.3 Experiments

In this section, we perform extensive experiments to assess the fairness and effectiveness of the proposed FUIREC model. Additionally, we conduct ablation studies and other analyses to gain deeper insights into the performance of the model. We specifically aim to address the following research questions (RSQs).

- **RQ1:** How does FUIREC perform compared to traditional and fairness-aware recommendation algorithms in terms of both recommendation quality and fairness?

- **RQ2:** How effectively do FUIREC's discriminators mitigate user gender bias and item popularity bias?
- **RQ3** How does the popularity regularization enhance the exposure of unpopular group and then enhance the item's fairness?
- **RQ4:** How does the user fairness constraint enhance the equality of treatment for similar users?
- **RQ5:** How do the weights of different FUIREC components impact the performance of our model?

4.3.1 Dataset Description

We conducted extensive experiments on three public recommendation datasets. The key statistics for each dataset are summarized in Table 5.2, and a brief description of each is provided below:

- **MovieLens 1M (ML-1M)**: This dataset contains 1 million ratings from 6,040 users on 3,952 movies. It is widely used for benchmarking in movie recommendation systems, offering a rich set of user-movie interactions. The dataset includes demographic information such as gender, which allows for fairness analysis.
- **MovieLens 100K (ML-100K)**: A smaller version of the MovieLens dataset, with 100,000 ratings from 943 users on 1,682 movies. Due to its manageable size and the availability of user demographic data, it is one of the most commonly used datasets for evaluating recommendation algorithms.
- **Book-crossing**: This dataset contains over 1 million ratings provided by 100,000 users on approximately 270,000 books. It is widely used for benchmarking in book RecSys, offering a rich set of user-book interactions. The dataset also includes demographic information such as user age and gender, which enables the analysis of fairness and bias in book recommendations.

Table 4.1: Dataset Statistics

Dataset	Users	Items	Interactions
ML-1M	6,040	3,952	1,000,000
ML-100K	943	1,682	100,000
Book-Crossing	278,858	271,379	1,149,780

To analyze our model's performance, we split the user set into a training set, verification set, and test set by a ratio of 7:1:2. Adopting the same setting as proposed in [69, 99], we allocate at least

10 ratings of a user for evaluation as query set and use the remaining ratings as support set for training the model.

In RecSys, an item is typically classified as "popular" based on its interaction frequency within the dataset. To establish this classification, a predefined threshold is applied. In line with [104] and the Pareto Rule according to [141], we consider the top 20% of items with the highest rating count to be "popular items", while the other 80% of the items are considered as "non-popular items" (see Fig. 4.4).

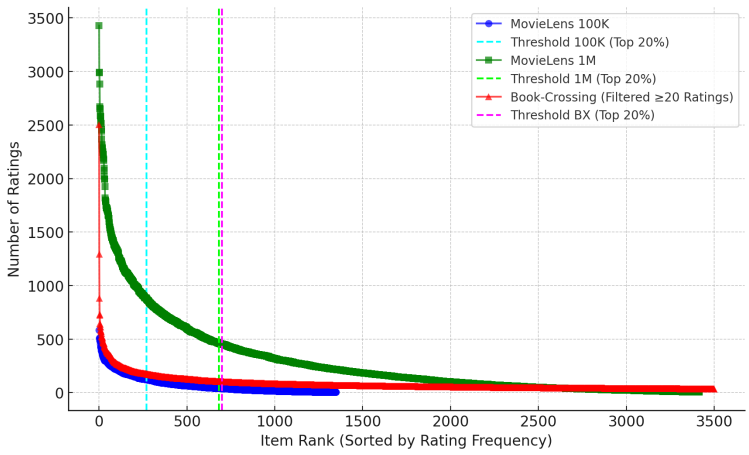


Figure 4.4: Rating Frequency and Popularity Threshold in Datasets.

In the context of fair recommendation, the gender attribute is selected as the sensitive attribute for the ML-100K and ML-1M datasets, while the age attribute is chosen as the sensitive attribute for the BookCrossing dataset. To evaluate individual fairness, we assess whether sensitive information is still embedded in the learned user representations. Following practices commonly adopted in fairness-aware recommendation studies [69, 99], we use the training users' attributes as ground truth. First, the system learns how to represent users (their preferences and behaviors) while aiming to ignore sensitive information like gender. Then, a simple linear classifier is trained to predict users' sensitive attributes (gender) using these learned representations as input. By testing this model on unseen users (those in the test set), we check if sensitive information can still be predicted. If the classifier struggles to guess sensitive attributes like gender accurately, it means the representations are fairer because they are not exposing that sensitive information. To measure how well this is working, we use a metric called AUC (Area Under the Curve), which helps evaluate the fairness of these user representations. Similarly, for item fairness, a simple linear classifier is trained to predict to which group an item belongs using these learned representations as input.

To facilitate reproducibility, we implemented the proposed framework using PyTorch 1.13 and Python 3.8. All experiments were conducted with an NVIDIA, 32GB memory and 64GB of RAM. We fixed the embedding dimension at $d = 48$ for both users and items.

Model Architecture:

- **Generator:** The recommendation model is implemented as a Multi-Layer Perceptron (MLP). It takes the concatenated user and item embeddings (input dimension $2d = 96$) and passes them through three fully connected hidden layers with sizes $\{64, 32, 16\}$, using ReLU activation functions. The final output layer is a single neuron with a Sigmoid activation to predict the interaction probability.
- **Discriminators:** Both the User-Gender Discriminator (D_g) and Item-Popularity Discriminator (D_p) consist of a two-layer MLP. Each discriminator accepts a 48-dimensional embedding as input, processes it through a hidden layer of size 16 with LeakyReLU activation, and ends with a Softmax output layer for binary classification.

We initialized all model parameters using Xavier initialization to ensure stable convergence. The model was optimized using the Adam optimizer with default betas ($\beta_1 = 0.9, \beta_2 = 0.999$) and a weight decay of 110^{-5} to prevent overfitting. We utilized a batch size of 256 for all datasets. The adversarial training followed a strict MinMax game strategy. As described in Algorithm 1, we employed an alternating update schedule:

- *Discriminator Phase:* For every iteration, we first updated the two discriminators (D_g and D_p) for $k_d = 5$ steps to ensure they provided a sufficiently strong adversarial signal.
- *Generator Phase:* Subsequently, we updated the recommendation model (Generator) for 1 step to minimize the combined recommendation and fairness losses.

To ensure the computational scalability of the proposed framework, particularly regarding the individual fairness constraint which theoretically requires $O(N^2)$ pairwise comparisons, we employed a mini-batch optimization strategy.

Instead of computing the fairness losses over the entire dataset globally, these constraints are enforced dynamically within each training batch $\mathcal{B}$ of size 256.

- User Fairness: The pairwise similarity $\text{sim}(e_u, e_{u'})$ and the subsequent penalty $\mathcal{L}_{ind}$ are computed exclusively for all user pairs (u, u') present within the current mini-batch. This reduces the complexity per iteration to $O(|\mathcal{B}|^2)$, ensuring linear scalability with respect to the total dataset size.
- Popularity Regularization: Similarly, the expected exposure scores E_{pop} and E_{unpop} are calculated as the mean predicted scores of the popular and unpopular items currently sampled in the batch.

Hyperparameter Search: We performed a hyperparameter search using grid search over a predefined range $\lambda_g, \lambda_p, \lambda_{\text{pop}}, \lambda_{\text{ind}} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We also tested embedding sizes in the

range $\{16, 32, 48, 64, 128\}$ and the learning rate values in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We then evaluated different combinations of these parameters using a validation set and selected the configuration that achieved the best trade-off between fairness metrics (AUC, GF) and recommendation accuracy metrics (NDCG, MAE). Based on this search, the optimal values used for our final experiments were: $\{\lambda_g = 0.1, \quad \lambda_p = 0.1, \quad \lambda_{\text{pop}} = 0.01, \quad \lambda_{\text{ind}} = 0.01\}$, with embedding size 48 and learning rate $\eta = 0.001$. We utilized an Early Stopping mechanism, monitoring the MAE score on the validation set;

4.3.2 Evaluation Metrics

For a fair comparison of accuracy and fairness with other models, we adopted the same metrics used by [69].

We evaluate the rating prediction and ranking performance of our FUIREC framework using MAE and NDCG metrics respectively. Higher values of NDCG and lower values of MAE represent more accurate and effective recommendations.

- **Normalized Discounted Cumulative Gain (NDCG):** NDCG assesses the ranking quality of recommendations by comparing predicted rankings with ideal rankings. Higher values indicate better ranking quality.

$$NDCG_k = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{DCG_u^k}{IDCG_u^k} \tag{4.17}$$

$$DCG_u^k = \sum_{r=1}^{k} \frac{2^{y_{ru}} - 1}{\log_2(1+r)} \tag{4.18}$$

- **Mean Absolute Error (MAE):** MAE measures the average absolute error between the predicted and true values. A lower MAE indicates better rating prediction.

$$MAE = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{1}{|D_u^q|} \sum_{i \in D_u^e} |y_{ui} - \hat{y}_{ui}| \tag{4.19}$$

- **Area Under the Curve (AUC):** The Area Under the Curve (AUC) is a widely used metric in classification tasks that evaluates performance across all possible threshold values by considering the True Positive Rate (TPR) and False Positive Rate (FPR). In the context of fair recommender systems, AUC is commonly employed to assess how effectively sensitive information has been removed from learned representations. Specifically, in this work, AUC is used to measure the extent to which sensitive attributes such as gender can still be inferred from user embeddings. To compute this, we train a simple binary classifier on the learned representations and evaluate its predictive performance using AUC.

- **Counterfactual Fairness (CF):** Evaluates a recommendation model's robustness to changes in a user's sensitive attribute while keeping other attributes constant. Specifically, it measures how much the predicted scores change when a user's sensitive attribute is altered to its counterfactual value.

  Let a_u denote the sensitive attribute of user u (e.g., gender), and a'_u represent the counterfactual value (e.g., switching from female to male). Let x_u denote the user's other attributes, which remain unchanged. The set D_q^u denotes the query set of items for user u, and U_f is the set of users evaluated for fairness.

  The CF metric is defined as:

$$\text{CF} = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{1}{|D_q^u|} \sum_{i \in D_q^u} |\hat{y}_{ui}(x_u, a_u) - \hat{y}_{ui}(x_u, a'_u)| \tag{4.20}$$

  where:

  - $\hat{y}_{ui}$ is the predicted score for user u and item i using the original sensitive attribute a_u,
  - $\hat{y}_{ui}^{a'_u}$ is the predicted score when the sensitive attribute is counterfactually changed to a'_u,
  - D_q^u is the set of query items for user u,
  - U_f is the fairness evaluation user set.

  A lower CF value indicates a fairer model that is less sensitive to changes in protected attributes.

- **Group Fairness (GF):** GF compares recommendation accuracy (MAE) across different groups (e.g., based on sensitive attributes such as gender or item popularity). A lower GF value indicates better fairness between groups.

$$GF = \left| \frac{1}{|A_1|} \sum_{u_1 \in A_1} R(u_1) - \frac{1}{|A_2|} \sum_{u_2 \in A_2} R(u_2) \right|$$

4.3.3 Baseline models

We compare FUIREC with several state-of-the-art models across two categories.

To evaluate the recommendation quality of our model, we use the following baselines:

- **PPR**: (Pairwise Preference Regression): This method estimates what users like by using a special kind of regression model called bilinear regression [142].

- **Wide & Deep**: This approach uses a deep neural network to predict whether a user will prefer a certain item [35].

- **DropoutNet:** By applying the dropout technique within a deep neural network, this model learns input features and addresses the cold-start problem, where little user data is available [41].
- **NLBA:** This is a cold-start neural network recommendation system where the weights of the model stay the same for all users, but the biases of the output and hidden layers are personalized for each user [42].
- **MELU:** A model that estimates user preferences using meta-learning, allowing it to quickly adjust to new users [143].

To evaluate the fairness of our model, we use the following baselines:

- **MACR:** It offers a general framework to reduce bias in RecSys. It focuses on capturing bias related to sensitive information in the given context [105].
- **CLOVER:** This is a comprehensive fair meta-learning framework focuses on ensuring three types of fairness individual, counterfactual, and group fairness by using a multi-task adversarial learning approach, making it adaptable to various meta-learned models [69].
- **DALFRec:** A dual-side based adversarial learning for fairness-aware recommendation algorithm that filters sensitive information from both user and item sides using adversarial learning [99].

4.4 Results and Discussion

To comprehensively evaluate our model's performance and understand the contributions of its individual components, we conducted a series of ablation studies. These experiments explored the impact of various factors, including the two adversarial components, the popularity regularizer, user fairness constraints, and different hyperparameter settings.

4.4.1 RQ1: Comparative Performance of FUIREC Across Fairness and Accuracy Metrics

Table 5.3 presents a comprehensive evaluation of FUIREC against traditional recommendation algorithms and fairness-aware frameworks across three benchmark datasets (ML-1M, BookCrossing, and ML-100K). The results demonstrate FUIREC's consistent superiority in both fairness metrics (AUC, GF, CF) and recommendation quality (MAE, NDCG), establishing new state-of-the-art performance.

Table 4.2: Performance comparison on ML-1M, BookCrossing, and ML-100k datasets. ↑: higher better, ↓: lower better.

Model	**MAE ↓**	**NDCG ↑**	**AUC ↓**	**GF ↓**	**CF ↓**
		ML-1M			
PPR	0.943	0.672	0.187	0.052	0.187
Wide&Deep	0.815	0.691	0.214	0.055	0.214
DropoutNet	0.813	0.702	0.204	0.063	0.204
NLBA	0.795	0.701	0.254	0.056	0.254
MELU	0.743	0.755	0.264	0.054	0.264
CLOVER	0.731	0.756	0.632	0.040	0.044
DALFRec	0.735	0.758	0.533	0.019	0.005
FUIREC	**0.723**	**0.772**	**0.520**	**0.017**	**0.003**
		BookCrossing			
PPR	2.374	0.598	0.694	0.101	0.194
Wide&Deep	1.945	0.627	0.156	0.098	0.156
DropoutNet	1.855	0.634	0.194	0.103	0.194
NLBA	1.718	0.651	0.556	0.112	0.156
MELU	1.332	0.723	0.559	0.103	0.259
CLOVER	1.352	0.723	0.546	0.103	0.027
DALFRec	1.330	0.725	0.544	0.081	0.002
FUIREC	**1.279**	**0.753**	**0.525**	**0.075**	**0.001**
		ML-100k			
PPR	1.398	0.486	0.772	0.051	0.095
Wide&Deep	1.256	0.863	0.112	0.046	0.112
DropoutNet	1.172	0.842	0.131	0.044	0.131
NLBA	1.213	0.553	0.891	0.047	0.129
MELU	0.892	0.652	0.974	0.048	0.157
CLOVER	0.880	0.666	0.562	0.036	0.032
DALFRec	0.869	0.677	**0.505**	0.026	0.007
FUIREC	**0.776**	**0.721**	0.511	**0.022**	**0.005**

The quantitative evidence reveals several key insights. First, FUIREC achieves superior fairness metrics compared to all baselines, with the lowest GF and CF values. This improvement is particularly notable given the simultaneous enhancement of recommendation quality, where FUIREC maintains the lowest MAE, and highest NDCG scores across all datasets.

A particularly revealing observation arises from comparing FUIREC to adversarial-only baselines like CLOVER and DALFRec. While these models integrate adversarial learning to reduce sensitive attribute leakage from embeddings, they often exhibit residual unfairness at the recommendation level. This is most apparent in their elevated GF values, indicating that adversarial training alone does not fully neutralize sensitive correlations. FUIREC addresses this limitation by introducing an

explicit fairness constraint, which operates directly on recommendation outputs to enforce equitable treatment for users with similar behavioral patterns. This is evidenced by the consistent reduction in GF compared to adversarial-only approaches across all datasets.

The framework's effectiveness stems from its multi-level fairness strategy. The adversarial components minimize attribute predictability in embeddings (AUC), while individual fairness constraints correct output-level disparities (GF). Moreover, these observations yield several salient implications for the design of fairness-aware recommender systems:

- Adversarial learning alone cannot eliminate completely the bias.
- The incorporation of explicit fairness constraints constitutes a necessary intervention to correct persistent disparities that adversarial learning alone fails to address.
- The integration of representation-level debiasing and output-level fairness constraints can simultaneously optimize equity and recommendation performance.

Table 5.3 confirms that FUIREC outperforms state-of-the-art baselines by addressing a fundamental gap in fairness-aware learning: the persistence of residual bias post-adversarial training. The performance of the proposed approach stems from the integration of representation-level and recommendation-level fairness mechanisms. By simultaneously addressing embedding biases through adversarial learning and output disparities through fairness constraints, it achieves a more comprehensive fairness mechanism than previous approaches while maintaining state-of-the-art recommendation quality.

4.4.2 RS2: Impact of the Dual Adversarial Components

In the ablation study presented in Table 4.3, we investigate the isolated effects of the user-gender (D_g) and item-popularity (D_p) discriminators by deactivating the explicit fairness constraints. This allows us to evaluate how effectively each adversarial component mitigates specific biases (both user and popularity constraints here are deactivated).

When only the user-gender discriminator is active (FUIREC($-D_p$)), we observe that user fairness metrics improve significantly, specifically, there is a sharp decrease in AUC-G ($0.859 \rightarrow 0.5211$ on ML-1M) and CF-G. However, the model performs poorly on item fairness, leaving popularity bias (high AUC-P and GF-p).

Conversely, when only the item discriminator is active (FUIREC ($-D_g$)), item fairness metrics are improved (AUC-P drops to 0.531 on ML-1M), but user fairness remains poor with high AUC,GF and CF. This confirms that debiasing one side does not implicitly correct the other.

When both discriminators are activated (Full FUIREC), we see simultaneous improvements for both metric sets, achieving balanced debiasing across users and items. However, a small fairness-

Table 4.3: Ablation study on dual adversarial discriminators (G: gender, P: popularity). Bold: best values.

Variant	NDCG	MAE	AUC (G)	AUC (P)	GF (G)	GF (P)	CF (G)	CF (P)
ML-1M								
FUIREC($-D_p$,$-D_g$)	0.781	**0.717**	0.859	0.716	0.052	0.084	0.056	0.071
FUIREC($-D_p$)	**0.777**	0.725	**0.521**	0.721	0.027	0.087	**0.002**	0.074
FUIREC($-D_g$)	0.769	0.735	0.844	**0.531**	0.020	0.038	0.005	**0.005**
FUIREC	0.774	0.721	0.524	0.537	**0.019**	**0.035**	0.003	0.005
BookCrossing								
FUIREC($-D_p$,$-D_g$)	0.769	**1.235**	0.869	0.693	0.124	0.107	0.094	0.109
FUIREC($-D_p$)	0.761	1.241	0.539	0.649	0.115	0.082	0.093	0.004
FUIREC($-D_g$)	0.764	1.253	0.841	**0.548**	0.088	0.091	0.005	0.097
FUIREC	**0.751**	1.264	**0.535**	0.552	**0.082**	**0.078**	**0.005**	**0.004**
ML-100K								
FUIREC($-D_p$,$-D_g$)	**0.731**	**0.755**	0.780	0.720	0.070	0.068	0.094	0.006
FUIREC($-D_p$)	0.727	0.759	0.512	0.724	0.057	0.068	**0.094**	0.006
FUIREC($-D_g$)	0.719	0.766	0.557	0.549	0.026	0.052	0.005	0.003
FUIREC	0.726	0.765	**0.511**	**0.544**	**0.022**	0.049	0.005	**0.003**

accuracy trade-off is observed in specific scenarios, such as on ML-100K where MAE increases slightly from 0.755 to 0.765, reflecting the marginal cost of debiasing.

4.4.3 RQ3: Impact of Popularity Regularization

Table 4.4 presents a systematic ablation study to isolate the contributions of the item-popularity discriminator (D_p) and the popularity regularizer (L_{pop}) (Here both user discriminator and user fairness constraints are deactivated).

The baseline configuration, which lacks both fairness components, exhibits severe exposure disparity, particularly on ML-1M where popular items dominate recommendations with an Exposure Ratio (ER) of 11.50. Introducing the adversarial discriminator alone reduces this ratio to 4.62, demonstrating that while representation-level debiasing effectively penalizes the encoding of popularity signals, it fails to fully eliminate implicit residual bias.

The complete FUIREC model, which combines adversarial learning with explicit regularization, achieves the optimal trade-off, further lowering the Exposure Ratio to 1.88 on ML-1M while maintaining high accuracy. Crucially, the impact of the explicit regularizer is most pronounced on the sparse Book-Crossing dataset. Here, the adversarial discriminator alone yields only marginal im-

Table 4.4: Impact of the Popularity Regularization term.

Variant	**MAE**	**NDCG**	**Popular Item Exposure %**	**Unpopular Item Exposure %**	**Exposure Ratio**
			ML-100K		
FUIREC(- D_p, - L_{pop})	0.755	0.731	81.0	19.0	4.26
FUIREC(- L_{pop})	0.774	0.723	67.0	33.0	2.03
FUIREC	0.771	0.726	57.5	42.5	1.35
			ML-1M		
FUIREC(- D_p, - L_{pop})	0.717	0.781	92.0	8.0	11.50
FUIREC(- L_{pop})	0.731	0.768	82.2	17.8	4.62
FUIREC	0.723	0.778	65.3	34.7	1.88
			BookCrossing		
FUIREC(- D_p, - L_{pop})	1.235	0.769	81.5	18.5	4.41
FUIREC(- L_{pop})	1.252	0.752	79.0	21.0	3.76
FUIREC	1.251	0.755	60.7	39.3	1.54

provements (4.41 → 3.76) due to the fuzzy latent boundaries caused by data sparsity. The addition of the popularity regularizer is decisive, driving the ratio down to 1.54.

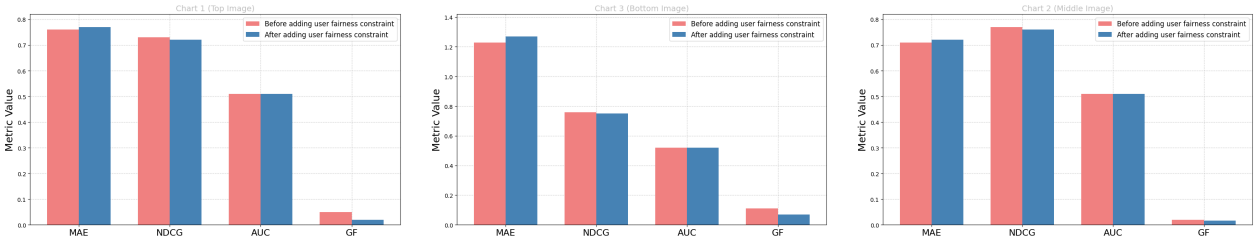
These findings validate the necessity of a hybrid fairness approach. While adversarial learning addresses the encoding of bias in latent representations, the popularity regularizer acts as a complementary safeguard on the output distribution. This outcome-level intervention is particularly indispensable for sparse datasets, ensuring equitable exposure without degrading recommendation utility.

4.4.4 RQ4: Impact of the User Fairness Constraint

Figure 4.5 illustrates the critical role of the user fairness constraint in addressing the limitations of adversarial learning. This fairness constraint complements this by operating directly at the recommendation output level, enforcing consistency for users with similar behavioral profiles regardless of their sensitive attributes. The empirical results confirm that this multi-level intervention improves both equity and utility. Across all three datasets, the inclusion of the user constraint consistently reduces Group Fairness disparities (e.g., ML-1M GF: 0.019→0.017; ML-100K: 0.05→0.02) while showing a small decrease in MAE and NDCG scores.

4.4.5 Visualization of Learned Representations

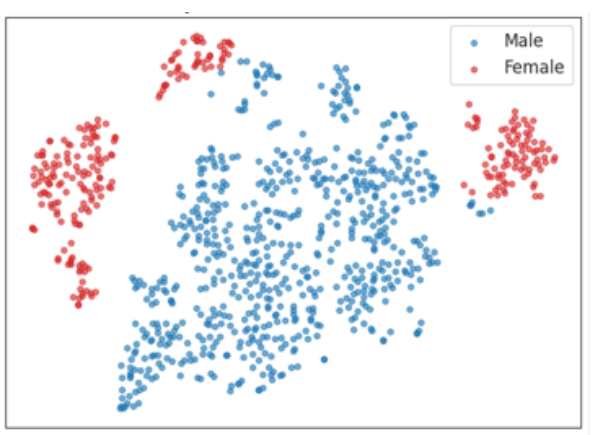
To provide a qualitative assessment of FUIREC's debiasing capability, we visualize the learned user and item embeddings using t-SNE visualization. Figure 4.6 compares the embedding distributions of



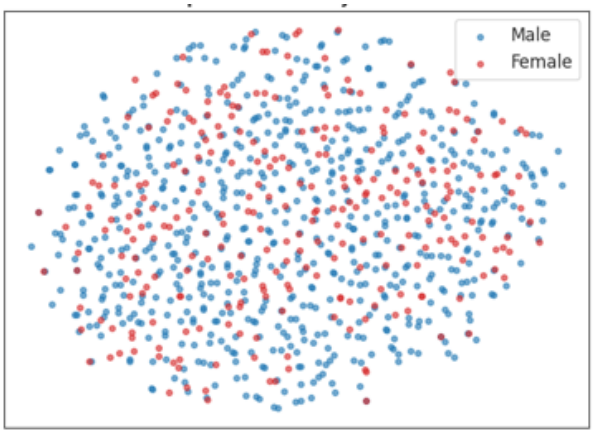
ML-100K BookCrossing ML-1M Dataset

Figure 4.5: The effect of the individual fairness constraint across three benchmark datasets.

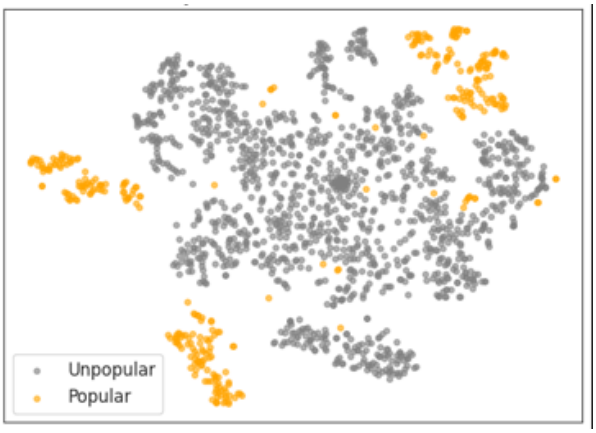
a standard backbone model (without fairness constraints) against our proposed FUIREC framework.



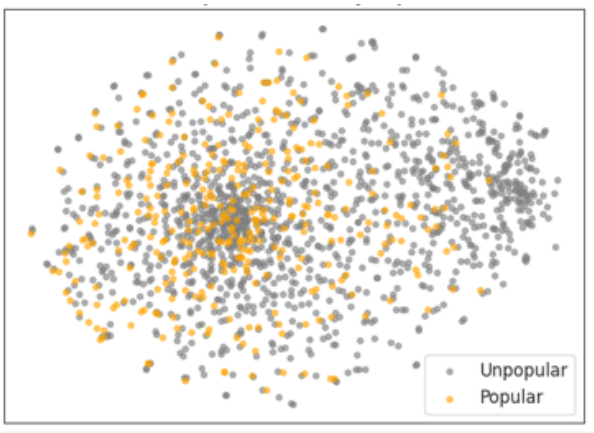
User Embeddings (Base Model)



User Embeddings (FUIREC)



Item Embeddings (Base Model)



Item Embeddings (FUIREC)

Figure 4.6: **t-SNE visualization of learned representations using the ML100K dataset.**

User Representation Analysis: As observed in Figure 4.6(a), the standard model exhibits

a strong correlation between latent representations and sensitive attributes, resulting in distinct clusters for Male and Female users. This separation allows downstream recommendation layers to easily exploit gender information, leading to biased outcomes. In contrast, Figure 4.6(b) shows that FUIREC effectively removes this structure. The user embeddings are uniformly mixed and entangled, making it difficult for any classifier to distinguish gender based on position in the latent space. This visual evidence aligns with the low AUC scores reported in Table 3, confirming that our user-gender discriminator successfully enforces gender-invariance.

Item Representation Analysis: A similar pattern is observed in the item space. Figure 4.6(c) reveals that the baseline model naturally clusters items by interaction volume, separating "Popular" items from "Unpopular" ones. This distinct separation explains the high Exposure Ratio observed in baseline models, as the system learns to prioritize the popular cluster. However, in Figure 4.6(d), FUIREC successfully disrupts this popularity-based clustering. Popular and unpopular items are projected into a shared semantic space, ensuring that the model relies on item features and user preferences rather than interaction volume. This qualitative analysis confirms that FUIREC effectively mitigates the "Matthew effect" at the representation level.

4.4.6 Hyperparameter Sensitivity and Trade-off Analysis

To determine the optimal configuration for the FUIREC framework, we conducted a comprehensive sensitivity analysis of the four key hyperparameters: the user-side weights (λ_g, λ_{ind}) and the item-side weights (λ_p, λ_{pop}). Figure 4.7 presents the 3D surface plots visualizing the impact of these parameters on both recommendation utility and fairness metrics . The results reveal a significant tension between minimizing error and maximizing fairness, requiring a careful balance to avoid degrading system utility.

When we set all lambda values to 1.0, the model focuses almost entirely on fairness. As shown in the plots, this gives the best fairness results (lowest GF and AUC scores). However, the recommendation accuracy gets much worse. The error (MAE) rises above 0.92, and the ranking score (NDCG) drops below 0.70.

On the other hand, when we use very small values (around 0.001), the model behaves like a standard recommender system. In this case, we get the best accuracy (lowest MAE and highest NDCG). However, the fairness scores are poor. This indicates that the model is giving accurate predictions but is failing to remove gender bias or popularity bias.

The experiments show that the best results come from using values between 0.01 and 0.1. This range provides the best trade-off where we can mitigate the bias without hurting the accuracy too much. Consequently, for the final FUIREC model, we adopted a configuration where the adversarial weights are set to $\lambda_g = 0.1$ and $\lambda_p = 0.1$, and the explicit constraint weights are set to $\lambda_{ind} = 0.01$and $\lambda_{pop} = 0.01$. This specific combination yields the most robust performance, maintaining high recommendation accuracy while ensuring that both user and item fairness metrics

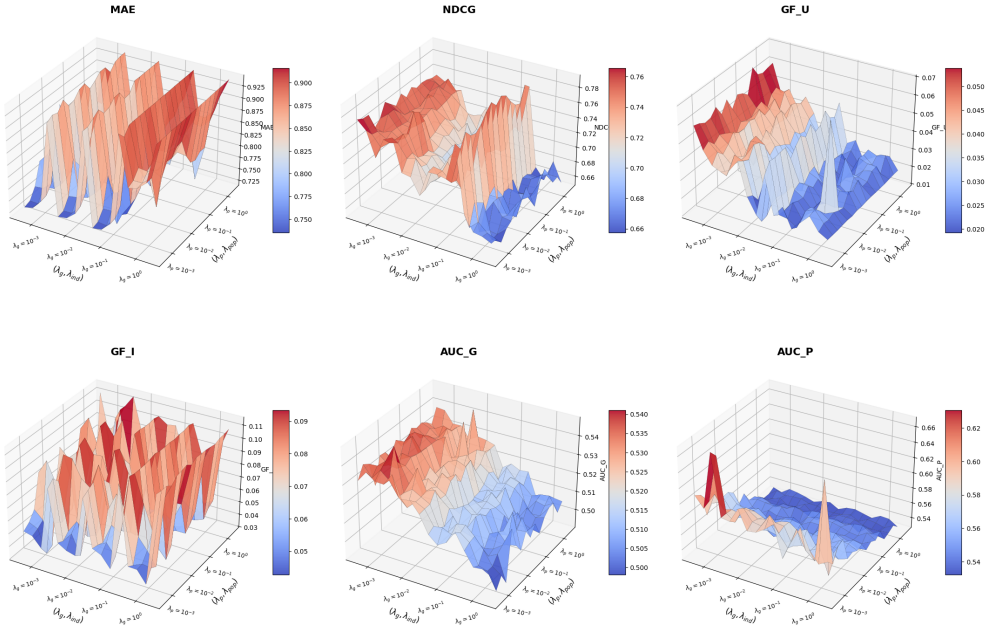


Figure 4.7: Hyperparameter sensitivity analysis.

remain within an optimal range.

4.5 Summary

In summary, this chapter presented FUIREC, a multi-stakeholder recommendation framework designed to mitigate user-side gender bias and item-side popularity bias. By combining dual adversarial discriminators with explicit output-level constraints, FUIREC effectively reduces bias at both the representation and ranking layers. Extensive experiments on three real-world datasets demonstrate that this hybrid approach achieves strong fairness guarantees for both users and items while maintaining high recommendation accuracy.

However, while output-level constraints successfully reduce downstream disparities, they do not enforce geometric separation within the latent space itself. This can leave the underlying embeddings vulnerable to non-linear attribute inference attacks. To address this limitation, the next chapter introduces ORAF (Orthogonal Regularization-based Adversarial Learning Framework). Chapter 5 advances this research by moving the entire intervention into the representation layer, replacing output-centric constraints with a strict orthogonal regularization penalty that removes bias directly inside the embedding space.

Chapter 5

ORAF: An Orthogonal Regularization-Based Adversarial Framework for Achieving Two-Sided Fairness in Recommender Systems

This chapter presents ORAF (Orthogonal Regularization-based Adversarial Learning Framework), an orthogonal framework that mitigates two-sided bias in recommender systems using dual adversarial discriminators and orthogonal regularization. ORAF combines a gender discriminator that removes sensitive attributes from user embeddings and a popularity discriminator that eliminates popularity signals from item representations, together with orthogonal constraints that enforce strict geometric perpendicularity between bias-aware and debiased embedding subspaces, anchored by dedicated attribute predictors. This hybrid adversarial-orthogonal approach transforms fairness from a statistical output-level correction into a deterministic geometric property of the latent space, providing a verifiable guarantee that adversarial training alone cannot deliver. Extensive experiments across three benchmark datasets and nine recommendation architectures demonstrate that ORAF significantly outperforms state-of-the-art baselines in both user-side and item-side fairness while maintaining competitive recommendation quality, with particular effectiveness under sparse data conditions and against non-linear attribute inference attacks.

5.1 Introduction

Building upon the multi-stakeholder foundations and dual-sided optimization challenges established in Chapter 4, this chapter introduces ORAF, which serves as an advanced architectural alternative and geometric enhancement to the FUIREC framework. While FUIREC successfully addresses multi-stakeholder equity by appending explicit, deterministic constraints at the final recommenda-

tion output, it leaves open the challenge of achieving complete, unentangled fairness purely within the latent space. Specifically, FUIREC's explicit constraints operate at the output level and do not enforce a geometric separation between biased and debiased representations. ORAF directly addresses this limitation by **replacing** these explicit output-level constraints with an orthogonal regularization mechanism that operates endogenously within the latent space. Thus, ORAF represents the natural completion of this research trajectory, shifting the operational mechanics entirely to an endogenous, representation-level intervention that purges bias structurally before downstream ranking occurs.

To address these critical gaps, we introduce ORAF (Orthogonal Regularization-based Adversarial Learning Framework), a novel framework that simultaneously tackles user-side demographic biases and item-side popularity biases through an integrated adversarial-orthogonal architecture. Unlike existing approaches that address these biases separately or rely on partial solutions, ORAF innovatively combines: (1) dual adversarial discriminators; a gender discriminator that adversarially removes gender-sensitive attributes from user embeddings, ensuring recommendations are invariant to the user's gender, and a popularity discriminator that eliminates popularity signals from item representations, decoupling relevance from historical popularity patterns; and (2) orthogonal constraints that enforce geometric separation between bias-explicit embeddings (learned by dedicated attribute predictors to capture sensitive signals) and debiased embeddings (learned through the adversarial process), promoting feature disentanglement.

To evaluate ORAF's effectiveness, we conducted comprehensive experiments across three real-world recommendation scenarios. Our results demonstrate that ORAF significantly outperforms existing methods in both user-side fairness (gender bias mitigation) and item-side fairness (popularity bias reduction), while maintaining competitive recommendation quality. Crucially, the model shows substantially reduced predictive accuracy when attempting to infer gender from debiased user embeddings or item popularity from debiased item embeddings, confirming effective bias removal. Further, ORAF exhibits robust counterfactual stability, where recommendations remain stable when synthetically altering user gender or item popularity. The framework also ensures strong group fairness, achieving near-identical performance between male and female users and between popular and unpopular items across all evaluated fairness metrics. Finally, ORAF preserves high recommendation accuracy, with consistently strong NDCG and MAE scores comparable to unbiased state-of-the-art recommenders.

In summary, our key contributions are:

- We propose ORAF, a novel unified framework that simultaneously addresses user gender bias and item popularity bias to ensure two-sided fairness through a hybrid adversarial-orthogonal approach.
- Our framework employs dual adversarial discriminators; a gender discriminator and a popu-

larity discriminator, that work in tandem to remove sensitive attributes from both user and item embeddings.

- We introduce orthogonal regularization that enforces strict geometric orthogonality between bias-explicit embeddings (learned by dedicated attribute predictors to capture sensitive signals) and debiased embeddings (learned through adversarial training), promoting complete feature disentanglement and complementing adversarial learning.
- We conduct extensive experiments on three real-world datasets, demonstrating that ORAF significantly outperforms state-of-the-art baselines in both user-side and item-side fairness while maintaining competitive recommendation quality.

5.2 The Proposed ORAF Framework

The core innovation of ORAF lies in its dual-component design. It combines a generator network responsible for learning recommendation representations with a debiasing module. The debiasing module simultaneously removes unwanted biases from both user and item representations. Before presenting the formal components, we provide a brief conceptual account of how the four building blocks of the framework interact. The aim is to make the overall dynamics accessible before introducing the mathematical details.

At a high level, ORAF involves four interacting components whose roles are designed to be complementary rather than redundant. The generator is the core recommendation model. Its objective is to produce debiased user and item embeddings from which accurate predictions can be made. At the same time, these embeddings must not reveal sensitive attributes.

The discriminators act as adversaries to the generator. The gender discriminator attempts to infer gender from the debiased user embeddings. The popularity discriminator attempts to infer item popularity from the debiased item embeddings. The generator is trained to confuse both discriminators through gradient reversal. This pushes the debiased embeddings toward distributional indistinguishability across sensitive attribute groups.

The predictors work in the complementary direction. They are trained to concentrate sensitive information into the bias-aware embeddings. This ensures that these representations become reliable encodings of the sensitive attributes. It also allows them to serve as semantically meaningful anchors for the orthogonal regularization. Without this supervised enrichment, the bias-aware subspaces would carry no structured sensitive signal. In that case, the orthogonal regularization would be enforced against an arbitrary direction in embedding space rather than against a meaningful bias direction.

Finally, the orthogonal regularization uses the bias-aware embeddings as a reference. It minimizes the inner product between the bias-aware and debiased embedding matrices. This enforces

geometric separation between the two subspaces, so that the debiased representations carry no component in the direction of the sensitive attribute signals concentrated in the bias-aware subspace. The key insight is that the adversarial and orthogonal components address different failure modes. The discriminators target distributional separability. The orthogonal regularization targets residual geometric correlation that the discriminators may fail to capture.

At the heart of the architecture is a dual embedding scheme. It maintains separate representation spaces for bias-related and bias-free features. Each user and item is represented by two distinct embeddings: bias-aware embeddings (u_b, i_b) that concentrate sensitive attribute information, and debiased embeddings (u_d, i_d) that are purified of such signals. The key components of the architecture are:

- **Generator**: the core recommendation model, which produces predictions for user-item interactions using the debiased embeddings. Detailed in Section 5.2.1.
- **Discriminators**: neural networks trained to predict sensitive attributes from the debiased embeddings. They consist of a gender discriminator operating on u_d and a popularity discriminator operating on i_d. Detailed in Section 5.2.2.
- **Predictors**: neural networks trained to predict sensitive attributes from the bias-aware embeddings. They consist of a gender predictor operating on u_b and a popularity predictor operating on i_b. Together, they ensure that the bias-aware subspaces are semantically meaningful anchors for the orthogonal regularization. Detailed in Section 5.2.3.
- **Orthogonal Regularization**: enforces a perpendicular configuration between the bias-aware subspace (u_b, i_b) and the debiased subspace (u_d, i_d). This geometrically decouples the sensitive attribute signals concentrated in the bias-aware embeddings from the representations used for recommendation. It thereby complements the distributional constraints imposed by the adversarial discriminators. Detailed in Section 5.2.4.

The notation adopted throughout this chapter is summarized in Table 5.1.

5.2.1 Recommender Module

The goal of this module is to generate high-quality debiased user and item embeddings ($\mathbf{u}^d$, $\mathbf{i}^d$). These embeddings must effectively capture user preferences and item characteristics while ensuring optimal recommendation quality. They provide personalized recommendations without incorporating biases related to sensitive attributes.

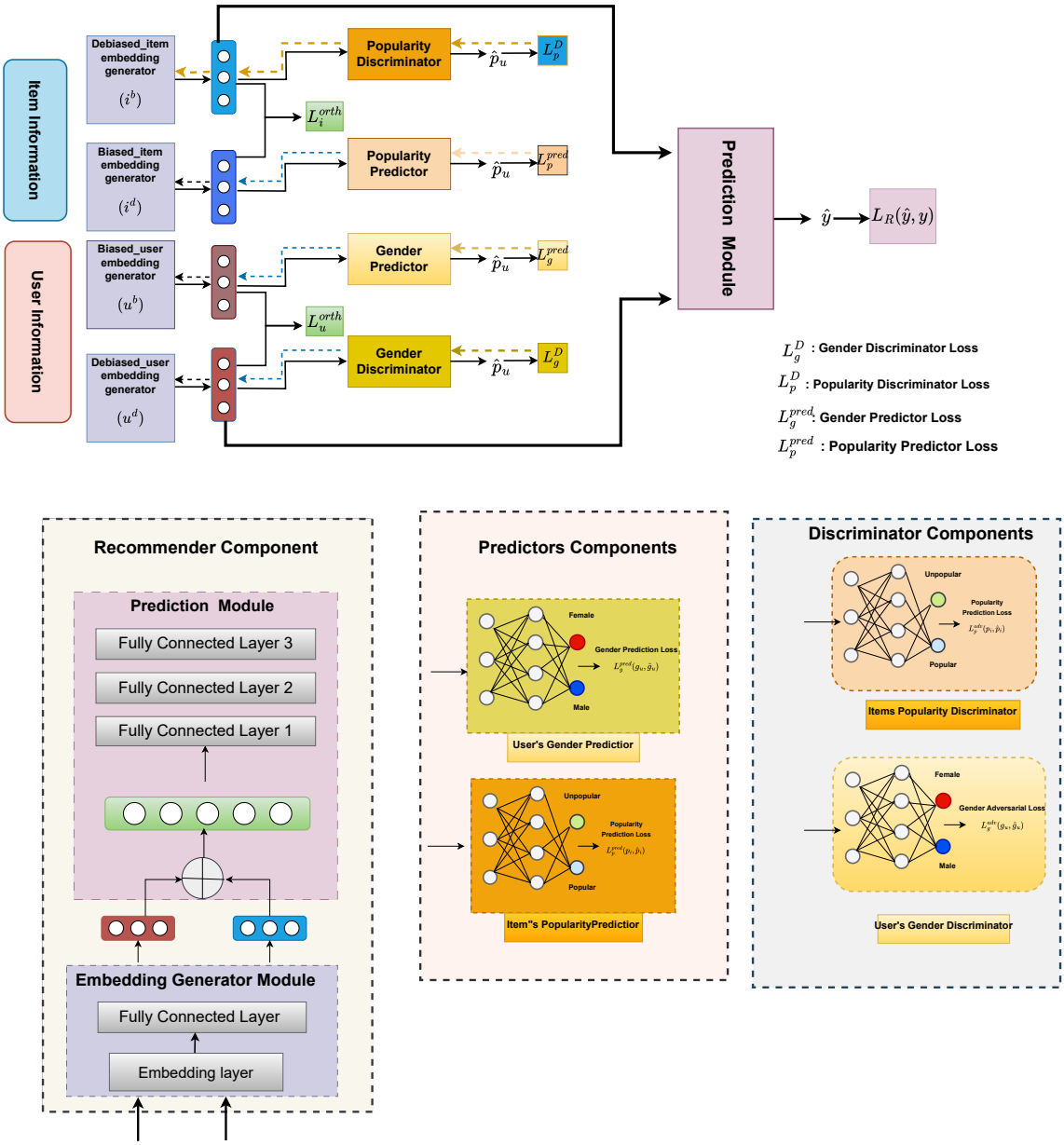


Figure 5.1: Overview of ORAF Architecture

Debiased Embedding Encoder

The goal of this component is to learn debiased user and item embeddings that are invariant to the sensitive attributes S_u and P_v. This is achieved by removing gender-related information from the user embeddings $\mathbf{u}^d$ and popularity signals from the item embeddings $\mathbf{i}^d$. These embeddings are trained adversarially to eliminate attribute-related signals. In addition, they are enforced to remain

Table 5.1: Notation used in the ORAF framework.

Symbol	Meaning
	Sets and indices
U, V	Set of users; set of items
u, v	A specific user; a specific item
$\mathcal{D}$	Set of training user-item interaction pairs
$y_{ui} \in \{0, 1\}$	Observed interaction label between user u and item i
$\hat{r}_{ui} \in [0, 1]$	Predicted interaction probability
	Sensitive attributes and labels
S_u	Sensitive attribute of user u (gender)
P_v	Sensitive attribute of item v (popularity)
$g_u \in \{0, 1\}$	Ground-truth gender label for user u
$p_v \in \{0, 1\}$	Ground-truth popularity label for item v
$\hat{g}_u$, $\hat{p}_v$	Predicted gender and popularity probabilities
	Embeddings (vectors and matrices)
$\mathbf{u}^d$, $\mathbf{i}^d$	Debiased user and item embeddings
$\mathbf{u}^b$, $\mathbf{i}^b$	Bias-aware user and item embeddings
$\mathbf{U}_d, \mathbf{U}_b \in \mathbb{R}^{nk}$	Debiased / bias-aware user embedding matrices (mini-batch)
$\mathbf{I}_d, \mathbf{I}_b \in \mathbb{R}^{nk}$	Debiased / bias-aware item embedding matrices (mini-batch)
$\mathbf{x}_u$, $\mathbf{x}_i$	Raw user and item feature vectors
$E_u(\cdot)$, $E_i(\cdot)$	User / item embedding layers
	Network parameters
Θ_G	Parameters of the generator (recommender)
Θ_{D_g}, Θ_{D_p}	Parameters of the gender / popularity discriminators
Θ_{P_g}, Θ_{P_p}	Parameters of the gender / popularity predictors
$\mathbf{W}_u^d, \mathbf{W}_i^d$	Weight matrices of the debiased encoders
$\mathbf{W}_u^b, \mathbf{W}_i^b$	Weight matrices of the bias-aware encoders
$\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$	Weight matrices of the prediction MLP
$\mathbf{b}^{(\cdot)}$, $\mathbf{c}^{(\cdot)}$	Bias vectors (subscripts identify the layer)
$\sigma(\cdot)$, ReLU$(\cdot)$	Sigmoid and ReLU activation functions
	Loss functions
$\mathcal{L}_{\text{rec}}$	Recommendation loss (binary cross-entropy)
$\mathcal{L}_g^{\text{adv}}$, $\mathcal{L}_p^{\text{adv}}$	Adversarial losses of the gender / popularity discriminators
$\mathcal{L}_g^{\text{pred}}$, $\mathcal{L}_p^{\text{pred}}$	Predictor losses for gender / popularity
$\mathcal{L}_{\text{orth}}^u$, $\mathcal{L}_{\text{orth}}^i$	Orthogonal regularization losses (user / item side)
$\mathcal{L}_{\text{total}}$	Total composite training objective
	Hyperparameters
λ_g, λ_p	Weights of the adversarial losses
λ_g^{pred}, λ_p^{pred}	Weights of the predictor losses
λ_{orth}^u, λ_{orth}^i	Weights of the orthogonal regularization losses
η_G, η_D, η_P	Learning rates of generator / discriminators / predictors
k	Number of discriminator updates per generator update
n	Mini-batch size
d	Embedding dimensionality

orthogonal to their bias-explicit counterparts ($\mathbf{u}^b$, $\mathbf{i}^b$). This debiasing process, detailed in Section 5.2.2, aims to reduce the impact of sensitive attributes and create embeddings that are independent of these biases.

The architecture consists of an embedding layer followed by a fully connected layer. User features (e.g., age, gender, occupation) and item features (e.g., genre, release year, popularity) are processed through these layers to generate the respective embeddings as follows:

$$\mathbf{u}^d = \text{ReLU}\left(\mathbf{W}_u^d \cdot E_u(\mathbf{x}_u) + \mathbf{b}_u^d\right) \tag{5.1}$$

$$\mathbf{i}^d = \text{ReLU}\left(\mathbf{W}_i^d \cdot E_i(\mathbf{x}_i) + \mathbf{b}_i^d\right) \tag{5.2}$$

Where:

- x_u, x_i are the user feature vector and item feature vector, respectively.
- $E_u(x_u), E_i(x_i)$ are the embedding layers that transform user and item features into latent vectors.
- W_u^d, W_i^d are the weight matrices for the fully connected layers applied to user and item embeddings.
- b_u^d, b_i^d are the bias terms for the fully connected layers.
- ReLU is the ReLU activation function, applied element-wise.

Bias-Aware Embedding Encoder: This component explicitly captures and isolates bias patterns associated with sensitive attributes in the embedding space. For users, the bias-aware embedding $\mathbf{u}^b$ encodes information correlated with protected characteristics such as gender. For items, $\mathbf{i}^b$ captures popularity-related patterns. These embeddings are optimized through supervised prediction tasks. In these tasks, separate predictors (detailed in Section 5.2.3) learn to accurately reconstruct the sensitive attributes from the embeddings. This explicit modeling concentrates bias-related signals within dedicated embedding subspaces. As a result, it prevents their propagation into the final recommendations.

The architecture mirrors that of the debiased encoder. It employs the same feature processing pipeline but optimizes for opposite objectives:

$$\mathbf{u}^b = \text{ReLU}\left(\mathbf{W}_u^b \cdot E_u(\mathbf{x}_u) + \mathbf{b}_u^b\right), \tag{5.3}$$

$$\mathbf{i}^b = \text{ReLU}\left(\mathbf{W}_i^b \cdot E_i(\mathbf{x}_i) + \mathbf{b}_i^b\right). \tag{5.4}$$

This symmetrical design ensures computational consistency. At the same time, it maintains the semantic separation between bias-aware and bias-free representations. The resulting embeddings serve two crucial purposes. First, they provide targets for orthogonal regularization, ensuring complete geometric separation between biased and debiased representations. Second, they enable quantitative measurement of the bias information captured within the model's representations.

Prediction Module

The prediction module synthesizes the debiased user and item representations to generate final recommendation scores. The debiased user embedding $\mathbf{u}^d$ and debiased item embedding $\mathbf{i}^d$ are first concatenated to form a unified interaction representation:

$$\mathbf{z} = [\mathbf{u}^d; \mathbf{i}^d] \tag{5.5}$$

Here, $\mathbf{z} \in \mathbb{R}^{2d}$ represents the combined feature vector. This vector encapsulates the user-item interaction patterns in the debiased embedding space.

This concatenated representation is then processed through a multi-layer perceptron (MLP). The MLP captures complex non-linear interactions between user and item features. Its architecture consists of:

$$\mathbf{h}_1 = \mathrm{ReLU}\left(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1\right), \quad \mathbf{h}_2 = \mathrm{ReLU}\left(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2\right), \quad \hat{r}_{ui} = \sigma\left(\mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3\right) \tag{5.6}$$

where:

- $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{R}^d$ are hidden layer activations,
- $\mathbf{W}_1 \in \mathbb{R}^{d2d}$, $\mathbf{W}_2 \in \mathbb{R}^{dd}$, $\mathbf{W}_3 \in \mathbb{R}^{1d}$ are weight matrices,
- $\mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^d$, $\mathbf{b}_3 \in \mathbb{R}$ are bias vectors,
- $\sigma(\cdot)$ denotes the sigmoid activation function, mapping the output to $[0, 1]$.

The final output $\hat{r}_{ui} \in [0, 1]$ represents the predicted probability of interaction between user u and item i.

The model is optimized using binary cross-entropy loss. This loss minimizes the discrepancy between predictions and observed interactions:

$$\mathcal{L}_{\text{rec}} = -\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left[y_{ui} \log(\hat{r}_{ui}) + (1 - y_{ui}) \log(1 - \hat{r}_{ui})\right] \tag{5.7}$$

where:

- $\mathcal{D}$ is the set of training user-item pairs,
- $y_{ui} \in \{0, 1\}$ is the ground-truth interaction label,
- $\hat{r}_{ui}$ is the predicted interaction probability.

This optimization ensures that the model accurately reconstructs user preferences from the debiased representations. At the same time, it maintains the fairness properties encoded in $\mathbf{u}^d$ and $\mathbf{i}^d$.

5.2.2 Adversarial Discriminators

To achieve attribute-invariant representations while preserving recommendation utility, we implement an adversarial learning framework with dual discriminators. These discriminators operate on the debiased embeddings. The framework establishes a min-max game between the generator G, which produces debiased embeddings ($\mathbf{u}^d$, $\mathbf{i}^d$), and the discriminators D_g and D_p, which attempt to recover sensitive attributes from these representations.

This formulation ensures two complementary outcomes. First, the generator learns to produce informative embeddings for recommendation tasks. Second, it simultaneously eliminates information about protected attributes through adversarial training.

Gender Discriminator

The gender discriminator D_g serves as a critical component for removing gender-related information from user embeddings. It is implemented as a multi-layer perceptron with dimension-reducing layers. The discriminator operates exclusively on the debiased user embeddings $\mathbf{u}^d$ to predict gender probabilities:

$$\mathbf{h}_g^{(1)} = \text{ReLU}(\mathbf{W}_{g1}\mathbf{u}^d + \mathbf{b}_{g1}) \tag{5.8}$$

$$\mathbf{h}_g^{(2)} = \text{ReLU}(\mathbf{W}_{g2}\mathbf{h}_g^{(1)} + \mathbf{b}_{g2}) \tag{5.9}$$

$$\hat{g}_u = \sigma(\mathbf{W}_{g3}\mathbf{h}_g^{(2)} + \mathbf{b}_{g3}) \tag{5.10}$$

where:

- $\mathbf{W}_{g1} \in \mathbb{R}^{dd/2}$, $\mathbf{W}_{g2} \in \mathbb{R}^{d/2d/4}$, $\mathbf{W}_{g3} \in \mathbb{R}^{d/41}$ are learnable weight matrices
- $\mathbf{b}_{g1} \in \mathbb{R}^{d/2}$, $\mathbf{b}_{g2} \in \mathbb{R}^{d/4}$, $\mathbf{b}_{g3} \in \mathbb{R}$ are bias terms
- $\hat{g}_u \in [0, 1]$ represents the probability of user u being female

The discriminator is trained to minimize the binary cross-entropy loss:

$$\mathcal{L}_g^{\text{adv}} = -\frac{1}{|U|}\sum_{u=1}^{|U|}[g_u \log(\hat{g}_u) + (1 - g_u)\log(1 - \hat{g}_u)] \tag{5.11}$$

To prevent the gender discriminator from inferring user attributes from the bias-free user embedding, we employ gradient reversal. Specifically, we use the negative gradients of the discriminator to penalize the model. This approach ensures that the generator learns to produce embeddings that confuse the discriminator. As a result, sensitive attribute information is removed from the learned representations.

Popularity Discriminator

The popularity discriminator D_p follows a symmetric architecture to remove popularity bias from item embeddings. It processes debiased item embeddings $\mathbf{i}^d$ to predict popularity scores:

$$\mathbf{h}_p^{(1)} = \text{ReLU}(\mathbf{W}_{p1}\mathbf{i}^d + \mathbf{b}_{p1}) \tag{5.12}$$

$$\mathbf{h}_p^{(2)} = \text{ReLU}(\mathbf{W}_{p2}\mathbf{h}_p^{(1)} + \mathbf{b}_{p2}) \tag{5.13}$$

$$\hat{p}_v = \sigma(\mathbf{W}_{p3}\mathbf{h}_p^{(2)} + \mathbf{b}_{p3}) \tag{5.14}$$

with parameters:

- $\mathbf{W}_{p1} \in \mathbb{R}^{dd/2}$, $\mathbf{W}_{p2} \in \mathbb{R}^{d/2d/4}$, $\mathbf{W}_{p3} \in \mathbb{R}^{d/41}$
- $\mathbf{b}_{p1} \in \mathbb{R}^{d/2}$, $\mathbf{b}_{p2} \in \mathbb{R}^{d/4}$, $\mathbf{b}_{p3} \in \mathbb{R}$

The popularity adversarial loss is given by:

$$\mathcal{L}_p^{\text{adv}} = -\frac{1}{|V|}\sum_{v=1}^{|V|}[p_v \log(\hat{p}_v) + (1 - p_v)\log(1 - \hat{p}_v)] \tag{5.15}$$

We use gradient reversal to prevent popularity leakage from debiased item embeddings. Specifically, negative discriminator gradients are used to train the generator against popularity prediction. This forces the generator to learn popularity-invariant representations. These representations maintain recommendation quality while confusing the popularity discriminator.

5.2.3 Attribute Predictors for Bias-Explicit Embeddings

The attribute prediction modules constitute a fundamental component of our fairness-aware recommendation framework. They serve to explicitly identify and isolate bias patterns within user and item representations. These supervised learning components are designed to extract protected attribute signals. They confine these signals to dedicated bias-explicit embeddings ($\mathbf{u}^b$ for users and $\mathbf{i}^b$ for items). By maximizing the bias information in these explicit embeddings, we create a bias proxy. This ensures that when we apply orthogonal regularization (Section 5.2.4), we are distancing the debiased embeddings from a representation that is known to capture the bias effectively.

Gender Predictor

The gender prediction component operates as a specialized neural network. Its role is to explicitly encode gender-related patterns within the bias-explicit user embeddings. The module is implemented as a three-layer perceptron. It processes bias-explicit user embeddings u^b through the following transformation:

$$h_1 = \text{ReLU}(W_1 u^b + b_1) \tag{5.16}$$

$$h_2 = \text{ReLU}(W_2 h_1 + b_2) \tag{5.17}$$

$$\hat{g}_i = \sigma(W_3 h_2 + b_3) \tag{5.18}$$

where W_1, W_2, W_3 represent weight matrices, b_1, b_2, b_3 are bias vectors, and σ denotes the sigmoid activation function. The model is optimized using the binary cross-entropy loss function:

$$\mathcal{L}_g^{\text{pred}} = -\frac{1}{|U|} \sum_{u \in U} [g_u \log \hat{g}_u + (1 - g_u) \log(1 - \hat{g}_u)] \tag{5.19}$$

In this loss, $g_u \in \{0, 1\}$ represents the ground truth gender label for user u, and $\hat{g}_u$ denotes the predicted gender probability. This optimization objective forces u^b to maintain strong gender-discriminative features. At the same time, it enables the debiased embeddings u^d to remain gender-neutral through adversarial training and orthogonal regularization.

Popularity Predictor

An analogous architecture is used for item popularity prediction. It operates on the bias-explicit item embeddings i^b to capture popularity-related patterns that need to be separated from intrinsic item properties.

$$h_1 = \text{ReLU}(V_1 i^b + c_1) \tag{5.20}$$

$$h_2 = \text{ReLU}(V_2 h_1 + c_2) \tag{5.21}$$

$$\hat{p}_i = \sigma(V_3 h_2 + c_3) \tag{5.22}$$

with corresponding parameters V_1, V_2, V_3 (weight matrices) and c_1, c_2, c_3 (bias vectors). The model is optimized via binary cross-entropy:

$$\mathcal{L}_p^{\text{pred}} = -\frac{1}{|I|} \sum_{i \in I} [p_i \log \hat{p}_i + (1 - p_i) \log(1 - \hat{p}_i)] \tag{5.23}$$

Here, $p_i \in \{0, 1\}$ represents the ground truth popularity label for item i, and $\hat{p}_i$ denotes the predicted popularity probability.

By minimizing this prediction error, we force $\mathbf{i}^b$ to become a highly accurate bias proxy for popularity. This ensures that when the orthogonal regularization minimizes the correlation between $\mathbf{i}^b$ and $\mathbf{i}^d$, it is effectively stripping popularity information from the debiased item representation. What remains is only the intrinsic content features necessary for recommendation.

5.2.4 Orthogonal Regularization

While adversarial training effectively mitigates direct bias, the minimax optimization inherent in adversarial frameworks can converge to local equilibria where sensitive information is merely obscured rather than fully removed from the learned embeddings. To address this limitation and eliminate residual bias in the debiased embeddings $\mathbf{u}_d$, we introduce orthogonal regularization as a complementary geometric constraint.

The core idea is to enforce an orthogonal decomposition of the embedding space for both users and items. We define the required orthogonality conditions as:

$$\mathbf{u}_b^\top \mathbf{u}_d = 0 \tag{5.24}$$

$$\mathbf{i}_b^\top \mathbf{i}_d = 0 \tag{5.25}$$

Here, $\mathbf{u}_b$ and $\mathbf{i}_b$ serve as the bias proxies, optimized by the attribute predictors in Section 5.2.3 to concentrate gender and popularity signals respectively. By enforcing orthogonality between $\mathbf{u}_b$ and $\mathbf{u}_d$, and analogously between $\mathbf{i}_b$ and $\mathbf{i}_d$, we ensure that the debiased representations lie in a subspace that is geometrically orthogonal to the bias-explicit subspace. The directions along which the bias proxy embeddings encode sensitive attribute signals therefore have zero projection onto the debiased embeddings.

To operationalise this constraint during training, we define the orthogonal regularization losses for users and items as the squared Frobenius norm of the cross-Gram matrices of the respective embedding matrices:

$$\mathcal{L}_{\text{orth}}^u = \frac{1}{n^2} \left\| \mathbf{U}_d^\top \mathbf{U}_b \right\|_F^2 \tag{5.26}$$

$$\mathcal{L}_{\text{orth}}^i = \frac{1}{n^2} \left\| \mathbf{I}_d^\top \mathbf{I}_b \right\|_F^2 \tag{5.27}$$

where $\mathbf{U}_b, \mathbf{U}_d \in \mathbb{R}^{nk}$ are the matrices of bias-aware and debiased user embeddings for a mini-batch of n users, $\mathbf{I}_b, \mathbf{I}_d \in \mathbb{R}^{nk}$ are the corresponding item embedding matrices, and $\|\cdot\|_F$ denotes the Frobenius norm. Minimizing these losses drives $\mathcal{L}_{\text{orth}}^u \to 0$ and $\mathcal{L}_{\text{orth}}^i \to 0$, progressively pushing the column spaces of $\mathbf{U}_d$ and $\mathbf{U}_b$ toward mutual orthogonality. When this constraint is fully satisfied, the two subspaces are orthogonal complements in the ambient space: any sensitive attribute signal encoded along the directions of the bias-aware subspace carries zero projection onto the debiased subspace and cannot be recovered from it by any operation confined to the span of the bias-aware embeddings.

This geometric property is deterministic and verifiable, since $\mathcal{L}_{\text{orth}}^u$ and $\mathcal{L}_{\text{orth}}^i$ are closed-form, differentiable quantities that can be monitored at every training step. This stands in contrast to

adversarial indistinguishability, whose effectiveness depends on discriminator capacity and training dynamics. The orthogonal regularizer therefore acts as a secondary safeguard, complementing adversarial training by enforcing a structural geometric separation between the two embedding subspaces that adversarial optimization alone cannot achieve.

5.3 ORAF Training

5.3.1 Theoretical Framework and Objective

The training paradigm for ORAF is founded on a minimax game. This game is played between the generator component, which produces debiased embeddings, and the adversarial discriminators. The complete objective function integrates four terms: recommendation accuracy, adversarial fairness, attribute prediction, and orthogonal regularization.

$$\min_{\substack{\Theta_G,\\ \Theta_{P_g},\\ \Theta_{P_p}}} \max_{\substack{\Theta_{D_g},\\ \Theta_{D_p}}} \overbrace{\underbrace{\mathcal{L}_{\text{rec}}}_{0} - \underbrace{\lambda_g \mathcal{L}_g^{\text{adv}} - \lambda_p \mathcal{L}_p^{\text{adv}}}_{0} + \underbrace{\lambda_g^{\text{pred}} \mathcal{L}_g^{\text{pred}} + \lambda_p^{\text{pred}} \mathcal{L}_p^{\text{pred}}}_{0} + \underbrace{\lambda_{\text{orth}}^{u} \mathcal{L}_{\text{orth}}^{u} + \lambda_{\text{orth}}^{i} \mathcal{L}_{\text{orth}}^{i}}_{0}}^{\mathcal{L}_{\text{total}}} \tag{5.28}$$

Equation 5.28 captures the adversarial nature of the training. The discriminators attempt to maximize their predictive accuracy. The generator aims to minimize the ability of the discriminators to predict sensitive attributes. At the same time, the generator minimizes both the recommendation loss and the orthogonal loss.

Discriminator Update

The discriminators are updated to maximize their predictive accuracy on the debiased embeddings:

$$\max_{\Theta_{D_g},\Theta_{D_p}} \underbrace{\mathbb{E}_{u\sim p_{\text{data}}(u)}[\log D_g(u^d)] + \mathbb{E}_{u\sim p_G(u)}[\log(1 - D_g(G(u)))]}_{\text{Gender discriminator}} + \underbrace{\mathbb{E}_{v\sim p_{\text{data}}(v)}[\log D_p(i^d)] + \mathbb{E}_{v\sim p_G(v)}[\log(1 - D_p(G(v}_{\text{Popularity discriminator}} \tag{5.29}$$

The update rules implement gradient descent on the discriminator losses:

$$\Theta_{D_g} \leftarrow \Theta_{D_g} - \eta_D \cdot \nabla_{\Theta_{D_g}} \mathcal{L}_g^{\text{adv}} \tag{5.30}$$

$$\Theta_{D_p} \leftarrow \Theta_{D_p} - \eta_D \cdot \nabla_{\Theta_{D_p}} \mathcal{L}_p^{\text{adv}} \tag{5.31}$$

Predictor Updates

The attribute predictors are updated using gradient descent. Their role is to ensure that the bias-aware embeddings retain sensitive information:

$$\Theta_{P_g} \leftarrow \Theta_{P_g} - \eta_P \cdot \nabla_{\Theta_{P_g}} \mathcal{L}_g^{\text{pred}} \tag{5.32}$$

$$\Theta_{P_p} \leftarrow \Theta_{P_p} - \eta_P \cdot \nabla_{\Theta_{P_p}} \mathcal{L}_p^{\text{pred}} \tag{5.33}$$

Generator Update

The generator learns to produce debiased embeddings that satisfy three goals simultaneously. They must maintain recommendation quality, fool the discriminators, and satisfy the orthogonal regularizations. The generator parameters are updated by minimizing the recommendation loss, the predictor losses, and the orthogonal losses. At the same time, they are updated to maximize the discriminator loss through Gradient Reversal:

$$\min_{\Theta_G} \mathcal{L}_{\text{rec}} - \lambda_g \mathcal{L}_g^{\text{adv}} - \lambda_p \mathcal{L}_p^{\text{adv}} + \lambda_g^{\text{pred}} \mathcal{L}_g^{\text{pred}} + \lambda_p^{\text{pred}} \mathcal{L}_p^{\text{pred}} + \lambda_{\text{orth}}^u \mathcal{L}_{\text{orth}}^u + \lambda_{\text{orth}}^i \mathcal{L}_{\text{orth}}^i \tag{5.34}$$

$$\Theta_G \leftarrow \Theta_G - \eta_G \cdot \nabla_{\Theta_G} \mathcal{L}_{\text{total}} \tag{5.35}$$

which is equivalent to:

$$\Theta_G - \eta_G \cdot \left[\frac{\partial \mathcal{L}_{\text{rec}}}{\partial \Theta_G} + \lambda_g^{\text{pred}} \frac{\partial \mathcal{L}_g^{\text{pred}}}{\partial \Theta_G} + \lambda_p^{\text{pred}} \frac{\partial \mathcal{L}_p^{\text{pred}}}{\partial \Theta_G} + \lambda_{\text{orth}}^u \frac{\partial \mathcal{L}_{\text{orth}}^u}{\partial \Theta_G} + \lambda_{\text{orth}}^i \frac{\partial \mathcal{L}_{\text{orth}}^i}{\partial \Theta_G} - \lambda_g \frac{\partial \mathcal{L}_g^{\text{adv}}}{\partial \Theta_G} - \lambda_p \frac{\partial \mathcal{L}_p^{\text{adv}}}{\partial \Theta_G} \right] \tag{5.36}$$

Algorithm 3 outlines the complete training procedure. It implements this adversarial game through careful alternation between component updates. The algorithm maintains stability by updating the discriminators multiple times (3 steps) per generator update. This ensures that the discriminators remain competitive adversaries throughout training. The hyperparameters λ_g, λ_p, and λ_{orth} control the trade-off between the competing objectives. The learning rates η_G, η_D, and η_P are tuned to maintain training stability.

The training proceeds iteratively through alternating optimization phases. This alternation maintains stability between the competing objectives. For each mini-batch sampled from the training data, the algorithm executes three sequential phases.

In the first phase, the adversarial discriminators are strengthened through multiple update steps. The generator parameters remain frozen during this phase. The gender discriminator D_g is updated to maximize its ability to predict user gender from the debiased embeddings u^d. The popularity discriminator D_p is trained to predict item popularity from i^d. Both updates are performed using equations 5.30 and 5.31.

Concurrently, the attribute predictors are refined. Their role is to ensure that the bias-aware embeddings retain sensitive attribute information. The gender predictor updates its parameters to

Algorithm 3 ORAF Minimax Optimization Training

Require: Interaction history D, user gender labels g_u, item popularity labels p_v, hyperparameters $\lambda_g, \lambda_p, \lambda_g^{\text{pred}}, \lambda_p^{\text{pred}}, \lambda_{\text{orth}}^u, \lambda_{\text{orth}}^i$, learning rates η_G, η_D, η_P, adversary steps k

Ensure: Trained parameters Θ_G, Θ_{D_g}, Θ_{D_p}, Θ_{P_g}, Θ_{P_p}

Initialize parameters $\Theta_G, \Theta_{D_g}, \Theta_{D_p}, \Theta_{P_g}, \Theta_{P_p}$
while not converged **do**
  for each mini-batch 0 **do**
    // Phase 1: train discriminators
    for 0 to 1 **do**
      Compute $\mathcal{L}_g^{\text{adv}}$ using equation 5.11
      Compute $\mathcal{L}_p^{\text{adv}}$ using equation 5.15
      $\Theta_{D_g} \leftarrow \Theta_{D_g} - \eta_D \cdot \nabla_{\Theta_{D_g}} \mathcal{L}_g^{\text{adv}}$
      $\Theta_{D_p} \leftarrow \Theta_{D_p} - \eta_D \cdot \nabla_{\Theta_{D_p}} \mathcal{L}_p^{\text{adv}}$
    end for
    // Phase 2: Update Predictors
    Compute $\mathcal{L}_g^{\text{pred}}$ using equation 5.19
    Compute $\mathcal{L}_p^{\text{pred}}$ using equation 5.23
    $\Theta_{P_g} \leftarrow \Theta_{P_g} - \eta_P \cdot \nabla_{\Theta_{P_g}} \mathcal{L}_g^{\text{pred}}$
    $\Theta_{P_p} \leftarrow \Theta_{P_p} - \eta_P \cdot \nabla_{\Theta_{P_p}} \mathcal{L}_p^{\text{pred}}$
    // Phase 3: Update Generator
    Compute $\mathcal{L}_{\text{rec}}$ using equation 5.7
    Compute $\mathcal{L}_{\text{orth}}^u$ using equation 5.26
    Compute $\mathcal{L}_{\text{orth}}^i$ using equation 5.27
    Compute $\mathcal{L}_{\text{total}}$ using equation 5.28
    $\Theta_G \leftarrow \Theta_G - \eta_G \cdot \nabla_{\Theta_G} \mathcal{L}_{\text{total}}$
  end for
end while
return updated $\Theta_G, \Theta_{D_g}, \Theta_{D_p}, \Theta_{P_g}, \Theta_{P_p}$

minimize prediction error on u^b. The popularity predictor does the same for i^b. Both updates are performed using equations 5.32 and 5.33.

The second phase focuses on generator optimization. During this phase, all adversarial components are frozen. The generator computes the composite loss. This loss incorporates recommendation accuracy, adversarial fairness through gradient reversal, and orthogonal regularization. The orthogonal terms $\mathcal{L}_{\text{orth}}^u$ and $\mathcal{L}_{\text{orth}}^i$ are computed as the Frobenius norm of the inner products between bias-aware and debiased embedding matrices. This computation enforces geometric separation between the two subspaces. The generator update employs gradient descent on the total objective

using equation 5.36.

This alternating optimization continues until convergence. Convergence is characterized by four key conditions: (1) discriminator losses plateau near the theoretical maximum of random guessing, indicating successful bias removal; (2) recommendation loss stabilizes, confirming maintained utility; (3) orthogonal regularization approaches zero, verifying geometric separation; and (4) predictor losses remain low, ensuring that bias-aware embeddings retain attribute information. The resulting equilibrium represents a Pareto-optimal solution. At this solution, the generator produces embeddings that are simultaneously debiased and highly effective for recommendation tasks.

5.3.2 Adversarial Training Analysis

To empirically validate the adversarial framework, Figure 5.2 visualizes the joint evolution of adversarial losses, utility, and orthogonality throughout training. Two milestones dominate the analysis: a *transition point* around epoch 17, at which the Generator begins to successfully obfuscate the sensitive signals, and an *equilibrium point* around epoch 36, at which all components stabilize simultaneously. Before the transition point, sensitive attributes are readily recoverable from the

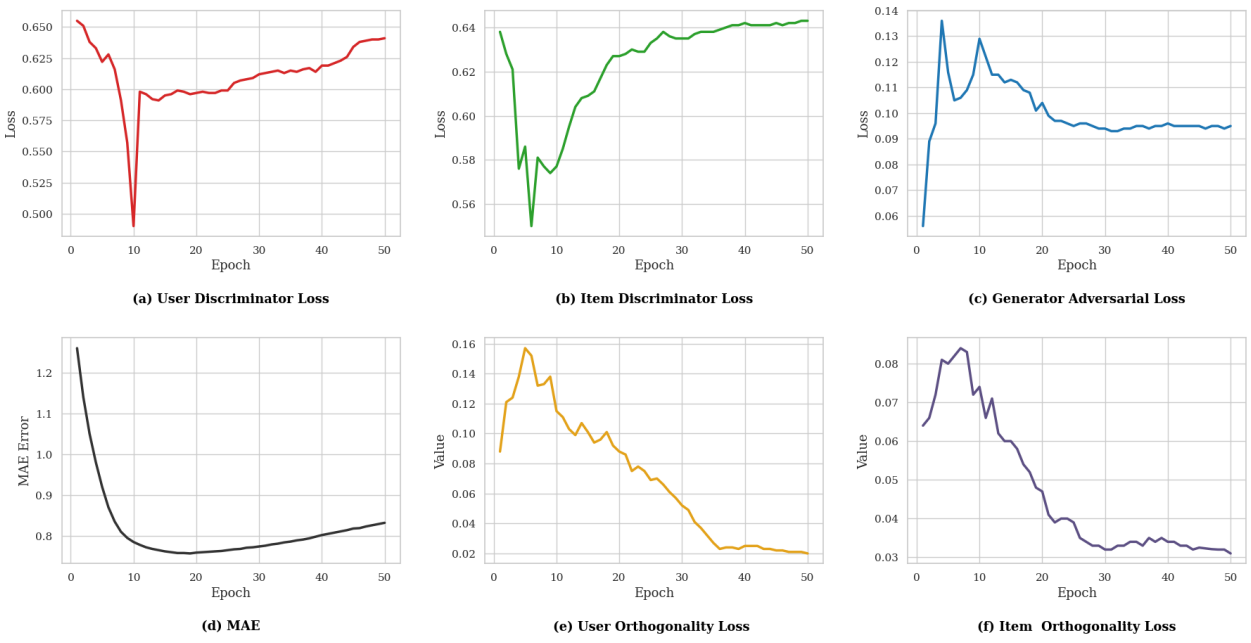


Figure 5.2: ORAF Training Process.

embeddings, as reflected by the low discriminator losses (≈ 0.48 for gender, ≈ 0.55 for popularity) and the high orthogonality losses (≈ 0.16). After epoch 17, the Generator progressively strips this information, causing the discriminator losses to rise toward the random-guessing level and the orthogonality losses to decay rapidly. By the equilibrium point, the discriminator losses plateau near ≈ 0.64 (indistinguishable from random guessing), the orthogonality losses flatten at negligible values

(≈ 0.02), and the utility stabilizes at MAE ≈ 0.83 after an initial minimum of ≈ 0.76 near epoch 18. The small utility gap between the unconstrained minimum and the equilibrium value reflects the expected fairness–accuracy trade-off, and the simultaneous stabilization of all six metrics confirms that the system reaches a Nash equilibrium without mode collapse.

5.3.3 Optimization Dynamics and Convergence Analysis

A valid concern in multi-objective adversarial frameworks is optimization instability, in particular mode collapse and destructive gradient conflict between competing objectives. To examine whether the composite ORAF objective in Equation 5.28 suffers from such issues, we track the L_2 gradient norms of the three main loss components, recommendation ($\mathcal{L}_{\text{rec}}$), adversarial fairness ($\mathcal{L}_{\text{adv}}$), and orthogonal regularization ($\mathcal{L}_{\text{orth}}$), across 50 training epochs (Figure 5.3).

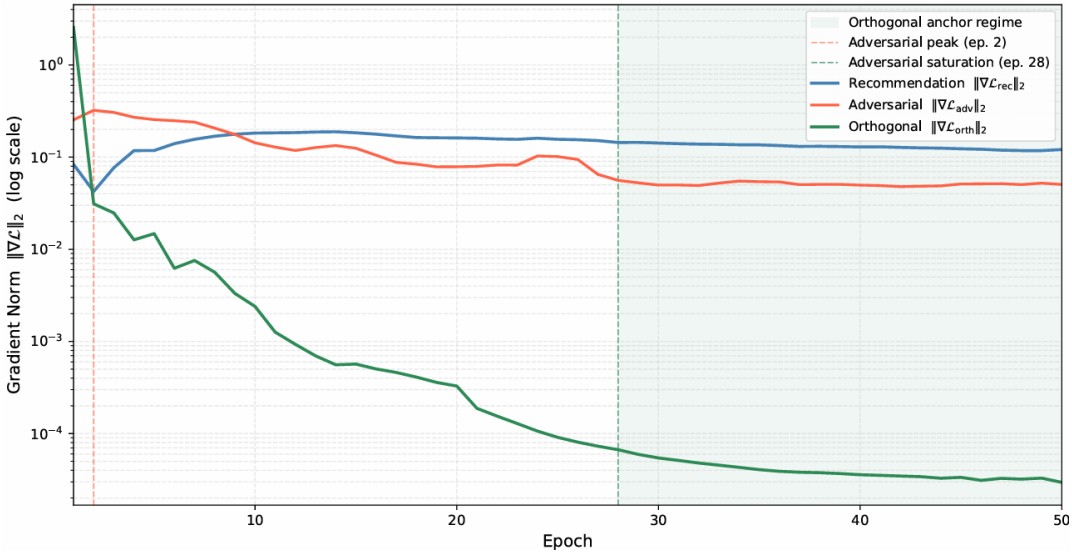


Figure 5.3: Gradient norm evolution per loss component across 50 training epochs on ML-100K.

The analysis points to a single equilibrium reached around epoch 28. The recommendation gradient remains stable and approximately flat throughout training, confirming that the primary objective is not destabilized by the fairness-related terms. The adversarial gradient peaks sharply at epoch 2, decreases progressively as the minimax game is played out, and stabilizes around epoch 28. The orthogonal gradient decays monotonically from epoch 1 and converges to near-zero values by the same epoch, after which it remains in a stable anchor regime. The fact that the orthogonal regularization is effectively satisfied well before the end of training, and that none of the three gradients exhibit oscillatory behavior at convergence, provides quantitative evidence that the three components play distinct and non-conflicting roles and that the composite objective converges in a stable, interpretable manner.

5.4 Experimental Details

Our experiments are conducted on three publicly available datasets: ML-1M, BookCrossing, and ML-100K. The detailed characteristics of these datasets are presented in Table 5.2.

- **MovieLens-1M (ML-1M)**: A benchmark dataset for Recommender systems containing approximately 1 million explicit ratings (on a 1-5 star scale) from 6, 040 users on 3, 952 movies. The dataset includes rich side information, such as user demographics (age, gender, occupation) and movie genres. Its density and clear user attributes make it ideal for studying demographic biases.
- **MovieLens-100K (ML-100K)**: A smaller subset of the MovieLens data containing 100, 000 ratings from 943 users on 1, 682 movies. While sharing the same structure as ML-1M, its smaller scale allows for efficient experimentation and hyperparameter tuning, providing a complementary view of model performance on datasets of different sizes within the same domain.
- **Book-Crossing (BookCrossing)**: A book rating dataset collected from the Book-Crossing community, containing 278, 858 ratings from 77, 805 users on 185, 973 books. This dataset is characterized by extreme sparsity and includes user age information and book metadata. It was selected to evaluate the generalizability of ORAF to a different domain (books versus movies) and to test its effectiveness with an alternative sensitive attribute (age instead of gender).

Table 5.2: Statistical Summary of Datasets Used

Property	**ML-100K**	**ML-1M**	**Book-Crossing**
Users	943	6, 040	77, 805
Items	1, 682	3, 952	185, 973
Interactions	100, 000	1, 000, 209	278, 858
Sparsity	93.70%	95.81%	99.98%
Domain	Movies	Movies	Books
User Attributes	Gender, Age, Occupation	Gender, Age, Occupation	Age, Location
Item Attributes	Genre, Release Year	Genre, Release Year	Author, Publisher, Year

The strategic selection of these datasets facilitates a comprehensive evaluation of the proposed framework. Specifically, the inclusion of both ML-100K and ML-1M enables an assessment of model scalability and performance consistency across varying data magnitudes within the movie domain. Complementarily, the Book-Crossing dataset interrogates the framework's robustness and cross-domain generalizability when subjected to extreme sparsity and a distinct sensitive attribute (age attribute).

To comprehensively evaluate the efficacy of the proposed ORAF framework, we adopt a two-stage comparative analysis: First, we benchmark ORAF against state-of-the-art user fairness-focused approaches to assess its capability in mitigating demographic biases (specifically gender bias). Subsequently, we compare ORAF against leading item fairness-focused methods to evaluate its effectiveness in addressing popularity bias and ensuring equitable item exposure.

To assess the performance of the proposed approach, we divided the users into training, validation, and test sets in a 7: 1: 2 ratio. The model was trained for up to 50 epochs with early stopping triggered if there was no improvement for 20 consecutive epochs or if performance decreased for 5 consecutive epochs. We used the Rectified Linear Unit (ReLU) as the activation function and the Adam optimizer. For the training dynamics, the generator was trained for 1 epoch per cycle, while the discriminators were trained for 3 epochs.

To determine the optimal hyperparameters, we performed a grid search over the values $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. for both λ_g, λ_p, λ_{orth}. The optimal values yielding the best performance are set for the final reported results to $\lambda_g = 0.1$, $\lambda_p = 0.1$, $\lambda_{orth} = 0.1$ respectively. Similarly, for the learning rate, we tested values in the same range and selected 10^{-2} as the best performing value. For embedding size, we tested values in the range $\{32, 48, 64, 125, 258\}$, and the final selected value was 48.

For user fairness, we use the gender attribute as the sensitive feature for the ML-100K and ML-1M datasets, and the age attribute as the sensitive feature for the BookCrossing dataset. Our goal is to assess whether sensitive information remains embedded in the newly learned fair user representation, or if it has been effectively eliminated. To achieve this, we treat the users' attributes (gender or age) from the training data as the ground truth and train a linear classifier using the fair user representations as input. The classification accuracy on the test set of users is reported to evaluate individual fairness, with the AUC metric used for performance assessment.

For item fairness, we use the group of items (popular or unpopular) as the sensitive attribute across all datasets. Our objective is to determine whether sensitive information related to item popularity has been effectively removed. To do this, we treat the popularity group of the items in the training data as the ground truth and train a linear classifier using the fair item representations as input. To construct item groups, for all datasets, a predefined threshold is applied. In accordance with [144] and the Pareto principle, we classify the top 20% of items with the highest rating counts

as "popular items" (group 1) and the remaining 80% as "non-popular items" (group 2) for the ML-1M and BookCrossing datasets. For the ML-100K dataset, we use the top 5% as popular items and the remaining 95% as non-popular items due to its smaller size and to match the baseline models we compare our framework with.

5.5 Evaluation Metrics

5.5.1 Recommendation Quality

We evaluate the rating prediction and ranking performance of our ORAF framework using MAE and NDCG metrics respectively. Higher values of NDCG and lower values of MAE represent more accurate and effective recommendations.

- **Normalized Discounted Cumulative Gain (NDCG):** NDCG assesses the ranking quality of recommendations by comparing predicted rankings with ideal rankings. Higher values indicate better ranking quality.

$$NDCG_k = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{DCG_u^k}{IDCG_u^k} \tag{5.37}$$

$$DCG_u^k = \sum_{r=1}^{k} \frac{2^{y_{ru}} - 1}{\log_2(1+r)} \tag{5.38}$$

- **Mean Absolute Error (MAE):** MAE measures the average absolute error between the predicted and true values. A lower MAE indicates better rating prediction.

$$MAE = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{1}{|D_u^q|} \sum_{i \in D_u^e} |y_{ui} - \hat{y}_{ui}| \tag{5.39}$$

5.5.2 User-side Fairness Metrics

To ensure rigorous comparison with baseline models, we adopted the user-side fairness metrics used in [98]

- **Area Under the Curve (AUC):** The Area Under the Curve (AUC) is a widely used metric in classification tasks that evaluates performance across all possible threshold values by considering the True Positive Rate (TPR) and False Positive Rate (FPR). In the context of fair recommender systems, AUC is commonly employed to assess how effectively sensitive information has been removed from learned representations. Specifically, in this work, AUC is used to measure the extent to which sensitive attributes such as gender can still be inferred from user embeddings. To compute this, we train a simple binary classifier on the learned representations and evaluate its predictive performance using AUC.

- **Attribute Substitution Stability (CF)**

  To evaluate the sensitivity of the model's recommendations to changes in a user's sensitive attribute, we adopt an attribute substitution stability metric, which we denote CF. This metric quantifies how much the predicted scores change when a user's sensitive attribute is synthetically flipped to its complementary value while all other user and item features are held constant. Formally, let S_u denote the sensitive attribute of user u (e.g., gender), and let S'_u denote its counterfactual value (e.g., switching from female to male). Let $\mathbf{x}_u$ denote the remaining user attributes, which are kept unchanged. The metric is defined as:

$$\mathrm{CF} = \frac{1}{|U_f|} \sum_{u \in U_f} \frac{1}{|D_q^u|} \sum_{i \in D_q^u} |\hat{y}_{ui}(\mathbf{x}_u, S_u) - \hat{y}_{ui}(\mathbf{x}_u, S'_u)| \tag{5.40}$$

  where $\hat{y}_{ui}$ is the predicted score for user u and item i, D_q^u is the query item set for user u, and U_f is the set of users evaluated for fairness. A lower CF value indicates that the model's recommendations are stable under attribute perturbation, which is a desirable fairness property.

- **Group Fairness (GF):** GF compares recommendation accuracy (MAE) across different groups (e.g., based on sensitive attributes such as gender or item popularity). A lower GF value indicates better fairness between groups.

$$GF = \left| \frac{1}{|A_1|} \sum_{u_1 \in A_1} R(u_1) - \frac{1}{|A_2|} \sum_{u_2 \in A_2} R(u_2) \right|$$

Relationships Among the User Fairness Metrics. The fairness indicators used in this work are not interchangeable; each captures fairness at a distinct analytical level. At the *representational level*, AUC measures how recoverable the sensitive attribute is from the learned embeddings by an external classifier, thereby quantifying attribute leakage in the representation space itself. An AUC close to 0.5 indicates that the embeddings carry no linear information that a classifier can exploit to recover the protected attribute. At the *individual level*, CF (Attribute Substitution Stability) measures the degree to which a recommendation for a single user remains invariant when their sensitive attribute is counterfactually flipped. An CF close to zero indicates that the system is "attribute-blind" at the point of decision-making. Finally, at the *group level*, GF (Group Fairness) evaluates the disparity in recommendation quality (e.g., MAE) between demographic groups, quantifying the equity of outcomes across the population.

In ORAF, these three levels do not stand in a state of mutual tension; rather, they are harmonized through the specific mechanistic design of the architecture. The adversarial discriminator targets the representational level (AUC), purging the "raw material" of bias from the embeddings. However, because adversarial training alone often leaves residual non-linear leakage pathways open,

a gap can emerge where AUC is low but behavioral metrics (CF and GF) remain high. ORAF resolves this by utilizing orthogonal regularization to close these exact pathways, ensuring that representational privacy translates directly into individual stability and group equity. The attribute predictors act as a semantic anchor for this process, ensuring that the geometric separation is precisely aligned with the sensitive-attribute direction.

Our experimental analysis confirms this synergy: representational, individual, and group fairness improve monotonically across the ORAF variants (see the ablation results in Table 5.4). By addressing the source of bias (representation) and its propagation (geometry), ORAF transforms what is traditionally viewed as a series of fairness trade-offs into a unified and verifiable debiasing guarantee. While this multi-level approach incurs a marginal utility cost, the resulting Pareto-optimal balance is highly acceptable and necessary for ethically-aligned, high-stakes recommendation environments.

5.5.3 Item-side Fairness Metrics

For the item-side evaluation, to compare ORAF with existing popularity bias mitigation models, we use the same metrics that were used in their respective works [145].

- **Precision@K** Precision@K measures the proportion of relevant items among the top-K recommended items for a user. It is defined as:

$$\text{Precision@K}(u) = \frac{|\{i \mid i \in R_u \cap L_u\}|}{K}, \tag{5.41}$$

where R_u is the set of recommended items for user u, L_u is the set of items that user u actually interacted with, and K is the number of recommended items.

- **Recall@K** Recall@K measures the proportion of relevant items among all possible relevant items for a user, considering only the top-K recommended items. It is defined as:

$$\text{Recall@K}(u) = \frac{|\{i \mid i \in R_u \cap L_u\}|}{|L_u|}, \tag{5.42}$$

where L_u is the set of all relevant items for user u.

- **Item Ranking-based Statistical Parity (RSP)** Measures the balance of exposure between item groups, whether items from the popular group and the unpopular group have an equal probability of appearing in recommendation lists, regardless of whether the user actually likes them. It is defined as:

$$\text{RSP@K} = \frac{1(P_1(K), P_2(K), \ldots, P_n(K))}{\text{mean}(P_1(K), P_2(K), \ldots, P_n(K))}, \tag{5.43}$$

where $P_j(K)$ is the probability of an item from group G_{Ij} appearing in the top-K recommendations.

- **Item Ranking-based Equal Opportunity (REO):** Assesses fairness by comparing the proportion of relevant items exposed to different item groups in the top-K recommendations. This ensures that a relevant unpopular item has the same probability of being recommended as a relevant popular item. It is defined as:

$$\text{REO@K} = \frac{1(P_1(K|y=1), P_2(K|y=1), \ldots, P_n(K|y=1))}{\text{mean}(P_1(K|y=1), P_2(K|y=1), \ldots, P_n(K|y=1))}, \tag{5.44}$$

where $P_j(K|y=1)$ is the probability that an item from group G_{Ij} is among the true positives in the top-K recommendations.

5.6 System Comparison Setup

5.6.1 User-Side Fairness

To analyze the effectiveness of ORAF on the user-side fairness, we compare ORAF with several state-of-the-art traditional and user fairness aware models across:

- **PPR**: (Pairwise Preference Regression): This method estimates what users like by using a special kind of regression model called bilinear regression [142].
- **Wide & Deep**: This approach uses a deep neural network to predict whether a user will prefer a certain item [35].
- **DropoutNet:** By applying the dropout technique within a deep neural network, this model learns input features and addresses the cold-start problem, where little user data is available [41].
- **NLBA:** This is a cold-start neural network recommender systems where the weights of the model stay the same for all users, but the biases of the output and hidden layers are personalized for each user [42].
- **MELU:** A model that estimates user preferences using meta-learning, allowing it to quickly adjust to new users [43].
- **MACR:** It offers a general framework to reduce bias in RecSys. It focuses on capturing bias related to sensitive information in the given context [105].
- **CLOVER:** This is a comprehensive fair meta-learning framework focuses on ensuring three types of fairness individual, counterfactual, and group fairness by using a multi-task adversarial learning approach, making it adaptable to various meta-learned models [98].

- **DALFRec:** A dual-side based adversarial learning for fairness-aware recommendation algorithm that filters sensitive information from both user and item sides using adversarial learning [99].

5.6.2 Item-side Fairness

To assess the efficacy of ORAF in mitigating popularity bias, we conduct a comparative analysis with the following models specifically designed to address popularity bias.

1. **DPR** [146]: An adversarial Bayesian model that reduces item-side popularity bias in recommendations using adversarial training.
2. **ADB** [147]: An end-to-end meta-learning framework that automatically mitigate different types biases, including popularity bias in the user interaction data without explicit bias modeling.
3. **UOF** [148]: Tackles popularity bias from the user side by ensuring balanced exposure of recommended items across different user groups, using reinforcement learning to optimize fairness.
4. **MACR** [105]: Eliminates popularity bias using model-agnostic counterfactual reasoning to simulate unbiased recommendations. to any recommendation model.
5. **FAiR**[145]: A fairness-centric model that adaptively mitigates the popularity bias in both users and items for recommendation.

5.7 Experimental Results and Analysis

In this section, we perform extensive experiments to assess the fairness and effectiveness of the proposed ORAF. Additionally, we conduct ablation studies and other analyzes to gain deeper insights into the performance of the model. We specifically aim to address the following research questions (RQs).

- **RQ1:** How Does the Integration of Two-Sided Fairness Mechanism Influence the Model Performance?
- **RQ2:** This RQ analyzes The performance of ORAF on the user-side fairness. It is devised into three sub-questions:
  - RQ2a: How does ORAF perform compared to traditional and user-fairness-aware recommendation models.

  - RQ2b: How Does the Gender Discriminator and user orthogonal regularization enhance the bias mitigation.
  - RQ2c: How Does ORAF perform for each user's group (male vs female)?

- **RQ3:** How ORAF performs on the item-side fairness? This RQ also has three sub-RQs similar to the previous RQ2.
- **RQ4: Ablation Study**
  - RQ4a: How does the complexity of the RSs algorithms impact the performance?
  - RQ4b: How do the discriminators' update frequencies affect the fairness and performance of ORAF?
  - RQ4c: The Impact of learning rates on the ORAF's Performance.
  - RQ4d: The Sensitivity of the Hyperparameters terms on the performance of ORAF?
- **RQ5:** How does ORAF perform in a real case study?

RSQ1: How Does the Integration of Two-Sided Fairness Mechanism Influence Model Performance?

To study the effect of integrating two-sided fairness mechanisms on recommendation performance, we compare four model variants: the baseline ORAF_BaseRec (no fairness mechanisms), ORAF_user (user-side fairness only), ORAF_item (item-side fairness only), and the full ORAF model (both fairness sides mechanisms). The comparative results across utility, item fairness, and user fairness metrics are presented in Figure 5.4.

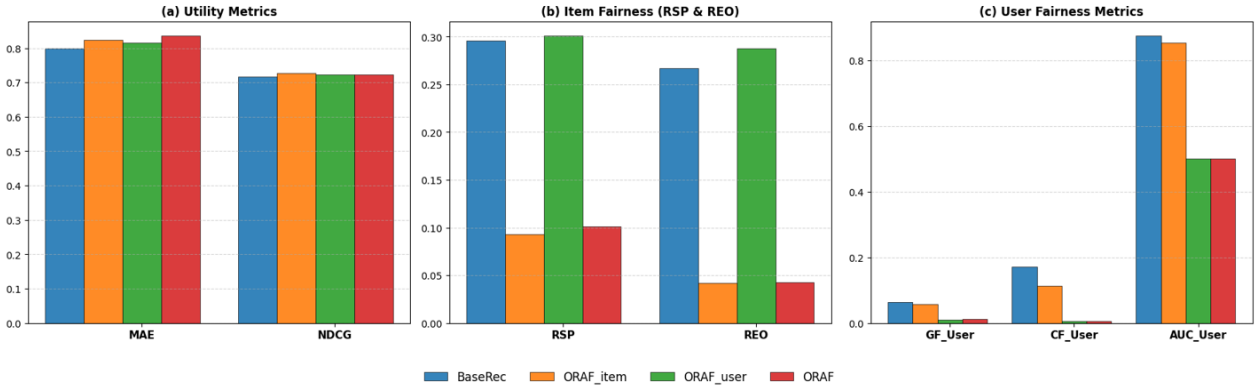


Figure 5.4: Impact of Two-sided Fairness Mechanisms on Recommender Systems' Performance

Our component-wise analysis reveals clear functional separation between the user-side and item-side debiasing mechanisms. The ORAF_item variant effectively addresses popularity bias, as shown

by substantial improvements in RSP and REO. However, it proves entirely ineffective against demographic leakage, maintaining high gender prediction AUC scores comparable to the baseline. In contrast, ORAF_user achieves near-complete demographic removal, reducing AUC to approximately 0.5, but shows no measurable effect on popularity bias metrics.

The full ORAF framework demonstrates successful integration of both mechanisms. Rather than experiencing destructive interference between competing adversarial objectives, the complete model inherits the strengths of each component, maintaining both the reduced ranking disparity characteristic of the item-fairness variant and the strong privacy protection (low AUC and ASS error) achieved by the user-fairness variant.

Regarding the impact on recommendation utility, the study highlights the intrinsic cost of two-sided debiasing. We observe that all fairness-aware variants, whether targeting user fairness, item fairness, or both, exhibit a decrease in NDCG and a marginal increase in MAE across all ORAF configurations compared to the unconstrained BaseRec model.

RSQ2: User-side Fairness Evaluation

A - Performance Comparison

This experiment evaluates the effectiveness of ORAF in achieving user-side fairness through comprehensive experiments across three datasets (ML-1M, BookCrossing, and ML-100k). As shown in Table 5.3, ORAF achieves the best balance between recommendation quality and user fairness.

Table 5.3: Comparison of User Side Performance on ML-1M, BookCrossing, and ML-100k.

	ML-1M					BookCrossing					ML-100K				
Model	**MAE**	**NDCG**	**AUC**	**GF**	**CF**	**MAE**	**NDCG**	**AUC**	**GF**	**CF**	**MAE**	**NDCG**	**AUC**	**GF**	**CF**
PPR	0.943	0.672	0.187	0.052	0.187	2.374	0.598	0.694	0.101	0.194	1.398	0.486	0.772	0.051	0.095
Wide&Deep	0.815	0.691	0.214	0.055	0.214	1.945	0.627	0.156	0.098	0.156	1.256	0.863	0.112	0.046	0.112
DropoutNet	0.813	0.702	0.204	0.063	0.204	1.855	0.634	0.194	0.103	0.194	1.172	0.842	0.131	0.044	0.131
NLBA	0.795	0.701	0.254	0.056	0.254	1.718	0.651	0.556	0.112	0.156	1.213	0.553	0.891	0.047	0.129
MELU	0.743	0.755	0.264	0.054	0.264	1.332	0.723	0.559	0.103	0.259	0.892	0.652	0.974	0.048	0.157
MELU+BS	0.745	0.744	1.000	0.049	0.277	1.343	0.712	1.000	0.101	0.291	0.891	0.651	1.000	0.046	0.174
MELU+Reg	0.762	0.748	0.894	0.048	0.145	1.357	0.713	0.912	0.097	0.232	0.914	0.648	0.873	0.045	0.132
MELU+IPW	0.751	0.751	0.975	0.047	0.187	1.365	0.707	1.000	0.096	0.245	0.903	0.643	0.946	0.047	0.117
MELU+MACR	0.744	0.750	0.878	0.051	0.116	1.352	0.715	0.867	0.097	0.179	0.887	0.656	0.815	0.044	0.108
CLOVER	**0.731**	0.756	0.632	0.040	0.044	1.352	0.723	0.546	0.103	0.027	0.880	0.666	0.562	0.036	0.032
DALFRec	0.735	0.758	0.533	0.019	0.005	1.330	0.725	0.544	0.081	0.002	0.869	0.677	0.505	0.026	0.007
ORAF	0.738	**0.766**	**0.513**	**0.012**	**0.003**	**1.279**	**0.736**	**0.525**	**0.061**	**0.001**	**0.836**	**0.721**	**0.501**	**0.022**	**0.004**

When compared to traditional accuracy-centric models (e.g., Wide&Deep, DropoutNet, MELU), the necessity for fairness constraints becomes evident. Traditional baselines exhibit significant sensitive information leakage, indicated by high gender AUC scores (approaching 1.0 for MELU+BS) and substantial group disparities. These results confirm that standard architectures inadvertently

capture and amplify demographic biases present in the training data.

Against user fairness-aware models, ORAF achieves substantial quantitative improvements in fairness metrics. While DALFRec achieves competitive accuracy through adversarial training alone (MAE: 0.735 on ML-1M), ORAF's dual-mechanism approach demonstrates superior fairness: 40% improvement in attribute substitution stability (CF: 0.003 vs. 0.005) and 37% improvement in group fairness (GF: 0.012 vs. 0.019).

The numerical superiority of ORAF over state-of-the-art baselines like DALFRec and CLOVER can be attributed to the specific architectural synergy between its adversarial and geometric components. While models like DALFRec utilize a dual-adversarial framework to minimize the distributional recoverability of sensitive attributes, they are fundamentally limited by the nature of the minimax game. In vanilla adversarial training, the generator can reach a local equilibrium where it hides sensitive information within complex, non-linear latent patterns that the capacity of the current discriminator cannot detect. This creates a residual bias leakage that explains why DALFRec maintains a higher AUC (0.533 on ML-1M) and higher instability (CF of 0.005).

In contrast, the improvement of ORAF is driven by the deterministic nature of the orthogonal regularization. Unlike the discriminator, which is a stochastic probe whose strength fluctuates during training, the orthogonal penalty is a closed-form geometric constraint that enforces zero projection between the debiased and bias-explicit subspaces. By ensuring that the representations share no linear component along the directions captured by the attribute predictors, ORAF fundamentally closes the disentanglement gap that purely adversarial models leave open. Furthermore, our use of dedicated attribute predictors serves to semantically ground this geometric separation. While vanilla models may orthogonalize against arbitrary directions, ORAF forces the separation to occur against a learned proxy of the true bias signal. This mechanistic combination, involving distributional unlearning via discriminators and structural decoupling via orthogonality, is the causal driver behind the near-perfect representational privacy (AUC ≈ 0.51) and significantly lower group disparity (GF ≈ 0.012) observed in our results.

B - Component-Wise Contribution Analysis for User-Side Fairness

This ablation study evaluates the user-side fairness components of ORAF through three progressive variants: **Base** (no fairness mechanism), **Base + Adv (U)** (Base with the adversarial gender discriminator on user embeddings), and **Base + Adv + Orth (U)** (Base + Adv (U) with the user-side orthogonal regularization added on top). Each variant is obtained by *adding* one component to the previous one, so the naming directly mirrors the architectural stack. Results are presented in Table 5.4.

The **Base** configuration exhibits clear gender bias across all datasets, with high gender predictability (AUC = 0.912 for ML-1M, 0.851 for BookCrossing) and notable counterfactual variance

Table 5.4: Ablation Study: Component-wise Analysis of User-Side Debiasing Mechanisms.

	ML-1M					BookCrossing					ML-100K				
Model	**MAE**	**NDCG**	**AUC**	**CF**	**GF**	**MAE**	**NDCG**	**AUC**	**CF**	**GF**	**MAE**	**NDCG**	**AUC**	**CF**	**GF**
Base	0.721	0.772	0.912	0.214	0.049	1.262	0.749	0.851	0.007	0.093	0.801	0.726	0.853	0.14	0.039
Base + Adv (U)	0.734	0.765	0.526	0.006	0.018	1.276	0.735	0.530	0.004	0.088	0.825	0.722	0.519	0.006	0.025
Base + Adv + Orth (U)	0.738	0.766	0.513	0.003	0.012	1.279	0.736	0.525	0.001	0.061	0.836	0.721	0.501	0.004	0.022

(CF = 0.214 for ML-1M). This confirms that the model encodes gender information in user representations when no fairness constraints are applied.

Adding the adversarial discriminator (**Base + Adv (U)**) produces a large reduction in gender predictability. AUC drops from 0.912 to 0.526 on ML-1M and CF improves from 0.214 to 0.006, which confirms that adversarial training effectively disrupts the linear separability of gender groups in the embedding space.

Incorporating the orthogonal regularization (**Base + Adv + Orth (U)**) yields a further reduction in AUC from 0.526 to 0.513 on ML-1M. The marginal nature of this gain is expected, since the adversarial component has already driven AUC close to the random-chance baseline of 0.5, leaving limited room for further reduction on this single metric. The value of the orthogonal regularization in this regime lies in enforcing a deterministic geometric separation between the bias-explicit and debiased embedding subspaces, which complements adversarial indistinguishability. The CF and GF metrics show more consistent gains from the orthogonal regularization across all three datasets, and the improvements are more pronounced on ML-100K and BookCrossing, where the adversarial component leaves more residual bias to address.

Sensitivity to Dataset Scale and Sparsity. The results reported in Table 5.4 reveal a consistent dataset-dependent pattern that warrants explicit analysis. We therefore conduct a comparative sensitivity analysis across the three benchmarks summarized in Table 5.2, which represent a diverse spectrum of scale and interaction density.

Our empirical results indicate that dataset scale is a primary determinant of the stability of the adversarial minimax game. On the larger ML-1M dataset (over one million interactions), the discriminators have access to a rich distribution of samples, allowing them to establish robust decision boundaries. This stronger adversarial signal forces the generator to produce highly refined, debiased embeddings, leading to near-optimal fairness results (AUC $\approx$ 0.513, CF $\approx$ 0.003, GF $\approx$ 0.012) with minimal utility loss ($\approx$ 2.3% MAE increase). In contrast, on the smaller ML-100K dataset, the reduced sample size leads to more stochastic discriminator updates. As a result, the generator faces a less consistent adversarial signal, which explains the slightly higher absolute error at smaller scales (MAE $\approx$ 0.836) while still achieving near-perfect representational privacy.

Furthermore, dataset sparsity significantly influences the ease of representation-level disentan-

glement. The Book-Crossing dataset, characterized by extreme sparsity, presents the most challenging environment for ORAF. In such sparse settings, user embeddings are estimated from a very limited number of interactions, which often produces noisy representations in which true preferences and demographic signals are tightly coupled. ORAF's orthogonal regularization nonetheless achieves a substantial reduction in attribute substitution stability ($0.007 \rightarrow 0.001$) and group fairness ($0.093 \rightarrow 0.061$) on this dataset, as reported in Table 5.4, although the residual bias remains slightly higher than on the denser movie benchmarks. This confirms that while ORAF's geometric constraints are effective across domains, their ability to close the disentanglement gap is most pronounced when the underlying embeddings are well-estimated through sufficient interaction density.

Conclusively, while ORAF exhibits a marginal performance decline on smaller or sparser datasets, its hybrid architecture, and in particular the deterministic nature of the orthogonal regularization, provides a vital safety net. The orthogonal regularization is a closed-form geometric quantity computed at every training step; it remains effective even when the adversarial signal is noisy, preventing the utility collapse often observed in purely adversarial models within sparse regimes. The monotonic improvement of the three user-side fairness metrics across our configurations (Base $<$ Base+Adv $<$ ORAF) is preserved across all datasets, as visible row by row in Table 5.4, demonstrating that the small-dataset and high-sparsity effects are properties of the operating regime rather than limitations of the framework. ORAF therefore maintains competitive accuracy and qualitatively consistent debiasing behavior across varying data magnitudes and densities.

Quantifying the Fairness-Accuracy Trade-off. We assess the utility-fairness trade-off of ORAF along two complementary axes: in absolute terms, by comparing the full configuration against the unconstrained Base recommender, and in relative terms, by benchmarking it against existing two-sided fairness methods.

Relative to the Base recommender. A utility-fairness trade-off does exist relative to the Base recommender, but it is very small and therefore acceptable. ORAF's full configuration introduces an MAE increase of only $\approx 2.3\%$ on ML-1M, $\approx 4.4\%$ on ML-100K, and $\approx 1.3\%$ on BookCrossing (Table 5.4), which is acceptable. In return, ORAF reduces gender leakage (AUC) by $\approx 44\%$ on ML-1M, $\approx 41\%$ on ML-100K, and $\approx 38\%$ on BookCrossing, with corresponding reductions in counterfactual instability (CF) of $\approx 99\%$ on ML-1M, $\approx 97\%$ on ML-100K, and $\approx 86\%$ on BookCrossing. The fairness gain is therefore one to two orders of magnitude larger than the utility cost on every dataset.

Relative to other fairness baselines (Table 5.3), ORAF achieves the best fairness on every dataset and on every fairness metric (AUC, GF, CF). On ML-1M its utility is competitive with the strongest baselines (MAE 0.738 vs. CLOVER 0.731 and DALFRec 0.735), while its GF and CF are substantially lower than those of every competing method; on ML-100K and BookCrossing, ORAF additionally achieves the lowest MAE and the highest NDCG, Pareto-dominating every competing

method.

We therefore conclude that ORAF's fairness-accuracy trade-off is both quantitatively acceptable in absolute terms and quantitatively favorable to the state of the art in relative terms.

C - Sensitive Attribute Inference Across Nonlinear Classifiers

To rigorously assess whether sensitive attribute information remains detectable in the debiased embeddings beyond the linear separability evaluated by the original AUC metric, we conduct an extended evaluation using four classifiers of increasing structural complexity: a Multi-Layer Perceptron (MLP), a Support Vector Machine with RBF kernel (SVM), a Random Forest (RF), and Gradient Boosting (GB). Each classifier is trained on frozen debiased user embeddings extracted from a fully trained model and evaluated on a stratified held-out split. All experiments are repeated over five independent runs with different random seeds, and statistical significance is assessed via paired t-tests across seeds. Results are reported in Table 5.5.

Table 5.5: Adversarial probing study: attribute inference AUC under nonlinear attack models (mean ± std over 5 seeds).

Dataset	**Attack Model**	**BaseRec**	**Full ORAF**	**P-value**
ML-100K	MLP	1.0000 ± 0.0001	0.5153 ± 0.0120	1.4210^{-7}
	SVM with RBF kernel	0.9999 ± 0.0002	0.5183 ± 0.0140	2.6510^{-7}
	Random Forest	0.9995 ± 0.0005	0.5053 ± 0.0044	1.8510^{-9}
	Gradient Boosting	0.9989 ± 0.0004	0.5087 ± 0.0107	8.7210^{-8}
ML-1M	MLP	0.9995 ± 0.0004	0.5050 ± 0.0062	1.1610^{-8}
	SVM with RBF kernel	0.9992 ± 0.0007	0.5066 ± 0.0047	5.1610^{-9}
	Random Forest (ensemble)	0.9977 ± 0.0018	0.5059 ± 0.0068	2.8910^{-8}
	Gradient Boosting (ensemble)	0.9991 ± 0.0008	0.5057 ± 0.0067	1.7210^{-8}

The results demonstrate a consistent and statistically significant reduction in attribute inference AUC across all four classifiers and both datasets, with all reductions significant at $p < 0.001$. On ML-1M, the debiasing effect is particularly pronounced: all four classifiers converge to AUC values in the narrow range of 0.5046 to 0.5066, with standard deviations below 0.007, indicating highly stable and near-chance-level performance across seeds. This convergence across classifiers of fundamentally different inductive biases, neural, kernel-based, and tree-based, provides strong empirical evidence that the reduction in gender detectability reflects a general disruption of the attribute-related structure in the embedding space.

On ML-100K, the pattern is directionally consistent but the residual AUC values are moderately

higher, ranging from 0.5053 to 0.5183, with larger standard deviations across seeds compared to ML-1M. This difference is expected: the smaller dataset provides a weaker training signal for both the adversarial discriminators and the orthogonal regularization, resulting in slightly less stable debiasing across runs. Nevertheless, the reductions from BaseRec to Full ORAF remain statistically significant across all classifiers, and the magnitude of the reduction from near-perfect attribute inference (AUC $\approx$ 1.0) to near-chance level is consistent with the ML-1M results.

D - User Group Fairness Analysis Across Demographic Segments

This analysis examines the equitable treatment of demographic groups to assess ORAF's effectiveness in achieving user-side fairness. We evaluate performance disparities between male and female user groups across three benchmark datasets, with detailed results presented in Table 5.6. The findings provide critical insights into the model's ability to maintain balanced performance across demographic segments.

Table 5.6: Performance comparison across demographic groups across five runs.

Metric	**ML-1M**		**BookCrossing**		**ML-100K**	
	Male	**Female**	**Young**	**Old**	**Male**	**Female**
MAE	0.732 ± 0.021	0.744 ± 0.013	1.298 ± 0.027	1.275 ± 0.027	0.825± 0.018	0.847 ± 0.016
NDCG@3	0.770 ± 0.013	0.762 ± 0.010	0.742 ± 0.06	0.733 ± 0.010	0.728 ± 0.016	0.714 ± 0.013
NDCG@5	0.791 ± 0.081	0.787 ± 0.06	0.757 ± 0.015	0.744± 0.015	0.764 ± 0.05	0.750 ± 0.01

The results demonstrate striking consistency in predictive accuracy across demographic groups. MAE values show minimal divergence, with relative differences remaining below 3% across all datasets. For ML-1M, the marginal disparity between male and female users (0.732 vs. 0.744, $\Delta = 1.6\%$) indicates near-perfect gender parity. Similarly, ML-100K exhibits comparable fairness (0.825 vs. 0.847, $\Delta = 2.6\%$), while BookCrossing reveals a slight advantage for older users (1.275 vs. 1.298, $\Delta = 1.8\%$), potentially attributable to more homogeneous preference patterns in this demographic. NDCG metrics confirm these findings, with gaps consistently below 0.014 points ($< 2\%$ relative difference), approaching the theoretical lower bound for group fairness.

RSQ3: Item-side Fairness Evaluation

A - Item-side Fairness Performance Comparison for Popularity Bias Mitigation

This evaluation assesses ORAF's effectiveness in addressing item-side fairness, specifically focusing on mitigating popularity bias and ensuring equitable exposure for items across popularity strata. Table 5.7 presents a comprehensive comparison of ORAF against state-of-the-art fairness-aware

recommendation models on the ML-100K dataset, evaluating both recommendation accuracy and item-side fairness metrics related to popularity bias.

Table 5.7: Item-side fairness performance comparison for popularity bias mitigation on ML-100K dataset.

Model	**Precision@10**	**Recall@10**	**RSP@10**	**REO@10**
DPR	0.509±.006	0.474±.006	0.108±.001	0.082±.009
ADB	0.496±.012	0.452±.007	0.141±.010	0.144±.007
UOF	0.478±.005	0.463±.009	0.183±.011	0.171±.013
MACR	0.518±.015	0.487±.010	0.126±.012	0.159±.008
FAIR	0.527±.019	0.497±0.018	**0.096±.010**	0.068±.014
ORAF	**0.529 ± .005**	**0.565 ±.008**	0.105 ± .032	**0.043 ± .037**

ORAF demonstrates superior performance in balancing recommendation accuracy with item-side fairness objectives related to popularity bias. The model achieves the highest Precision@10 (0.529) and Recall@10 (0.565) among all evaluated fairness-aware approaches, indicating that its popularity debiasing mechanisms do not compromise the identification of relevant items. This preservation of predictive performance while implementing item-side fairness constraints represents a significant advancement over existing methods.

In terms of statistical parity for item exposure, ORAF maintains competitive performance with an RSP@10 value of 0.105, closely approaching FAiR's best performance (0.096). The RSP value of 0.105 indicates that the exposure distribution between popular and unpopular items differs by approximately 10.5%, meaning that the proportion of popular versus unpopular items appearing in recommendation lists deviates only marginally from a perfectly balanced distribution.

Regarding relevance-aware fairness, ORAF achieves the best Ranking-based Equal Opportunity (REO@10) value of 0.043 among all compared models. This low REO value indicates that the disparity in ranking positions between popular and unpopular relevant items in user recommendation lists is minimal, specifically, popular and unpopular relevant items differ by only approximately 4% in their average ranking positions. This demonstrates that relevant items from unpopular groups receive equitable exposure opportunities comparable to their popular counterparts, ensuring that exposure fairness aligns with genuine user preferences rather than superficial popularity metrics.

B – Sensitivity Analysis of the Popularity Classification Threshold

To ensure our item-side fairness findings are not sensitive to the choice of popularity threshold, we repeat the evaluation across five thresholds ranging from 10% to 30% on both ML-100K and ML-1M, comparing BaseRec and Full ORAF at each setting.

Table 5.8: Popularity-threshold sensitivity analysis: for ML-100K and ML-1M.

		ML-100K				**ML-1M**			
Threshold	Model	Prec@10↑	Rec@10↑	RSP@10↓	REO@10↓	Prec@10↑	Rec@10↑	RSP@10↓	REO@10↓
Top-10%	BaseRec	0.5590	0.6030	0.3072	0.2351	0.6666	0.6321	0.4530	0.3725
	Full ORAF	0.5022	0.5578	0.1196	0.0555	0.6335	0.6088	0.1522	0.0717
Top-15%	BaseRec	0.5567	0.5998	0.3098	0.2349	0.6671	0.6321	0.4619	0.3879
	Full ORAF	0.4984	0.5564	0.1214	0.0592	0.6455	0.6150	0.1146	0.0335
Top-20%	BaseRec	0.5560	0.5992	0.3475	0.2642	0.6672	0.6325	0.4557	0.3842
	Full ORAF	0.5040	0.5605	0.1476	0.0810	0.6454	0.6156	0.1367	0.0538
Top-25%	BaseRec	0.5569	0.6009	0.3234	0.2326	0.6669	0.6319	0.4248	0.3550
	Full ORAF	0.5207	0.5719	0.1566	0.0763	0.6498	0.6194	0.1547	0.0707
Top-30%	BaseRec	0.5577	0.6010	0.3244	0.2266	0.6672	0.6323	0.4513	0.3770
	Full ORAF	0.5228	0.5745	0.1647	0.0794	0.6472	0.6177	0.1345	0.0503

Table 5.8 reports the results across all five threshold configurations. Full ORAF consistently incurs a reduction in Prec@10 and Rec@10 relative to BaseRec across both datasets, reflecting the well-known fairness-accuracy trade-off that is inherent to representation-level debiasing. Crucially, however, this utility cost is accompanied by substantial and consistent gains in item-side fairness. On ML-100K, RSP@10 drops from a range of 0.307-0.324 (BaseRec) to 0.120-0.165 (Full ORAF), representing a reduction of approximately 50-61%, while REO@10 falls from 0.226-0.264 to 0.056-0.081, a reduction of 54-76%. On the larger ML-1M dataset, the improvements are equally consistent: RSP@10 is reduced by 60-75% and REO@10 by 62-79% across all threshold levels. Importantly, the magnitude and direction of these improvements remain stable regardless of whether the popularity boundary is set at 10%, 15%, 20%, 25%, or 30%, confirming that the item-side fairness gains reported in the main experiments are robust and do not depend on any particular threshold choice.

C - Component Ablation Analysis for Item Popularity Bias Mitigation

The ablation study presented in Table 5.9 systematically evaluates the incremental contributions of the item-side fairness mechanisms through three progressive variants: **Base** (no fairness intervention), **Base + Adv (I)** (Base with the adversarial popularity discriminator D_p on item embeddings), and **Base + Adv + Orth (I)** (Base + Adv (I) with the item-side orthogonal regularization added on top). The naming mirrors that of the user-side ablation in Table 5.4, so the reader can interpret both studies with the same mental model.

The **Base** configuration exhibits substantial popularity bias across all datasets, with high RSP and REO values. This confirms that, without explicit fairness mechanisms, the model disproportionately favors popular items. While achieving strong utility metrics, the baseline shows significant

Table 5.9: Ablation Study: Component-wise Analysis of Item-Side Debiasing Mechanisms.

	ML-1M				BookCrossing				ML-100K			
Model	**rec@10**	**prec@10**	**RSP@10**	**REO@10**	**rec@10**	**prec@10**	**RSP@10**	**REO@10**	**rec@10**	**prec@10**	**RSP@10**	**REO@10**
Base	0.485 ± 0.001	0.654 ± 0.001	0.206 ± 0.002	0.216 ± 0.003	0.521 ± 0.006	0.634 ± 0.003	0.107 ± 0.003	0.134 ± 0.005	0.554 ± 0.002	0.605 ± 0.002	0.276 ± 0.005	0.220 ± 0.010
Base + Adv (I)	0.464 ± 0.002	0.614 ± 0.006	0.023 ± 0.021	0.023 ± 0.017	0.511 ± 0.006	0.622 ± 0.002	0.087 ± 0.003	0.111 ± 0.004	0.543 ± 0.004	0.589 ± 0.011	0.133 ± 0.098	0.122 ± 0.055
Base + Adv + Orth (I)	0.477 ± 0.003	0.643 ± 0.003	0.031 ± 0.023	0.018 ± 0.020	0.503 ± 0.006	0.614 ± 0.003	0.079 ± 0.004	0.097 ± 0.006	0.527 ± 0.005	0.565 ± 0.008	0.105 ± 0.032	0.043 ± 0.037

fairness deficits.

Adding the popularity discriminator (**Base + Adv (I)**) yields dramatic fairness improvements across all datasets. RSP and REO metrics decrease substantially, particularly on ML-1M where both drop by approximately 89%. This demonstrates the effectiveness of adversarial training at mitigating popularity bias, at a modest utility cost of 4-6% in recall and precision.

Incorporating the orthogonal regularization (**Base + Adv + Orth (I)**) provides complementary benefits by enforcing geometric separation between the bias-aware embeddings (which concentrate popularity information) and the debiased embeddings (which are used for recommendation). The orthogonality constraint guarantees that popularity information isolated in the bias-aware subspace cannot leak into the debiased subspace. Overall, the progressive analysis reveals that while adversarial training reduces popularity bias, the orthogonal regularization ensures complete isolation by enforcing independent subspaces for bias-aware and debiased item embeddings.

Effect of Attribute Predictors on Orthogonal Regularization

To empirically verify the necessity of the attribute predictor component, we introduce a controlled variant named **Base + Adv + Orth** (both sides, *without* attribute predictors), in which both the user-side gender predictor and the item-side popularity predictor are deactivated while the adversarial discriminators and the orthogonal regularization remain fully active on both sides. This variant is compared against the **Full ORAF** configuration (which adds the attribute predictors on top) over five independent runs, with results reported in Table 5.10. Detailed statistical significance results from the progressive component-wise ablation across all four model variants (Base, Base + Adv, Base + Adv + Orth, and Full ORAF) are further provided in Table B.3

The rationale for this experiment stems from a fundamental dependency within the ORAF architecture. In Full ORAF, the gender predictor optimizes the bias-aware user embeddings to encode maximum gender information, while the popularity predictor performs the analogous role for bias-aware item embeddings. Together, they establish well-defined bias subspaces on both sides that serve as references for the orthogonal regularizations. The orthogonal regularization then drives the debiased user and item embeddings away from their respective bias subspaces, ensuring that the separation is geometrically meaningful with respect to the sensitive attributes.

When both predictors are removed, neither the bias-aware user embeddings nor the bias-aware item embeddings receive any supervision toward capturing their respective sensitive attributes, and

consequently both encode mixtures of preference and content signals with no structured relationship to gender or popularity. Under these conditions, the orthogonal regularization is applied against uninformative, effectively random subspaces on both sides, rendering the geometric separation unrelated to the bias directions it was designed to address.

The experimental results substantiate this analysis. In the absence of the predictors, all five inference classifiers report AUC values in the range of 0.6561 to 0.6747, indicating that a substantial degree of gender information remains present in the debiased user embeddings. Full ORAF, by contrast, reduces all five classifiers to near-chance level, with AUC values between 0.5102 and 0.5132, and all differences are statistically significant across seeds. The attribute substitution stability metric (ASS) deteriorates from 0.0040 in Full ORAF to 0.0090 in **Base + Adv + Orth** ($p = 2.3810^{-4}$), reflecting a measurable reduction in the robustness of recommendations to sensitive attribute perturbation. Group fairness similarly worsens, increasing from 0.0228 in Full ORAF to 0.0406 in the no-predictor variant ($p = 1.2210^{-4}$).

As expected, recommendation quality metrics exhibit the inverse pattern. The no-predictor variant achieves lower MAE (0.8259 vs. 0.8356) and higher NDCG (0.7382 vs. 0.7288) than Full ORAF, since removing both predictors reduces the fairness constraint pressure on the recommendation objective and allows the model to optimize more freely for accuracy. Item-side fairness metrics show a clear improvement in Full ORAF, with RSP reducing from 0.1694 to 0.103 and REO from 0.1407 to 0.0401, consistent with the popularity predictor establishing a more structured item-side bias subspace against which the orthogonal regularization can operate effectively. Collectively, these findings establish that both attribute predictors are functionally indispensable components of ORAF: without them, the orthogonal regularization on both the user and item sides loses its directional grounding and ceases to perform targeted sensitive attribute removal from the debiased representations.

Ablation Study Experiments

RQ4a: Performance Comparison of ORAF Across Diverse Recommendation Architectures

To evaluate the generalizability of the proposed debiasing framework, we extend the analysis to nine recommendation architectures spanning a broad range of structural designs: classical collaborative filtering (CF), matrix factorization (MF), neural collaborative filtering (NCF), NeuMF, DeepFM, Wide&Deep, AutoRec, and two graph-based models, LightGCN and NGCF. The inclusion of LightGCN and NGCF is particularly relevant, as these models propagate information through a user-item interaction graph using graph convolution operations, representing a fundamentally different inductive bias from the embedding-lookup architectures that dominate the earlier comparisons. The comparative results before and after applying ORAF's debiasing module are visualized

Table 5.10: An ablation study on the role of attribute predictors in the orthogonal regularization. p-values from paired t-tests across 5 seeds.

Metric	**Base + Adv + Orth (no Pred)**	**Full ORAF**	**p-value**
MAE ↓	0.8259 ± 0.0140	0.8356 ± 0.0097	2.2410^{-4}
NDCG ↑	0.7382 ± 0.0165	0.7288 ± 0.0161	2.1810^{-3}
AUC(LR) ↓	0.6694 ± 0.0960	0.5124 ± 0.0322	9.4910^{-3}
AUC(SVM) ↓	0.6580 ± 0.0607	0.5102 ± 0.0252	2.0510^{-4}
AUC(RF) ↓	0.6561 ± 0.0684	0.5132 ± 0.0401	9.8410^{-3}
AUC(GB) ↓	0.6747 ± 0.0626	0.5105 ± 0.0200	1.5510^{-3}
AUC(MLP) ↓	0.6709 ± 0.0728	0.5102 ± 0.0307	6.3510^{-3}
CF ↓	0.0090 ± 0.0222	0.0040 ± 0.0158	2.3810^{-4}
GF ↓	0.0406 ± 0.0119	0.0228 ± 0.0143	1.2210^{-4}
RSP ↓	0.1694 ± 0.0600	0.103 ± 0.0368	6.8210^{-2}
REO ↓	0.1407 ± 0.0701	0.0401 ± 0.0376	2.8610^{-2}

in Figure 5.5.

Across all nine architectures, the integration of ORAF consistently reduces bias-related metrics while introducing a modest and predictable cost in recommendation accuracy. On the user-side, AUC drops from values close to 1.0 in the biased configurations to lower values after debiasing across all architectures, confirming that gender information becomes largely unrecoverable from the learned representations regardless of the underlying model type. The attribute substitution stability (CF) and group fairness (GF) metrics follow a similar pattern, with reductions observed across all models. On the item-side, REO and RSP both decrease substantially after debiasing, with most architectures moving from values in the range of 0.20-0.30 down to approximately 0.03-0.10.

Regarding the graph-based models specifically, ORAF successfully reduces AUC from its biased baseline value for both LightGCN and NGCF, confirming that the debiasing mechanism transfers to these structurally distinct architectures. However, the residual AUC values after debiasing remain somewhat higher for these graph-based models compared to the embedding-based architectures. This is consistent with the expectation that graph convolutional propagation, by aggregating neighborhood information across the interaction graph, provides an additional indirect channel through which sensitive attribute information can be encoded in the representations, making complete removal more challenging. Regarding recommendation quality, MAE and NDCG show modest changes

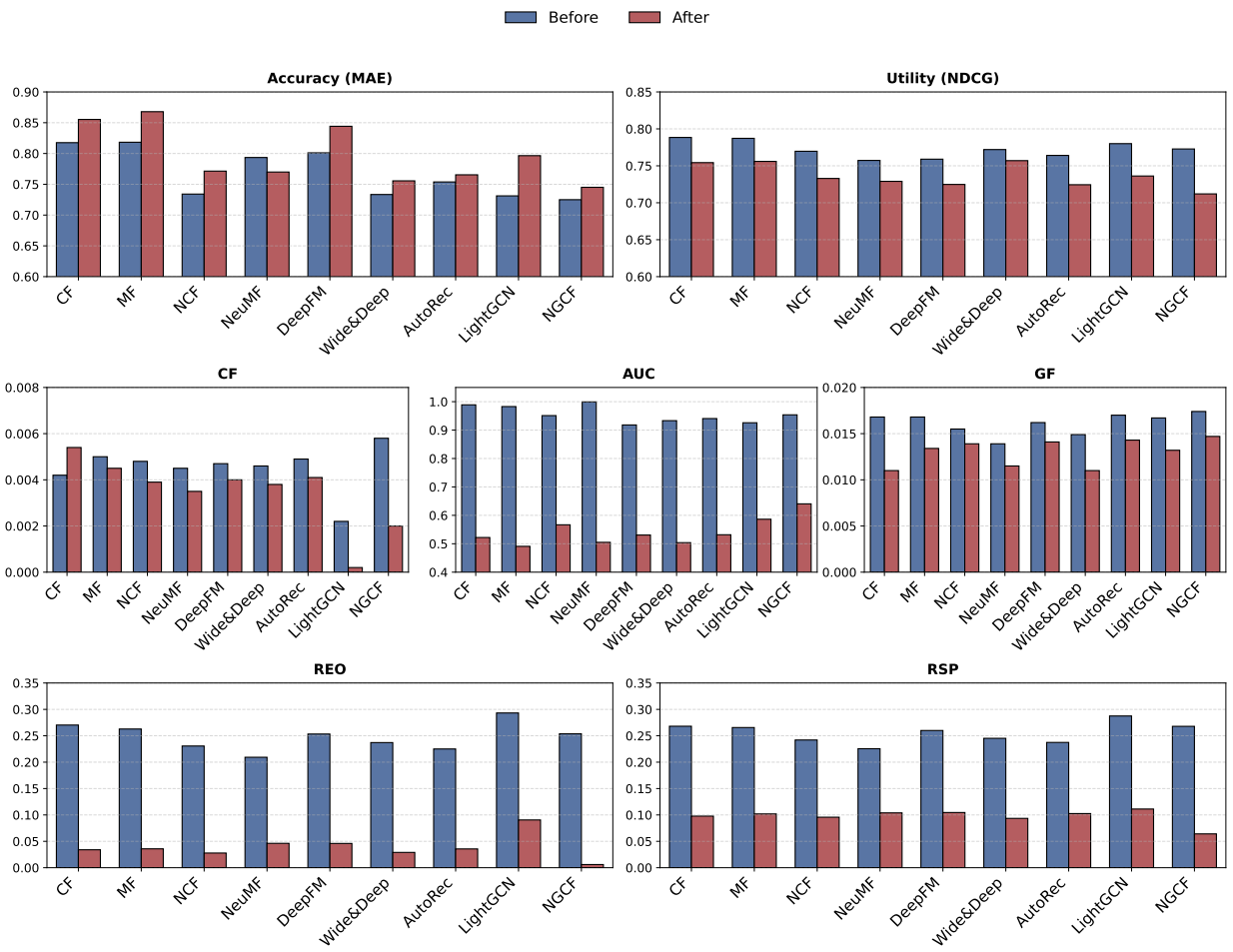


Figure 5.5: Performance comparison of the proposed ORAF framework across diverse recommendation architectures on the ML100k dataset

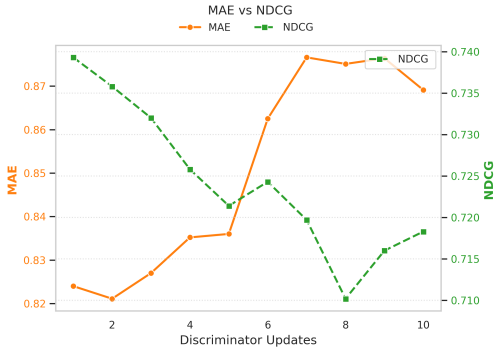
after debiasing that are comparable in magnitude across both embedding-based and graph-based models, suggesting that the fairness-accuracy trade-off does not substantially worsen when ORAF is applied to graph convolutional architectures.

These results provide reasonable evidence that ORAF's debiasing mechanism is broadly compatible with diverse recommendation architectures, including graph-based ones. We nonetheless moderate the claim of cross-architecture robustness: the degree of fairness improvement is architecture-dependent, and the interaction between graph convolutional propagation and the adversarial-orthogonal debiasing process warrants further investigation, which we identify as a direction for future work.

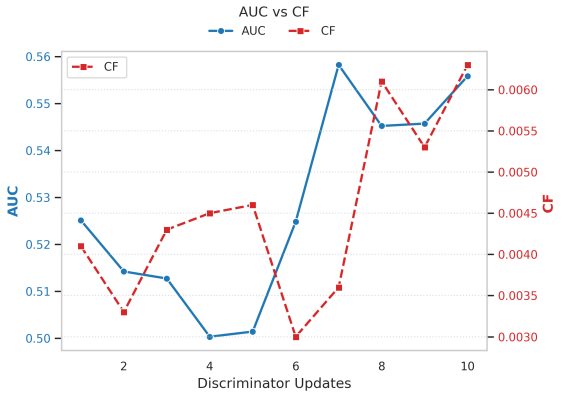
RQ4b: Ablation Study on Discriminator Update Frequency

To evaluate the stability of the minimax game inherent in our adversarial framework, we conducted a sensitivity analysis on the number of discriminator updates performed per generator step (n_{disc}),

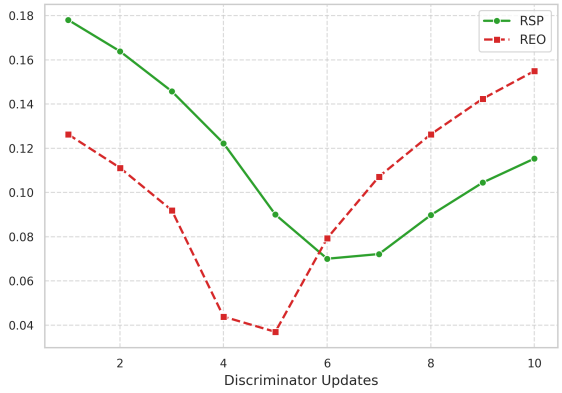
as illustrated in Figure 5.6. The training dynamics were analyzed across three distinct regimes:



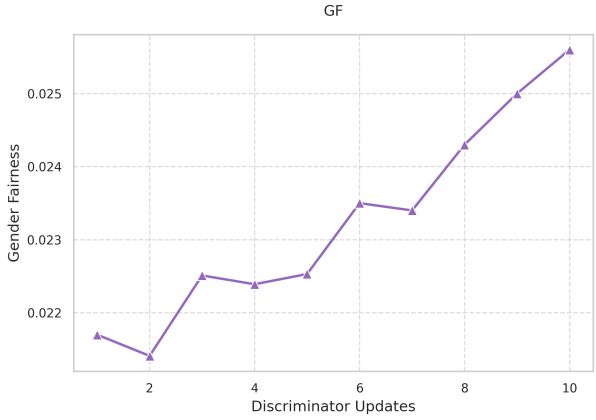
NDCG vs. MAE



AUC vs. CF



RSP vs. REO



GF

Figure 5.6: Impact of discriminator update frequency on fairness-utility trade-off

RQ4c: Ablation Study on Discriminator Update Frequency

To evaluate the stability of the minimax game inherent in our adversarial framework, we conducted a sensitivity analysis on the number of discriminator updates performed per generator step (n_{disc}). The training dynamics were analyzed across three distinct regimes:

In the initial range of 1 to 3 updates, the discriminators are *under-trained* relative to the main recommendation model. During this phase, the adversarial components do not update frequently enough to establish precise decision boundaries across the latent space. Consequently, the feedback signal provided to the encoder is too weak to enforce robust feature disentanglement. The generator effectively bypasses the adversarial penalty, retaining sensitive information within the embeddings to minimize reconstruction loss without being impeded by the sluggish discriminators, resulting in suboptimal fairness metrics. The interval between 4 and 5 updates marks the convergence of the

minimax game. Here, the discriminators are sufficiently robust to detect subtle traces of bias within the representations, yet they do not overpower the generator. This balance forces the encoder to aggressively orthogonalize the feature space to survive the combined adversarial loss. The results observe a global minimum in both the fairness gap and error rates in this region, suggesting that the discriminators successfully act as regularizers, stripping away spurious correlations and noise, thereby enhancing the model's generalizability.

Extending updates beyond 6 results in gradient domination, where the discriminators collectively overpower the generator. When the discriminators are updated too frequently, they converge to a near-perfect classification state before the generator can adapt. This creates a scenario where the generator faces vanishing gradients or conflicting optimization objectives, as it cannot simultaneously fool multiple highly optimized adversaries while maintaining utility. The consequent degradation in performance metrics indicates that the optimization landscape has destabilized, causing the model to lose the semantic integrity required for recommendation.

Consequently, we select $n_{disc} = 4$ as the optimal hyperparameter for the proposed framework, as it achieves the Pareto frontier, simultaneously minimizing the fairness gap and recommendation error while successfully orthogonalizing the bias-free and bias-aware embedding spaces.

RQ4d: Learning Rate Optimization for Multi-Objective Training

Figure 5.7 presents the ablation study characterizing the stability of the proposed framework under varying learning rate schedules. It evaluates the ORAF's performance across a grid of learning rates for the Recommender (η_{rec}), User Discriminator (η_{disc_u}), and Item Discriminator (η_{disc_i}). The tested range for all parameters encompasses values in $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

The magnitude of η_{rec} is the primary determinant of model stability. At lower values ($\eta_{rec} = 0.001$), the model successfully converges, achieving a Mean Absolute Error (MAE) of approximately 0.82. However, increasing η_{rec} leads to rapid divergence. As shown in the bottom rows of Figure 1, setting $\eta_{rec} \geq 0.1$ causes the optimization to fail, resulting in a substantial increase in error with MAE exceeding 1.14 (peaking at 1.55). In these divergent states, fairness metrics such as Counterfactual Fairness (CF) and Group Fairness (GF) flatten to near-zero, indicating a total loss of feature representation.

In the convergent regime ($\eta_{rec} = 0.001$), increasing the discriminator learning rates affects the accuracy-fairness trade-off. The results show that when η_{disc_u} or η_{disc_i} are increased significantly relative to η_{rec} (e.g., to 0.01 or 0.1), the MAE increases. This indicates that if the discriminators update too aggressively, they degrade the utility of the learned embeddings. Consequently, the lowest error rates are observed when η_{disc_u} and η_{disc_i} are kept small and synchronized with the recommender.

To maximize predictive accuracy while maintaining stable training dynamics, the following con-

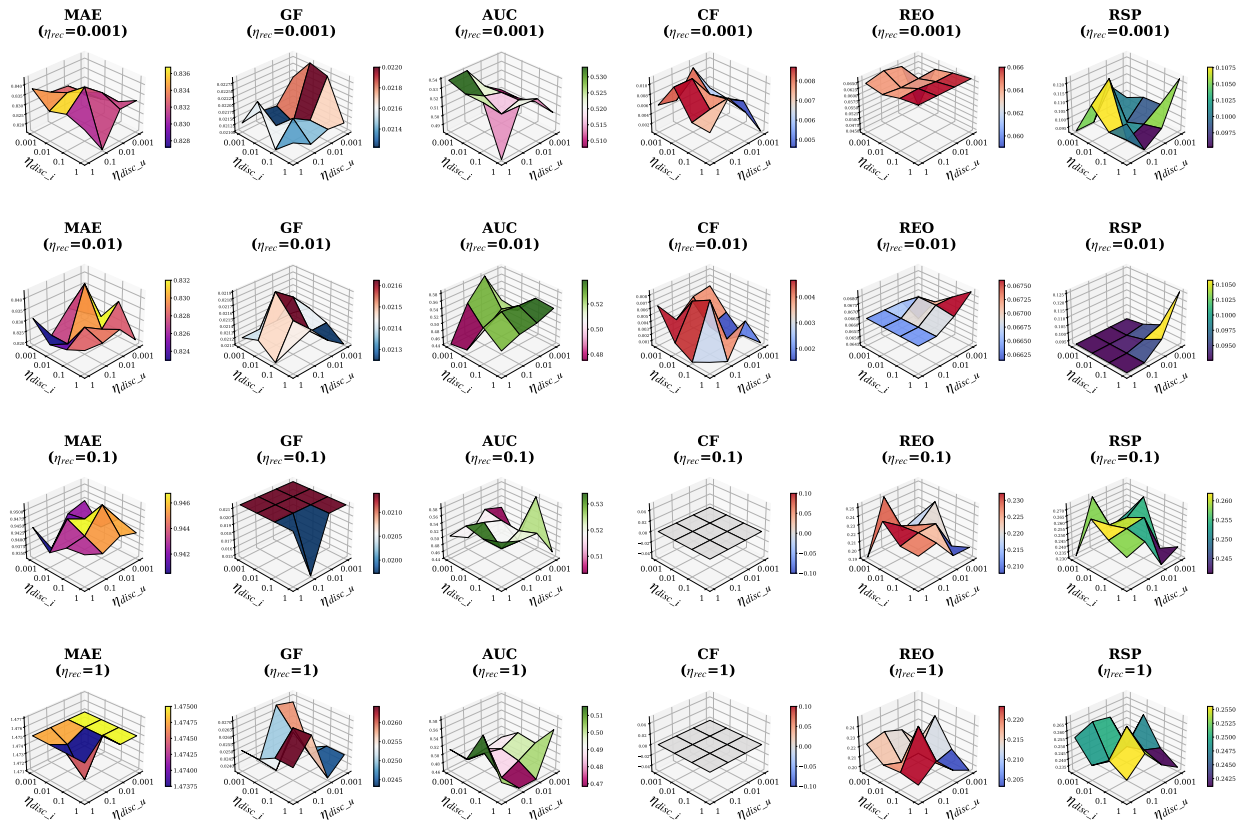


Figure 5.7: Learning rate optimization landscape for accuracy-fairness trade-offs on ML-100K dataset

figuration is used: $\eta_{rec} = 0.001$, $\eta_{disc_u} = 0.001$, $\eta_{disc_i} = 0.001$

RQ4-e: Regularization Sensitivity and the Fairness-Utility Trade-off

The second phase analyzed the sensitivity of the model to the orthogonal (λ_{orth}) and adversarial (λ_g, λ_p) regularization terms. Figure 5.8 illustrates the inherent tension between minimizing reconstruction error and satisfying user and item fairness constraints, revealing three distinct behaviors:

When the regularization weights are minimal ($\lambda \in \{0.001, 0.01\}$), the optimization is dominated by the primary recommendation task. While this yields the most accurate rating predictions (MAE ≈ 0.81), the adversarial unlearning is insufficient. Consequently, the user and item fairness metrics remain suboptimal, indicating the retention of sensitive bias in the latent space.

Increasing the magnitude of any regularization term to 1.0 effectively enforces strict fairness constraints. While this produces low bias scores, it imposes a severe penalty on the feature encoder. This "over-correction" degrades the semantic quality of the embeddings, causing the predictive error to deteriorate significantly (MAE > 0.85).

The empirical results identify the intermediate value of **0** as the optimal compromise. In this

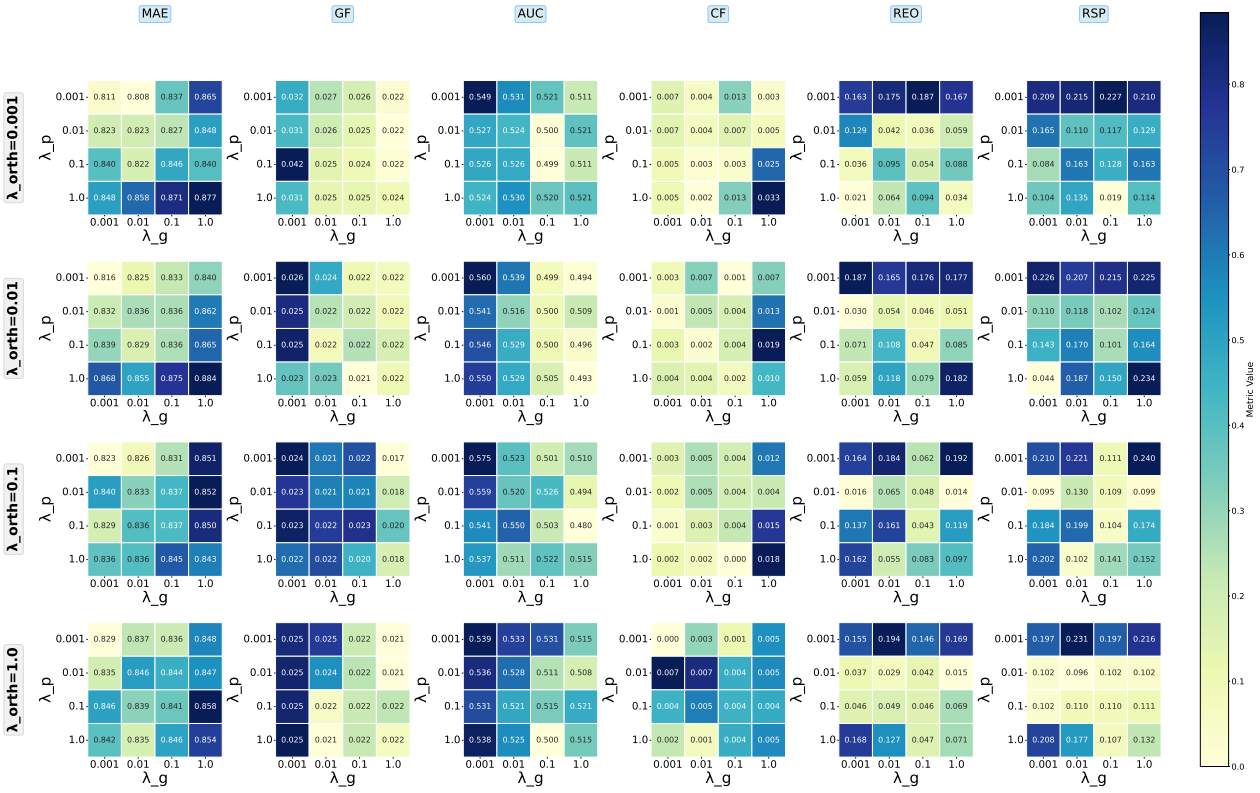


Figure 5.8: Hyperparameter sensitivity analysis

configuration, the system achieves a robust balance: the internal discriminator AUC converges to 0.503, confirming successful feature disentanglement. This leads to substantial fairness gains, stabilizing Group GF at 0.023 and CF at 0.004. Furthermore, popularity bias is significantly mitigated, with REO reduced to 0.046 and RSP to 0.101, all without the utility collapse observed at higher magnitudes ($MAE \approx 0.836$).

Based on this analysis, we select the symmetric configuration of $\lambda_{orth} = \lambda_g = \lambda_p = 0.1$ for all subsequent experiments

RQ4-f: Computational Complexity and Scalability Analysis

To empirically assess the practical deployability of ORAF, we conduct a computational complexity analysis comparing three progressive model variants across both benchmark datasets. **Base** is a standard recommender trained without any fairness mechanisms and serves as the computational baseline. **Base + Adv** extends the baseline with the dual-embedding structure and adversarial discriminators, while excluding orthogonal regularization. **Full ORAF** represents the complete proposed framework. All experiments were conducted on a server equipped with an NVIDIA A100 40GB GPU, an 8-core CPU, and 32 GB of RAM. The full complexity profile is reported in Table 5.11.

The introduction of the dual-path and adversarial components (**Base + Adv**) adds a fixed over-

Table 5.11: Computational complexity comparison across model variants.

Model	**Params (M)**	**Size (MB)**	**t/epoch (s)**	**Mem (MB)**	**Lat (ms)**
			ML-100K		
Base	0.152	0.579	3.916	41.5	0.636
Base + Adv	0.173	0.660	5.814	53.1	0.962
Full ORAF†	0.173	0.660	6.628	54.5	1.094
			ML-1M		
Base	0.494	1.883	39.256	49.2	0.651
Base + Adv	0.515	1.964	52.153	63.5	1.008
Full ORAF†	0.515	1.964	46.206	64.0	1.134

head of exactly 0.021M parameters, regardless of the dataset scale. Notably, progressing from **Base + Adv** to **Full ORAF** requires zero additional parameters (remaining at 0.173M for ML-100K and 0.515M for ML-1M). This is because the final components of ORAF, such as the orthogonality regularization, are purely mathematical constraints applied to the loss function during training, and require no new neural network layers. On ML-1M, Full ORAF represents only a 4.2% total parameter increase compared to **Base**.

Visual Analysis of Debiasing Effectiveness

To provide a rigorous and quantitative assessment of ORAF's effectiveness in removing sensitive information from learned representations, we complement the visual embedding analysis with three cluster separation metrics computed directly on the projected embeddings: the Silhouette Score (SS), the Davies-Bouldin Index (DBI), and the Inter-Centroid Distance (ICD). The Silhouette Score measures how well each point fits its own cluster relative to others, with values close to 1 indicating strong separation and values near 0 indicating overlap. The Davies-Bouldin Index measures the average ratio of within-cluster scatter to between-cluster separation, where lower values indicate better separation. The Inter-Centroid Distance measures the Euclidean distance between the centroids of the two sensitive attribute groups in the projected space, where a smaller value indicates that the two groups occupy closer and more overlapping regions. Together, these three metrics provide complementary quantitative evidence that is independent of the known perceptual limitations of dimensionality reduction techniques such as t-SNE, directly addressing the concern that visual conclusions may be subjective or distorted. We employ both linear (Principal Component

Analysis) and non-linear (t-SNE) dimensionality reduction techniques to project high-dimensional embeddings into 2D spaces. The complete visualization results are presented across Figures 5.9-5.11, with male/female users coded blue/red and popular/unpopular items coded yellow/purple, and all quantitative metrics are consolidated in Table 5.12.

The visualization of biased user representations reveals pronounced group segregation across all learned embedding spaces. In both ML-100K (Figure 5.10) and ML-1M (Figure 5.9), male (blue) and female (red) representations form clearly separated clusters in both PCA and t-SNE projections, with Silhouette Scores of 0.2200 and 0.1252 respectively and Inter-Centroid Distances of 1.2255 and 1.4266. This geometric separation confirms that the original learned representations encode strong demographic signals that make gender trivially predictable using simple linear classifiers.

After ORAF processing, both sets of representations undergo a substantial transformation toward group neutrality. The debiased embeddings show thorough mixing of blue and red representations throughout the projection space, with SS dropping to 0.0155 and 0.0181 and ICD collapsing to 0.3228 and 0.2202 on ML-100K and ML-1M respectively. The DBI increases markedly in both cases, from 1.9522 to 6.3091 on ML-100K and from 2.2799 to 9.2469 on ML-1M, further confirming that the two groups are no longer geometrically separable in the debiased space. The t-SNE projections corroborate this finding, with every local neighborhood containing balanced proportions of male and female representations after debiasing.

Parallel analysis of item representations reveals similar patterns for popularity bias. On ML-100K, bias-aware item embeddings show an SS of 0.2489 and an ICD of 1.2174, reflecting clear stratification between popular (yellow) and unpopular (purple) items. After debiasing, SS drops to -0.0104 and ICD collapses to 0.1887, with DBI rising from 1.6502 to 13.1255. A negative Silhouette Score indicates that points are on average more similar to points in the opposite group than to those in their own, reflecting near-complete interleaving rather than mere overlap. On ML-1M, item-side results follow the same pattern, with SS decreasing from 0.2646 to -0.0029 and ICD dropping from 1.7190 to 0.2810 (Figure 5.9).

The BookCrossing dataset (Figure 5.11), which uses age as the sensitive user attribute and presents the additional challenge of extreme data sparsity (99.98%), shows results broadly consistent with the other two datasets. User-side SS and ICD reduce from 0.1969 and 1.7051 to 0.0158 and 0.3165 respectively, with DBI rising from 1.7463 to 6.5680. The slightly higher residual SS compared to the movie datasets is consistent with the expected difficulty of achieving complete geometric separation under extreme sparsity. On the item side, the debiased embeddings achieve an ICD of 0.0652, which is the lowest value observed across all three datasets, with DBI rising substantially from 2.1480 to 15.2538, suggesting particularly strong popularity signal removal in this domain.

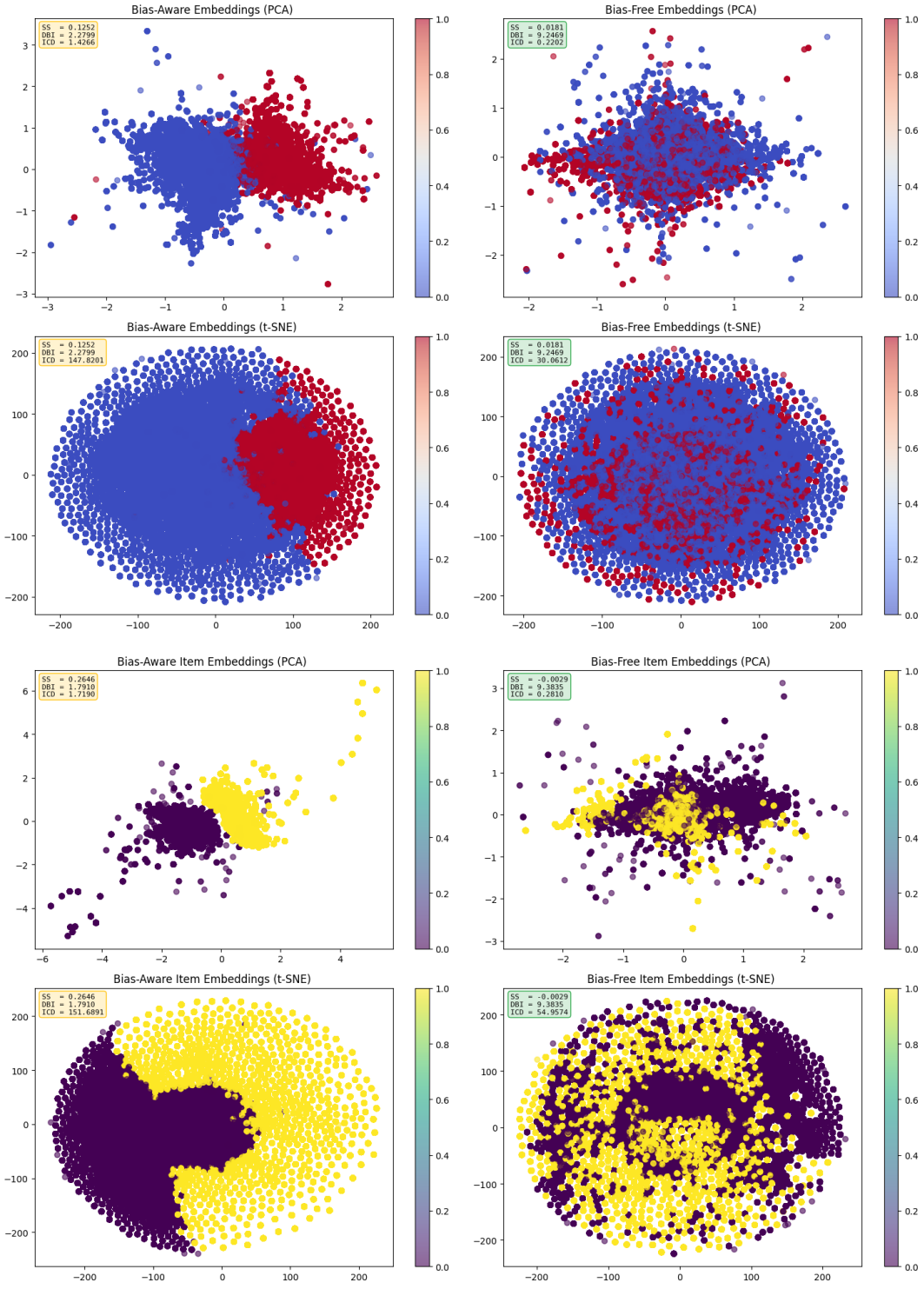


Figure 5.9: ML-1M dataset

RQ5: Real Study Case

To validate the effectiveness of our model ORAF in ensuring two-sided fairness and mitigating both gender and popularity bias, we conducted an experiment using a female user (ID 366) from

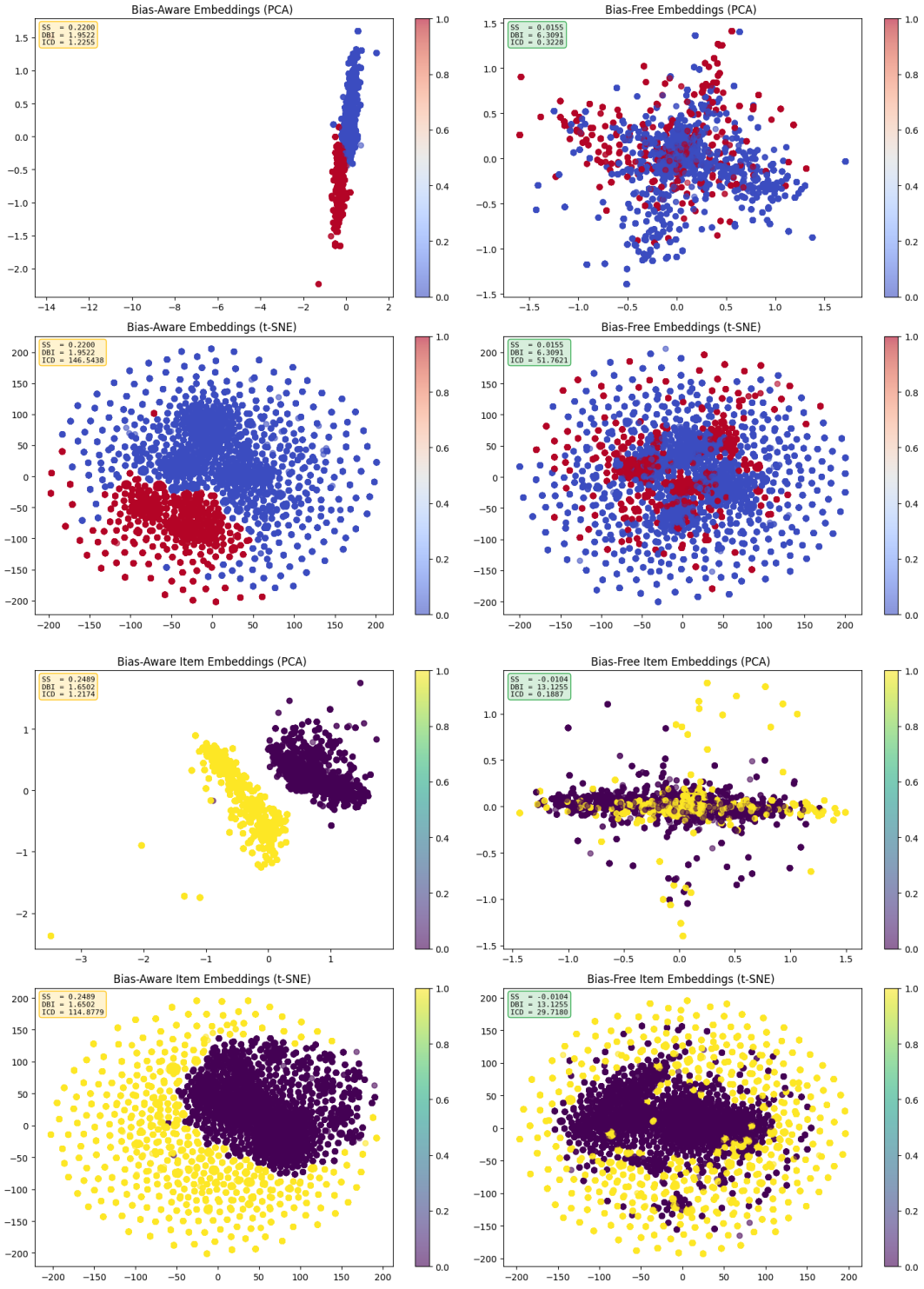


Figure 5.10: ML-100K dataset

the MovieLens 100K dataset, a female user whose historical preferences demonstrate a pronounced inclination toward male-associated genres. This user's preferences reveal a diverse range of movie genres, with a significant number of highly rated films typically perceived as male-oriented. The

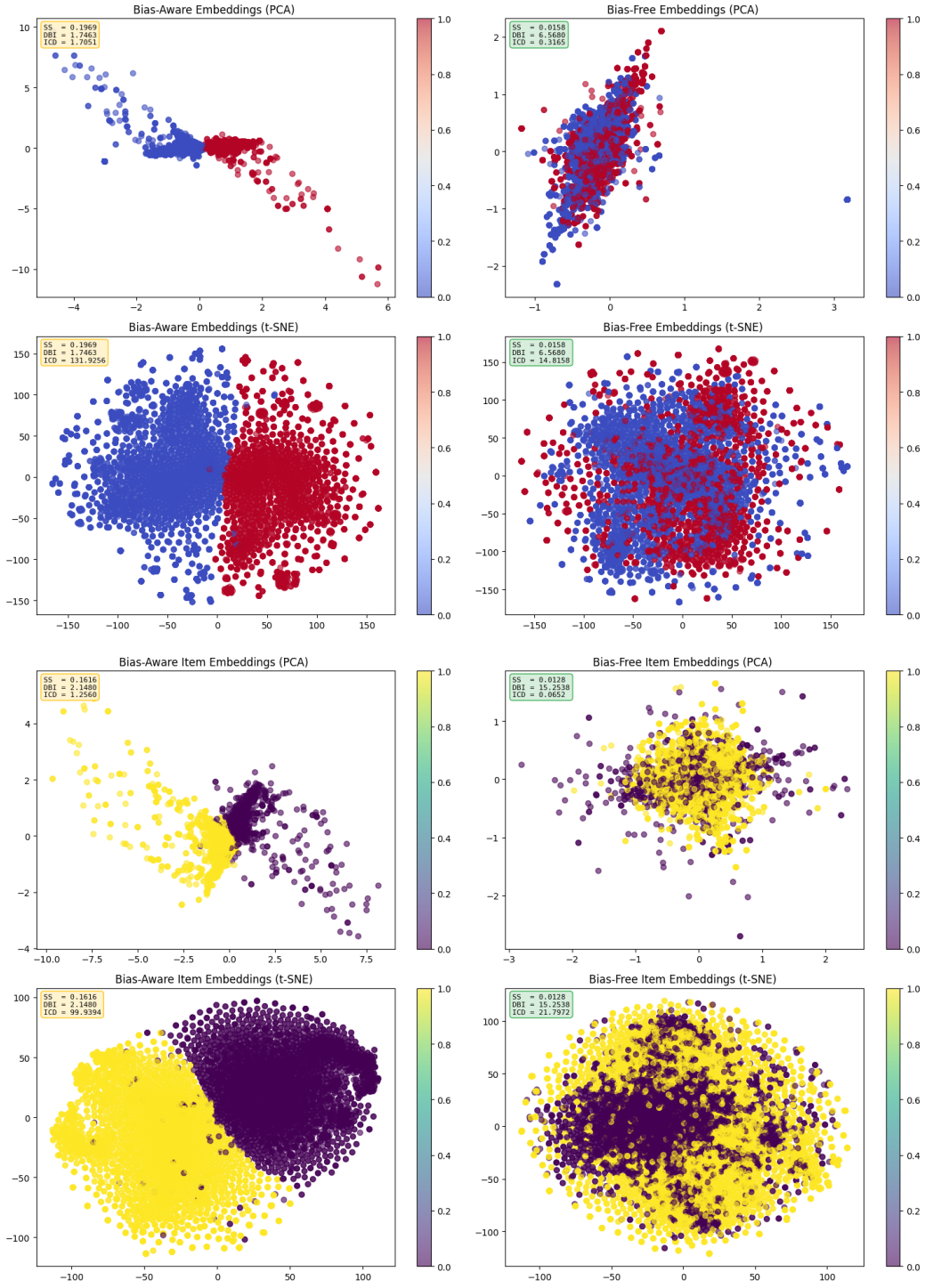


Figure 5.11: BookCrossing dataset

top-rated movies for this user, shown in the figures, demonstrate a strong preference for genres like Horror, Thriller, Comedy, Action, and Sci-Fi.

Our analysis compared the recommendation outputs between the standard BaseModel, which

Table 5.12: Quantitative embedding separation metrics for bias-aware and bias-free representations across all datasets.

Dataset	Embedding	SS	DBI	ICD
ML-100K	User Bias-Aware	0.2200	1.9522	1.2255
	User Bias-Free	0.0155	6.3091	0.3228
	Item Bias-Aware	0.2489	1.6502	1.2174
	Item Bias-Free	−0.0104	13.1255	0.1887
ML-1M	User Bias-Aware	0.1252	2.2799	1.4266
	User Bias-Free	0.0181	9.2469	0.2202
	Item Bias-Aware	0.2646	1.7910	1.7190
	Item Bias-Free	−0.0029	9.3835	0.2810
BookCrossing	User Bias-Aware	0.1969	1.7463	1.7051
	User Bias-Free	0.0158	6.5680	0.3165
	Item Bias-Aware	0.1616	2.1480	1.2560
	Item Bias-Free	0.0128	15.2538	0.0652

does not incorporate fairness constraints, and our fairness-aware ORAF framework (Per-user results are reported in Table A.1 and aggregated statistics are summarized in Table A.2 of the appendix .)

A closer look at the top 10 recommended movies by the BaseModel for User 366 reveals significant gender stereotyping, with 7 out of the 10 suggested movies belonging to the Drama and Romance genres. These genres, however, only represent 3.1% (Romance) and 7.7% (Drama) of the user's historical preferences, while genres like Horror (38.5%) and Thriller (16.9%) account for a larger portion of the user's ratings. This overrepresentation of "female-coded" genres in the recommendations aligns with stereotypical gender biases, reflecting common associations between these genres and female preferences in film consumption. This suggests that the model is relying more on broad demographic assumptions than on the individual user's specific tastes. Furthermore, the uniformity of these recommendations makes the user's gender easily identifiable from the genre distribution alone, highlighting that the model prioritizes demographic patterns over a deeper understanding of the user's diverse preferences. Upon examining the recommendation list, it becomes evident that the user's gender can be easily inferred as female, which underscores the model's reliance on demographic stereotypes rather than a nuanced understanding of the user's true

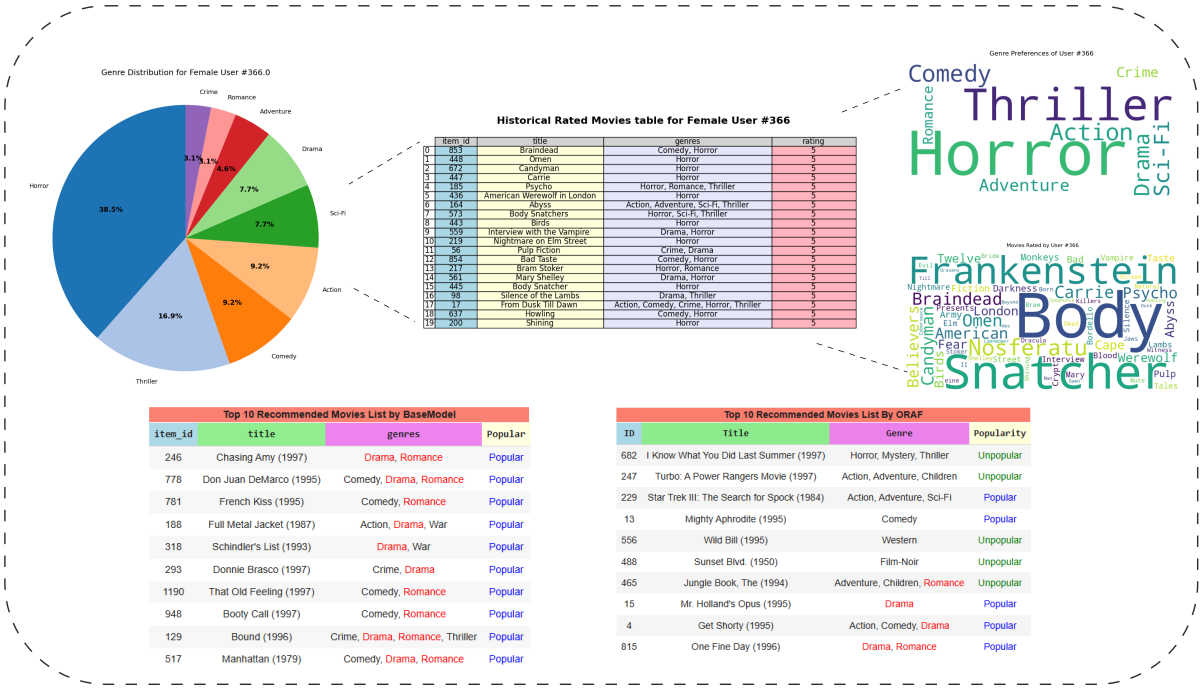


Figure 5.12: Study case of female user 366

preferences. Furthermore, all recommended movies are classified as "Popular", despite not aligning with the user's true preferences demonstrating the model's strong popularity bias and its failure to suggest niche titles that might better match the user's unique preferences.

In contrast, the recommendations generated by ORAF show no clear gender-based bias, resulting in a more diversified and user-specific output. The suggested titles cover a broader range of genres, including horror (I Know What You Did Last Summer), action (Star Trek III), western (Wild Bill), and film-noir (Sunset Blvd), with only 30% (3 out of 10) of the movies belonging to drama/romance categories. This variety in the recommended list makes it considerably more difficult to predict the user's gender based on the recommendations alone, thereby demonstrating that ORAF effectively decouples its recommendations from gender-based stereotypes.

ORAF, on the other hand, achieved a better balance, with 5 popular and 5 less popular movies appearing in its top 10. Crucially, most of these titles aligned with the user's genre preferences, encompassing both mainstream and niche selections. This confirms that ORAF successfully mitigates popularity bias while still catering to the user's true interests, offering a more nuanced and personalized recommendation.

5.8 Summary

We presented in this chapter ORAF, a novel orthogonal regularization-based adversarial framework designed to achieve comprehensive two-sided fairness in recommender systems. Our approach addressed a key limitation of existing methods that treat user and item fairness as independent objectives, often resulting in compensatory biases and uneven outcomes. By integrating dual adversarial discriminators with rigorous orthogonal constraints, ORAF simultaneously mitigated demographic biases in user representations and popularity biases in item recommendations.

The key innovation of our framework lies in its hybrid architecture, which combined the representation-level invariance of adversarial learning with the geometric guarantees of orthogonal regularization. This dual mechanism ensured clear disentanglement between sensitive attributes and recommendation-relevant features, providing mathematically grounded assurance of bias removal while preserving recommendation utility. Extensive experiments across three real-world datasets demonstrated that ORAF significantly outperformed state-of-the-art baselines in both fairness metrics and accuracy based measures.

General Conclusion and Future Perspectives

The proliferation of algorithmic decision-making in digital ecosystems has elevated recommender systems from mere engineering conveniences to powerful socio-technical arbiters. Modern recommender algorithms, spanning matrix factorization, deep neural networks, sequence-aware models, graph neural networks, and semantic content-based pipelines, act as primary algorithmic gatekeepers. They filter massive information streams and directly shape both consumer experiences and item provider livelihood.

However, conventional recommender architectures optimize almost exclusively for accuracy metrics. By treating historical interaction logs and behavioral data as ground-truth representations of user utility, these systems inherently inherit, amplify, and perpetuate systemic biases.

Addressing these challenges requires overcoming a fundamental multi-stakeholder dilemma. Algorithmic bias in recommendation operates along two distinct, frequently conflicting dimensions: the consumer-side, where recommendations can unfairly exploit sensitive demographic attributes (such as user gender, age, or ethnicity), creating disparate treatment, reinforcing echo chambers, or violating core privacy boundaries; and the provider-side, where long-tail items suffer from severe exposure stagnation due to rampant popularity bias, concentrating economic rewards onto a tiny fraction of dominant suppliers.

Simultaneously mitigating both consumer-side and provider-side biases presents a severe mathematical and structural bottleneck. Traditional approaches rely heavily on output-level interventions, such as post-hoc re-ranking or score-calibration algorithms. While computationally straightforward, these post-processing methods treat the underlying recommender as an immutable black box, attempting to restore fairness by structurally altering the final recommendation lists. This superficial correction inevitably destroys recommendation utility, introduces severe Pareto inefficiencies, and fails to purge the underlying latent representations of non-linear bias shortcuts.

This thesis is motivated by a foundational scientific conviction: truly robust, trustworthy, and multi-stakeholder algorithmic fairness cannot be achieved by retrofitting output predictions. Instead, fairness must be treated as an intrinsic structural property of the learned latent spaces

across recommender algorithms. By shifting the analytical focus from output correction to representation governance, this doctoral work investigates how joint adversarial optimization, explicit regularization, and geometric factorization can be leveraged to engineer inherently unbiased latent embeddings.

Thesis Overview and Scientific Motivation The central argument advanced throughout this research is that algorithmic fairness in recommender systems cannot be achieved solely through output-level correction mechanisms. Instead, fairness must be addressed as a latent representation learning and geometric structuring problem, where sensitive information is explicitly controlled within the representation space itself. This representation-centric theory of fairness constitutes the conceptual backbone of the dissertation and distinguishes the work from prior research that treats fairness as an external constraint applied post hoc.

The core scientific inquiry addresses whether fairness can be engineered directly into the representation space itself, ensuring that recommendations are fair by design rather than fair by correction. Relying on output modifications leaves the underlying latent representations untouched, allowing biased artifacts to persist in the embedding space and rendering any observed output equity fragile. Consequently, establishing an architectural ecosystem where fairness is treated as an intrinsic, structural property of the learned latent spaces becomes the paramount objective.

Summary of Core Contributions This dissertation delivers three principal methodological contributions, each addressing a distinct stage in the scientific evolution toward representation-centric fairness.

The first contribution, presented in Chapter 3, is DS-DVAE, a framework that establishes the principle of fairness by architectural design within deep generative architectures. DS-DVAE explicitly factorizes user and item embeddings into task-relevant preference subspaces and sensitive bias-associated subspaces. By optimizing an augmented Evidence Lower Bound integrated with a Kullback-Leibler divergence regularizer, the framework forces the model to compress sensitive attributes into an isolated channel. At inference time, recommendations are generated solely from the task subspace, ensuring that sensitive attributes have no path to influence the output, thereby guaranteeing two-sided fairness without requiring post-hoc corrections.

The second contribution, presented in Chapter 4, transitions the research into active optimization and geometric governance. FUIREC (Fairness-aware Unified framework with adversarial and Explicit Regularization for two-sided rECommendation) is a hybrid framework that deploys dual-adversarial discriminators to suppress consumer gender and provider popularity signals distributionally. To overcome the limitation that adversarial training alone leaves residual bias, FUIREC integrates explicit, differentiable regularization penalties directly into the loss function to mathematically bound demographic disparity and exposure variance, creating an exceptionally robust and

highly tunable fairness-utility trade-off.

The third contribution, presented in Chapter 5, reaches the logical conclusion of this research trajectory with ORAF (Orthogonal Regularization-Based Adversarial Framework). ORAF replaces soft structural constraints with a strict geometric mandate, enforcing geometric orthogonality between preference and bias embedding subspaces through a closed-form, differentiable orthogonality penalty. This step successfully transforms fairness from a statistical distribution-matching exercise into a verifiable, deterministic geometric property of the latent space.

Unified Theoretical Perspective: A Representation-Centric Theory of Fairness The individual frameworks presented in this dissertation provide the empirical validation of an overarching, unified scientific paradigm: The Representation-Centric Theory of Fairness. This theory posits that algorithmic bias in recommender systems is fundamentally a problem of latent representation corruption. In standard recommender algorithms, user and item embeddings are optimized solely for predictive accuracy. Because sensitive attributes correlate tightly with interaction patterns in most historical datasets, the optimization process implicitly encodes these attributes into the embeddings as computational shortcuts to predict user utility. Any downstream recommendation generated from such entangled embeddings inherits the demographic signal. Consequently, traditional fairness approaches that operate exclusively at the output level are fundamentally flawed; they merely mask the symptoms of bias while leaving the corrupted representation space intact, forcing an unnecessarily steep Pareto trade-off that severely damages recommendation accuracy.

The Representation-Centric Theory of Fairness completely flips this paradigm by asserting that fairness must be enforced as a structural, geometric constraint during the representation learning phase. Where traditional fairness perspectives isolate intervention at the output layer through ranking debiasing and score adjustments based on static metric targets, the representation-centric theory shifts the intervention locus directly to the latent space architecture through representation purification, dual-adversarial filtering, and geometric orthogonal separation. By pursuing joint structural optimization via augmented ELBOs, multi-stage MinMax games, and closed-form geometric regularizers, this paradigm treats latent space organization as highly malleable, governing fairness as an intrinsic architectural property. When the embedding generation phase is mathematically constrained, the model is forced to learn true preference markers that are statistically independent of sensitive shortcuts. This representation purification ensures that downstream prediction layers are physically incapable of exploiting biased heuristics, resulting in an optimal multi-stakeholder fairness-utility balance that output-level mechanisms cannot theoretically achieve.

This overarching theory rests upon three foundational principles. First, bias is encoded in the latent representation, not merely in the output, meaning that any intervention leaving the representation untouched cannot eliminate downstream demographic leakage. Second, statistical debiasing without geometric control leaves residual leakage; adversarial learning enforces that the

distribution of embeddings is approximately equal across sensitive groups, but a generator can easily hide sensitive information in non-linear ways that a stronger probe can recover. Third, geometric separation provides a verifiable fairness guarantee. Orthogonal regularization requires that debiased embeddings be strictly perpendicular to bias-aware embeddings. When this condition holds, sensitive information cannot be recovered from the debiased representation by any operation confined to that subspace, rendering the guarantee deterministic and verifiable at training time.

The three developed contributions represent a natural progression, each addressing a specific limitation of its predecessor. Stage one, DS-DVAE, establishes that fairness can be achieved by architectural design through explicit subspace factorization, creating fairness by construction rather than correction. The discovered limitation is practical: the factorized VAE requires careful tuning and assumes structural separability that may not hold under severe sparsity. Stage two, FUIREC, addresses these practical limitations through adversarial suppression rather than strict architectural separation, using a single embedding space rendered approximately invariant through a minimax game. The discovered limitation is that adversarial training alone leaves residual bias, as the generator learns to hide sensitive information in ways that fool the current discriminator but remain detectable to downstream probing attacks. Stage three, ORAF, resolves the residual bias problem completely by replacing output-level constraints with geometric orthogonal regularization, enforcing that the debiased embedding subspace must be geometrically perpendicular to the bias direction to provide a verifiable guarantee.

Synthesizing these insights, this dissertation proposes a comprehensive paradigm of latent fairness governance. Under this framework, recommendation fairness is governed through explicit control over the organization of the latent representation space, rather than through corrective interventions applied to outputs. This framework comprises three complementary mechanisms: subspace factorization (DS-DVAE), adversarial suppression (FUIREC), and geometric orthogonalization (ORAF). The key thesis-level claim is that any durable approach to fairness in recommendation must operate at the representation level.

Key Findings and Scientific Insights The experiments executed across this doctoral research yield several high-level scientific insights. The first insight is that fairness and recommendation utility are not strictly antagonistic; when fairness is integrated into the latent space as a structural constraint, the removal of spurious demographic correlations allows the model to learn more robust, generalized features that preserve or even enhance long-term utility. The second insight is that latent representations are the primary carriers of demographic leakage, confirming that output-level interventions leave the underlying system vulnerable to adversarial probing. The third insight is that disentanglement alone is insufficient when task-relevant features strongly correlate with sensitive attributes, as perfect separation becomes impossible without discarding predictive signal unless reinforced by active optimization. The fourth insight is that geometric orthogonal constraints sig-

nificantly improve fairness robustness, providing deterministic guarantees that survive sophisticated downstream attribute inference attacks. The fifth insight is that multi-stakeholder fairness is inherently multi-objective, meaning user-side and item-side fairness can structurally conflict, requiring explicit, unified loss functions to balance the feedback loops between consumer and provider vectors.

Limitations of the Thesis First, fairness is deeply contextual and non-universal; algorithmic fairness cannot be reduced to a single universal optimization objective. Different fairness definitions may conflict, and the appropriate notion depends heavily on the application domain, societal context, and shifting stakeholder priorities. A mathematical balance achieved within a latent vector space cannot fully capture the fluid, dynamic definitions of justice required across varying human communities. Second, metric incompatibility remains a fundamental challenge. Different fairness metrics may present conflicting narratives about the same system, meaning there is no single metric that captures all normative concerns, and metric selection itself implicitly encodes value judgments.

Data dependencies and optimization complexities introduce additional constraints. Fairness behavior varies significantly across datasets, making claims contingent on the underlying data distribution. Furthermore, the developed in-processing models assume that ground-truth sensitive demographic labels are fully available during offline training to guide the adversaries and regularizers; in real-world settings, these labels are frequently missing, incomplete, or corrupted due to strict privacy regulations, posing a major deployability constraint. Additionally, scalability and adversarial instability constrain industrial deployment. The integration of dual-adversarial discriminators in FUIREC and ORAF requires executing a complex MinMax optimization game that is highly sensitive to hyperparameter configurations, increasing computational overhead and time-to-convergence compared to standard pipelines. Finally, fairness generalization to unseen groups remains an open problem, as the frameworks assume that sensitive attribute labels are stable and represented during the training phase.

Future Research Directions To extend the representation-centric theory established in this dissertation, several ambitious directions extend beyond incremental improvements. Foundation models and generative recommendation represent a primary new frontier. The recommendation landscape is rapidly transitioning to generative paradigms powered by Large Language Models and multi-modal foundation architectures. Adapting representation-centric fairness to this setting is a priority, specifically investigating how token-level representations and self-attention matrices can be geometrically constrained to prevent generative models from propagating demographic stereotypes or hallucinating extreme popularity monopolies during item generation.

Simultaneously, integrating causal fairness in recommendation is essential to move beyond correlational patterns. A causal perspective would distinguish between legitimate and illegitimate sources of association using structural causal models and counterfactual reasoning, ensuring deep counterfactual fairness. To address privacy concerns, federated fairness learning represents a vital path forward; combining the hybrid adversarial mechanisms of FUIREC with decentralized learning would enable privacy-preserving fairness training across distributed edge devices without requiring centralized access to sensitive attributes.

Furthermore, explainable fair recommendation is essential for user trust, requiring explanation modules that highlight which features contributed to a prediction to ensure transparency in high-stakes domains. Future systems must also support dynamic fairness adaptation to handle temporal drift, extending regularizers to handle streaming data environments where item popularity lifecycles and user behaviors shift non-linearly over time. Finally, future work must expand into cross-cultural fairness studies to adapt latent geometric constraints to regional and cultural norms, while extending fairness beyond demographics to encompass multiple dimensions of socioeconomic, geographic, and behavioral bias.

Appendix A

1- Aggregated User-Level Deviation Analysis

To move beyond single-user illustration, we randomly sampled 20 users from the test set and evaluated both BaseRec and Full ORAF on an identical candidate item pool for each user, covering both user-side gender fairness and item-side popularity fairness at the individual recommendation list level. For each sampled user, we compute four metrics: the attribute substitution stability under gender flipping (ASS-Gen), which measures how much predicted scores change when the user gender attribute is substituted while all other inputs remain fixed; the attribute substitution stability under popularity flipping (ASS-Pop), which measures score sensitivity to item popularity label substitution; the top-10 recommendation overlap under gender substitution (Overlap-Gen), which counts how many items appear in both the original and the gender-substituted top-10 list; and the fraction of popular items in the top-10 (Pop%), which quantifies popularity concentration in the recommendation output. Per-user results are reported in Supplementary Table A.1 and aggregated statistics are summarized in Supplementary Table A.2.

User-side fairness. Under BaseRec, the mean ASS-Gen across the 20 users is 0.4300, with individual values ranging from 0.1799 to 0.8419, confirming that the baseline model encodes gender as a strong signal in its predicted scores. The top-10 overlap under gender substitution averages 5.65 out of 10, meaning that on average nearly half of the recommended items change when only the gender attribute is altered, a clear indication of gender-dependent recommendation behavior. Under Full ORAF, the mean ASS-Gen drops to 0.0440, an 89.8% reduction, and the top-10 overlap rises to 9.80 out of 10. Twelve of the 20 users achieve a perfect overlap of 10 out of 10 under ORAF, compared to none under BaseRec, indicating that for the majority of sampled users the ranked recommendation list is effectively identical regardless of the gender value supplied to the model.

Item-side fairness. Under BaseRec, the mean ASS-Pop is 0.5770, with individual values as high as 0.9094, confirming that predicted scores are strongly influenced by item popularity labels. The average fraction of popular items in the top-10 is 24.0%, with some users receiving up to 60% popular items, indicating a pronounced concentration of recommendations toward historically

popular content. Under Full ORAF, ASS-Pop drops to 0.0673, an 88.3% reduction, and the average top-10 popularity concentration falls to 9.5%, confirming that ORAF substantially increases the exposure of non-popular and niche items across all sampled users.

Consistency across users. The improvements are consistent across all 20 sampled users regardless of gender. Both male and female users show large reductions in ASS scores and large increases in top-10 overlap under ORAF, with no user subgroup showing a substantially weaker improvement. This individual-level consistency confirms that the debiasing effect operates uniformly across the population rather than being concentrated in a subset of users.

Table A.1: Per-user deviation analysis over 20 randomly sampled users from ML-100K.

User	**Gender**	**BaseRec**				**Full ORAF**			
		ASS-Gen	ASS-Pop	Overlap	Pop%	ASS-Gen	ASS-Pop	Overlap	Pop%
654	F	0.8419	0.3520	3	10%	0.0731	0.0646	10	20%
114	M	0.1915	0.7837	9	50%	0.0000	0.0769	10	20%
25	M	0.4940	0.7969	6	30%	0.0231	0.0667	10	10%
759	F	0.3890	0.2785	2	20%	0.0000	0.0698	10	20%
281	M	0.4804	0.7287	8	20%	0.0017	0.1021	10	0%
250	M	0.4678	0.9094	6	40%	0.0000	0.0941	10	10%
228	F	0.3956	0.7121	5	0%	0.2344	0.0908	9	0%
142	M	0.6167	0.5563	7	20%	0.0061	0.0278	10	10%
754	F	0.3422	0.4520	3	30%	0.0000	0.0644	10	0%
104	M	0.7030	0.8246	8	10%	0.0253	0.0600	10	0%
692	F	0.4607	0.6105	4	40%	0.0000	0.0786	10	20%
758	F	0.4566	0.3843	3	60%	0.0892	0.0576	10	0%
913	F	0.4215	0.3900	3	10%	0.0000	0.0986	10	0%
558	M	0.2954	0.3708	5	10%	0.1096	0.0397	10	0%
89	M	0.1799	0.3623	6	10%	0.0752	0.0496	9	20%
604	M	0.3097	0.7523	5	10%	0.0552	0.0697	9	20%
432	M	0.7921	0.6492	7	20%	0.0000	0.0942	10	0%
32	M	0.2069	0.4116	9	10%	0.0905	0.0343	10	0%
30	M	0.2823	0.7809	7	40%	0.0854	0.0532	9	30%
95	F	0.2728	0.4335	7	40%	0.0118	0.0532	10	10%
Avg		0.4300	0.5770	5.65	24.0%	0.0440	0.0673	9.80	9.5%

Table A.2: Aggregated deviation analysis over 20 randomly sampled users.

Model	**Avg ASS-Gen ↓**	**Avg ASS-Pop ↓**	**Avg Overlap ↑**	**Avg Pop% ↓**
BaseRec	0.4300	0.5770	5.65	24.0%
Full ORAF	0.0440	0.0673	9.80	9.5%
Reduction	*−89.8%*	*−88.3%*	*+73.5%*	*−60.4%*

Appendix B: Statistical Significance Test

To evaluate the individual contribution of each component of ORAF, we conduct a progressive ablation study in which components are added incrementally across four variants, evaluated over five independent runs on ML-100K with statistical significance assessed via paired t-tests across seeds. The four variants follow the same additive naming convention used throughout the chapter 5:

- **Base**: the recommendation baseline with no fairness constraints.
- **Base + Adv**: Base augmented with the dual adversarial discriminators for both gender and popularity.
- **Base + Adv + Orth**: Base + Adv augmented with the orthogonal regularization on both sides, but *without* the attribute predictors. The orthogonal regularization is therefore applied against bias-aware embeddings that receive no supervised signal to concentrate sensitive attribute information.
- **Full ORAF (Base + Adv + Orth + Pred)**: the complete framework, adding the attribute predictors that supervise the bias-aware subspaces to explicitly encode gender and popularity signals and give the orthogonal regularization a semantically meaningful anchor.

Results are reported in Table B.3, and two significance tests are reported for each metric. The column *p-val (vs Prev)* reports a paired t-test comparing each variant against the immediately preceding one, isolating the marginal contribution of the component added at that step. The column *p-val (vs Full)* reports a paired t-test comparing each variant directly against Full ORAF, confirming whether the complete model is statistically superior to each partial configuration.

Table B.3: Component-wise ablation study on ML-100K over five independent seeds. p-val (vs Full): paired t-test against Full ORAF. p-val (vs Prev): paired t-test against immediately preceding variant. $p < 0.05$ (significant). n.s.: $p \geq 0.05$ (not significant).

Variant	**Metric**	**p-val (vs Full)**	**p-val (vs Prev)**
Base	MAE	1.8310^{-7}	—
	NDCG	5.2510^{-5}	—
	AUC(LR)	1.3010^{-8}	—
	AUC(MLP)	1.1810^{-7}	—
	AUC(SVM)	5.6210^{-9}	—
	AUC(GB)	5.2810^{-9}	—
	ASS	2.2710^{-5}	—
	RSP	6.1810^{-4}	—
	REO	1.0410^{-4}	—
Base + Adv	MAE	1.8810^{-4}	2.9810^{-5}
	NDCG	7.8210^{-4}	6.0310^{-3}
	AUC(LR)	1.1610^{-5}	3.6510^{-8}
	AUC(MLP)	2.2810^{-5}	1.5510^{-6}
	AUC(SVM)	3.5610^{-5}	1.0110^{-5}
	AUC(GB)	2.8010^{-5}	2.8210^{-5}
	ASS	9.4710^{-3}	1.9310^{-5}
	RSP	6.6110^{-4}	1.7710^{-8}
	REO	3.3810^{-11}	4.0310^{-4}
Base + Adv + Orth	MAE	6.3310^{-3}	4.1310^{-2}
	NDCG	5.3610^{-3}	4.4310^{-2}
	AUC(LR)	9.3110^{-3}	4.2310^{-2}
	AUC(MLP)	2.9210^{-3}	8.2410^{-3}
	AUC(SVM)	3.5410^{-2}	6.4410^{-3}
	AUC(GB)	1.0110^{-3}	4.4410^{-2}
	ASS	7.5810^{-3}	5.7010^{-3}
	RSP	9.3410^{-3}	1.2210^{-2}
	REO	1.5010^{-2}	1.0610^{-2}
Full ORAF	MAE	—	6.3310^{-3}
	NDCG	—	5.3610^{-3}
	AUC(LR)	—	9.3110^{-4}
	AUC(MLP)	—	2.9210^{-4}
	AUC(SVM)	—	3.5410^{-3}
	AUC(GB)	—	1.0110^{-3}
	ASS	—	7.5810^{-3}
	RSP	—	9.3410^{-3}
	REO	—	1.5010^{-3}

Bibliography

[1] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43, 2023.

[2] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

[3] Yunqi Li, Michiharu Yamashita, Hanxiong Chen, Dongwon Lee, and Yongfeng Zhang. Fairness in job recommendation under quantity constraints. In *AAAI-23 Workshop on AI for Web Advertising*, volume 181, 2023.

[4] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions: Y. deldjoo et al. *User Modeling and User-Adapted Interaction*, 34(1):59–108, 2024.

[5] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–39, 2023.

[6] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. A survey on fairness-aware recommender systems. *Information Fusion*, 100:101906, 2023.

[7] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in recommender systems. In *Recommender systems handbook*, pages 679–707. Springer, 2012.

[8] Jiqun Liu. Toward a two-sided fairness framework in search and recommendation. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*, pages 1–11. ACM, 2023. doi: 10.1145/3576840.3578332.

[9] Filippo Carnovalini, Antonio Rodà, and Geraint A Wiggins. Popularity bias in recommender systems: The search for fairness in the long tail. *Information*, 16(2):151, 2025.

[10] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions: H. abdollahpouri et al. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.

[11] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. *ACM Transactions on Information Systems*, 41(2):1–29, 2022.

[12] 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, 2022.

[13] Xiangmeng Zhao, Lisi Chen, Jie Shao, Peng Han, and Jialie Shen. Investigating user-side fairness in outcome and process for multi-type sensitive attributes in recommendations. *ACM Transactions on Recommender Systems*, 2025. doi: 10.1145/3731568.

[14] Zhuoyang Xia, Meng Jian, Zihan Liu, Yulong Bai, Lifang Wu, and Shaona Wang. Mitigating long-tail bias in recommendations via graph diffusion. *Multimedia Systems*, 31(6):1–11, 2025.

[15] Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems: A. klimashevskaia et al. *User Modeling and User-Adapted Interaction*, 34(5):1777–1834, 2024.

[16] Jie Zhang, Sheng Wu, Tao Wang, Yang Li, and Wei Chen. Relieving popularity bias in recommendation via debiasing representation enhancement. *Complex and Intelligent Systems*, 11:34, 2025. doi: 10.1007/s40747-024-01649-z.

[17] Yao Wu, Jian Cao, and Guandong Xu. Fairness in recommender systems: evaluation approaches and assurance strategies. *ACM Transactions on Knowledge Discovery from Data*, 18 (1):1–37, 2023.

[18] Dominykas Seputis, Alexander Timans, and Rajeev Verma. Equity by design: Fairness-driven recommendation in heterogeneous two-sided markets. *arXiv preprint arXiv:2602.10739*, 2026.

[19] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems–survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227, 2018.

[20] Chuang Xia, Xu Shi, Haoran Xie, Yi Lu, Pan Li, and Mingsheng Shang. Beyond trade-offs: Leveraging spatiotemporal heterogeneity of user preference for long-term fairness and accuracy in interactive recommendation. *ACM Transactions on the Web*, 2025. doi: 10.1145/3769471.

[21] Liu Zhongzhou, Fang Yuan, and Wu Min. Mitigating popularity bias for users and items with fairness-centric adaptive recommendation. volume 41, pages 1–27. ACM, 2023. doi: 10.1145/3563783.

[22] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 11–21, 2018.

[23] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[24] Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):408–424, 2022.

[25] F. Ricci, L. Rokach, and B. Shapira, editors. *Recommender Systems Handbook*. Springer, 3rd edition, 2022.

[26] Yifan Zhang, Xu Chen, Yunfei Wang, Yueming Li, and Xiangyu Zhao. A survey on multi-stakeholder fairness in recommender systems. *ACM Computing Surveys*, 56(3):1–37, 2023.

[27] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:1–19, 2009.

[28] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[29] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in NeurIPS 20*, pages 1257–1264, 2008.

[30] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of UAI 2009*, pages 452–461, 2009.

[31] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[32] S. Rendle. Factorization machines. In *Proceedings of ICDM 2010*, pages 995–1000, 2010.

[33] S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1):1–38, 2019.

[34] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

[35] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[36] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie. AutoRec: Autoencoders meet collaborative filtering. In *Proceedings of WWW 2015*, pages 111–112, 2015.

[37] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 153–162, 2016.

[38] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698, 2018.

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[40] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proceedings of SIGIR 2002*, pages 253–260, 2002.

[41] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. *Advances in neural information processing systems*, 30, 2017.

[42] Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. A meta-learning perspective on cold-start recommendations for items. *Advances in neural information processing systems*, 30, 2017.

[43] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1073–1082, 2019.

[44] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. 2015.

[45] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018.

[46] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In

Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pages 1441–1450, 2019.

[47] X. Wang, X. He, M. Wang, F. Feng, and T. S. Chua. Neural graph collaborative filtering. In *Proceedings of SIGIR 2019*, pages 165–174, 2019.

[48] 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*, pages 639–648, 2020.

[49] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983, 2018.

[50] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5):102666, 2021.

[51] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 249–254, 2021.

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

[53] Shan Yao, Xianyi Huang, Lei Chen, and Zenglin Wu. Beyond parity: Fairness objectives for collaborative filtering. *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, pages 2009–2018, 2017.

[54] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In *Conference on fairness, accountability and transparency*, pages 187–201. PMLR, 2018.

[55] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[56] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

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

[58] Benjamin Marlin, Richard S. Zemel, Sam Roweis, and Malcolm Slaney. Collaborative filtering and the missing-at-random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 267–275, 2007.

[59] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1670–1679, 2016.

[60] Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th international conference on World Wide Web*, pages 951–961, 2016.

[61] Jin Huang, Harrie Oosterhuis, and Maarten De Rijke. It is different when items are older: Debiasing recommendations when selection bias and user preferences are dynamic. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 381–389, 2022.

[62] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 501–509, 2020.

[63] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[64] Sanjay Krishnan, Jay Patel, Michael J Franklin, and Ken Goldberg. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 137–144, 2014.

[65] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[66] Harald Steck. Item popularity and recommendation accuracy. pages 125–132, 2011.

[67] Yuying Zhao, Yu Wang, Yi Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[68] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 46–54, 2018.

[69] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the 32nd International FLAIRS Conference*, pages 413–418, 2019.

[70] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2021)*, pages 2439–2449, 2021.

[71] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 242–250, 2018.

[72] Andres Ferraro, Michael D Ekstrand, and Christine Bauer. It's not you, it's me: the impact of choice models and ranking strategies on gender imbalance in music recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 884–889, 2024.

[73] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA, 2017.

[74] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. Estimating position bias without intrusive interventions. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 474–482, 2019.

[75] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*, pages 224–232, 2018.

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

[77] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414, June 2018.

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

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

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

[81] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021 (WWW 2021)*, pages 2198–2208, 2021.

[82] Christine Bauer and Andres Ferraro. Strategies for mitigating artist gender bias in music recommendation: a simulation study. In *Proceedings of the Music Recommender Systems Workshop (Singapore)(MuRS 2023). Zenodo*, 2023.

[83] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Provider fairness across continents in collaborative recommender systems. *Information Processing & Management*, 59(1):102719, 2022.

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

[85] Basileal Imana, Aleksandra Korolova, and John Heidemann. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the web conference 2021*, pages 3767–3778, 2021.

[86] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*, pages 1334–1345. IEEE, 2019.

[87] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2221–2231, 2019.

[88] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[89] Shoujin Wang, Xiuzhen Zhang, Yan Wang, and Francesco Ricci. Trustworthy recommender systems. *ACM Transactions on Intelligent Systems and Technology*, 15(4):1–20, 2024.

[90] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education*, 32(3):636–684, 2022.

[91] 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*, pages 1808–1812, 2021.

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

[93] Fatemeh Nazary, Yashar Deldjoo, and Tommaso Di Noia. Chatgpt-healthprompt. harnessing the power of xai in prompt-based healthcare decision support using chatgpt. In *European Conference on Artificial Intelligence*, pages 382–397. Springer, 2023.

[94] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[95] Ludovico Boratto, Gianni Fenu, and Mirko Marras. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Modeling and User-Adapted Interaction*, 31(3):421–455, 2021.

[96] Lei Chen, Le Wu, Kun Zhang, Richang Hong, Defu Lian, Zhiqiang Zhang, Jun Zhou, and Meng Wang. Improving recommendation fairness via data augmentation. In *Proceedings of the ACM Web Conference 2023*, pages 1012–1020, 2023.

[97] Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In *International conference on machine learning*, pages 715–724. PMLR, 2019.

[98] Tianxin Wei and Jingrui He. Comprehensive fair meta-learned recommender system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1989–1999, 2022.

[99] Sheng Liu, Yi Zhang, Lichao Yi, Xianhua Deng, Lei T. Yang, and Bo Wang. Dual-side adversarial learning based fair recommendation for sensitive attribute filtering. *ACM Transactions on Knowledge Discovery from Data*, 18(4):1–20, 2024.

[100] Ludovico Boratto, Francesco Fabbri, Gianni Fenu, Mirko Marras, and Giacomo Medda. Counterfactual graph augmentation for consumer unfairness mitigation in recommender systems. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 3753–3757, 2023.

[101] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, July 2018.

[102] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, pages 1569–1578, November 2017.

[103] Saeedeh Karimi, Hossein A Rahmani, Mohammadmehdi Naghiaei, and Leila Safari. Provider fairness and beyond-accuracy trade-offs in recommender systems. *arXiv preprint arXiv:2309.04250*, 2023.

[104] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 42–46, 2017.

[105] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1791–1800, 2021.

[106] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021 (WWW 2021)*, pages 2980–2991, 2021.

[107] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys 2018)*, pages 154–162, 2018.

[108] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. FAR: A fairness-aware re-ranking algorithm for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems (RecSys)*, 2020.

[109] Robin Burke. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*, 2017.

[110] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference 2020*, pages 1194–1204, 2020.

[111] Qinyi Chen, Jason Cheuk Nam Liang, Negin Golrezaei, and Djallel Bouneffouf. Interpolating item and user fairness in multi-sided recommendations. *Advances in Neural Information Processing Systems*, 37:50189–50229, 2024.

[112] Yao Wu, Jian Cao, Guandong Xu, and Yong Tan. Tfrom: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1022, July 2021.

[113] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Yashar Deldjoo. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779, 2022.

[114] Guoli Wu, Zhiyong Feng, Shizhan Chen, Hongyue Wu, Xiao Xue, Jianmao Xiao, Guodong Fan, Hongqi Chen, and Jingyu Li. FairSort: Learning to fair rank for personalized recommendations in two-sided platforms. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[115] Yifan Wang, Peijie Sun, Weizhi Ma, Min Zhang, Yuan Zhang, Peng Jiang, and Shaoping Ma. Intersectional two-sided fairness in recommendation. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 3609–3620, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3589334.3645518. URL https://doi.org/10.1145/3589334.3645518.

[116] Qinyi Chen, Jason Cheuk Nam Liang, Negin Golrezaei, and Djallel Bouneffouf. Interpolating item and user fairness in multi-sided recommendations. *Advances in Neural Information Processing Systems*, 37:50189–50229, 2024.

[117] Sophie Greenwood, Sudalakshmee Chiniah, and Nikhil Garg. User-item fairness tradeoffs in recommendations. volume 37, pages 114236–114288, 2024.

[118] Chenyang Wang, Yankai Liu, Yuanqing Yu, Weizhi Ma, Min Zhang, Yiqun Liu, Haitao Zeng, Junlan Feng, and Chao Deng. Two-sided calibration for quality-aware responsible recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 223–233, 2023.

[119] Yiwen Wang and Lisa Singh. Impact on bias mitigation algorithms to variations in inferred sensitive attribute uncertainty. *Frontiers in Artificial Intelligence*, 8:1520330, 2025. doi: 10.3389/frai.2025.1520330.

[120] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[121] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2212–2220, 2019.

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

[123] Òscar Celma. Music recommendation and discovery in the long tail. In *Springer*, 2010.

[124] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, pages 689–698. ACM, 2018. doi: 10.1145/3178876.3186150.

[125] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. RecVAE: A new variational autoencoder for top-N recommendations with implicit feedback. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM '20)*, pages 528–536. ACM, 2020. doi: 10.1145/3336191.3371831.

[126] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[127] Noveen Sachdeva, Giuseppe Manco, Ettore Ritacco, and Vikram Pudi. Sequential variational autoencoders for collaborative filtering. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*, pages 600–608. ACM, 2019. doi: 10.1145/3289600.3291007.

[128] Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. Bilateral variational autoencoder for collaborative filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*, pages 292–300. ACM, 2021. doi: 10.1145/3437963.3441759.

[129] Yi Zhang, Yiwen Zhang, Dengcheng Yan, Shuiguang Deng, and Yun Yang. Revisiting graph-based recommender systems from the perspective of variational auto-encoder. *ACM Transactions on Information Systems*, 41(3):81:1–81:28, 2023. doi: 10.1145/3573385.

[130] Zhe Xie, Chengxuan Liu, Yichi Zhang, Hongtao Lu, Dong Wang, and Yue Ding. Adversarial and contrastive variational autoencoder for sequential recommendation. In *Proceedings of*

the Web Conference 2021 (WWW '21), pages 449–459. ACM, 2021. doi: 10.1145/3442381.3449873.

[131] Chuan He, Yongchao Liu, Qiang Li, Chuntao Hong, Wenliang Zhong, and Xin-Wei Yao. M^2VAE: Multi-modal multi-view variational autoencoder for cold-start item recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14811–14819, 2026. doi: 10.1609/aaai.v40i17.38501.

[132] Yaxiong Wu, Craig Macdonald, and Iadh Ounis. A hybrid conditional variational autoencoder model for personalised top-n recommendation. In *Proceedings of the 2020 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '20)*, pages 89–96. ACM, 2020. doi: 10.1145/3409256.3409835.

[133] Xiaopeng Li and James She. Collaborative variational autoencoder for recommender systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, pages 305–314. ACM, 2017. doi: 10.1145/3097983.3098077.

[134] Andrei Hagiu and Julian Wright. Multi-sided platforms. *International journal of industrial organization*, 43:162–174, 2015.

[135] Sabrina Karboua, Fouzi Harrag, Farid Meziane, and Amal Boutadjine. Mitigation of popularity bias in recommendation systems. In *Tunisian-Algerian Joint Conference on Applied Computing*, 2022.

[136] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization*, pages 119–129, 2021.

[137] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023.

[138] Hengchang Hu, Yiming Cao, Zhankui He, Samson Tan, and Min-Yen Kan. Automatic feature fairness in recommendation via adversaries. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 245–252, 2023.

[139] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

[140] Ramon Lopes, Rodrigo Alves, Antoine Ledent, Rodrygo LT Santos, and Marius Kloft. Recommendations with minimum exposure guarantees: A post-processing framework. *Expert Systems with Applications*, 236:121164, 2024.

[141] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. Challenging the long tail recommendation. *arXiv preprint arXiv:1205.6700*, 2012.

[142] Seung-Taek Park and Wei Chu. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 21–28, 2009.

[143] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1073–1082, 2019.

[144] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments (RMSE) at RecSys 2019*, volume 2440 of *CEUR Workshop Proceedings*, 2019. URL http://ceur-ws.org/Vol-2440/paper4.pdf.

[145] Zhongzhou Liu, Yuan Fang, and Min Wu. Mitigating popularity bias for users and items with fairness-centric adaptive recommendation. *ACM Transactions on Information Systems*, 41(3):1–27, 2023.

[146] Ziwei Zhu, Jianling Wang, and James Caverlee. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 449–458, 2020.

[147] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 21–30, 2021.

[148] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In *Proceedings of the web conference 2021*, pages 624–632, 2021.