

UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI



This is the Author's *accepted* manuscript version of the following contribution:

Ludovico Boratto, Gianni Fenu, Mirko Marras and Giacomo Medda, *Practical Perspectives of Consumer Fairness in Recommendation* in ***Information Processing and Management***, Volume 60 (2023), Issue 2, Article number 103208.

The publisher's version is available at:
https://doi.org/10.1016/j.ipm.2022.103208

When citing, please refer to the published version.

© 2022. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/

This full text was downloaded from UNICA IRIS https://iris.unica.it/

Practical Perspectives of Consumer Fairness in Recommendation

Ludovico Boratto, Gianni Fenu, Mirko Marras and Giacomo Medda

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale, 72, Cagliari, 09124, Italy

ARTICLE INFO

Keywords:
Recommender Systems
Consumer Fairness
Mitigation Procedure
Systematic Mapping
Evaluation Protocol
Reproducibility
Benchmark

Abstract

In recent years, there has been an increasing number of mitigation procedures against consumer unfairness in personalized rankings. However, the experimental protocols adopted so far for evaluating a mitigation procedure were often fundamentally different (e.g., with respect to the fairness definitions, data sets, data splits, and evaluation metrics) and limited to a narrow set of perspectives (e.g., focusing on a single demographic attribute and/or not reporting any analysis on efficiency). This situation makes it challenging for scientists to consciously decide which mitigation procedure better suits their practical setting. In this paper, we investigated the properties a given mitigation procedure against consumer unfairness should be evaluated on, to provide a more holistic view on its effectiveness. We first identified eight technical properties and evaluated the extent to which existing mitigation procedures against consumer unfairness met these properties, qualitatively and quantitatively (when possible), on two public data sets. Then, we outlined the main trends and open issues emerged from our multi-dimensional analysis and provided key practical recommendations for future research. The source code accompanying this paper is available at https://github.com/jackmedda/Perspective-C-Fairness-RecSys.

1. Introduction

With the large adoption of decision-support systems, humans' intervention has been increasingly supported by automated intelligence in various real world, high-stakes environments. Entrusting significant choices to intelligent systems unfortunately increases unawareness of how these technologies behave and why a certain outcome has been produced. In order for people to gain trust in artificial intelligence (AI), it is thus fundamental to account for the trustworthiness of these systems. It is now commonplace to see evidence on the harmful impacts of AI systems, highlighting the risks we are exposed to, not only as users playing with Web tools but also as human beings. Governments are hence establishing regulations to shape initiatives towards the social responsibility of AI technologies.

Human-Centered Artificial Intelligence (HCAI) is an emerging field related to the trustworthiness and safety of intelligent systems. A notable example of system based on HCAI is represented by recommender systems, where people are provided with personalized suggestions generated by a certain model (e.g., (Armentano, Monteserin, Berdun, Bongiorno and Coussirat, 2018; Mauro, Ardissono, Cocomazzi and Cena, 2022)). Prior studies have however shown that recommender systems often lead to discriminatory outcomes (e.g., Dinnissen and Bauer; Lesota, Melchiorre, Rekabsaz, Brandl, Kowald, Lex and Schedl (2021); Neidhardt and Sertkan (2022)). These phenomena can occur in the form of unfavorable outcomes, affecting the entity being ranked or the users the recommendations are targeted to (consumers) (Burke, Sonboli and Ordonez-Gauger, 2018; Chen, Dong, Wang, Feng, Wang and He, 2020; Abdollahpouri, Adomavicius, Burke, Guy, Jannach, Kamishima, Krasnodebski and Pizzato, 2020). An abundance of consumer fairness notions have been consequently proposed, along with procedures for unfairness mitigation.

Despite the growing interest in providing fair recommendations to consumers, the landscape is convoluted, with often diverging definitions. The fragmented conception of consumer fairness has led to unfairness mitigation procedures built on top of heterogeneous evaluation protocols. Indeed, most of the works have been limited to showing that the proposed mitigation was better than doing nothing. In this direction, Boratto, Fenu, Marras and Medda (2022) have recently reproduced mitigation procedures against consumer unfairness in personalized rankings, emphasizing that why, when, and how to apply a certain mitigation procedure, another, or all of them is still unclear. The current state of progress on consumer unfairness mitigation therefore calls for a discussion on what consumer fairness is and

ludovico.boratto@acm.org (L. Boratto); fenu@unica.it (G. Fenu); mirko.marras@acm.org (M. Marras); giacomo.medda@unica.it (G. Medda)

ORCID(s): 0000-0002-6053-3015 (L. Boratto); 0000-0003-4668-2476 (G. Fenu); 0000-0003-1989-6057 (M. Marras); 0000-0002-1300-1876 (G. Medda)

which properties a mitigation procedure against consumer unfairness should be evaluated on to let scientists select more consciously the procedure to apply according to the circumstances and conditions they experience.

In our study in this paper, we aim to build a common ground that can act as a basis for the evaluation of consumer unfairness mitigation procedures. To this end, we first collected the papers addressing consumer unfairness mitigation in the literature and categorized them according to the definition, mitigation, and assessment strategies for consumer (un)fairness they proposed, highlighting the main similarities and differences (Section 2). We reproduced the mitigation procedure introduced in each paper and defined a unified experimental protocol for their evaluation (Section 3). We then evaluated the reproduced mitigation procedures on the basis of the defined evaluation properties (Section 4). Finally, we discussed the current trends, open issues, and future research related to consumer fairness in recommendation, according to the findings of our analysis (Section 5). Our study in this paper builds upon the research methodology and findings outlined in Boratto et al. (2022) and provides the following novel contributions:

- We identify and define a set of eight technical properties (applicability, coherence, consistency, data robustness, reproducibility, scalability, trade-off management, transferability) a given mitigation procedure against consumer unfairness should meet for being effective in practice. This contribution extends Boratto et al. (2022)'s work from a conceptual perspective, since in the original paper no extensive discussion about the perception of fairness definition, mitigation, and assessment was provided, no taxonomy of properties was included, and the framework was limited to considering what we identified in this paper as consistency and reproducibility.
- We assessed the extent to which existing mitigation procedures against consumer unfairness meet the defined properties, qualitatively and quantitatively (when possible), on two public data sets. Although we adopted the same research method (mitigations, recommendation models, data sets, and evaluation metrics) of Boratto et al. (2022), the assessment in the original paper considered only consistency and reproducibility, not covering the other six properties; each required a wide range of new experiments (computation, results, and discussion).
- We summarize the main trends and open issues emerged from our multi-dimensional analysis and provide key practical elements that should be considered while designing mitigation procedures able to better meet the defined properties. Boratto et al. (2022)'s work was extended from this angle by providing a more holistic view of the current state on mitigation procedure development, touching on more properties and their interdependencies as well as contextualizing the findings with respect to the reproduced works more comprehensively.

2. Perception of the State of the Art

Before delving into the unified experimental protocol, in this section, we describe the process we followed for collecting papers that propose consumer unfairness mitigation procedures and then categorizing these procedures based on how unfairness was defined, mitigated, and operationally assessed.

2.1. Paper Collection Process

Mitigation procedures against consumer unfairness proposed so far in the literature were collected by scanning Information Retrieval conferences and workshops proceedings with high impact, including CIKM, ECIR, ECML-PKDD, FAccT, KDD, RecSys, SIGIR, UMAP, WSDM, WWW, and scientific journals edited by ACM, Elsevier, IEEE, and Springer. To filter the appears to analyze, we ran queries that included a technical term, "*Recommender System*" or "*Recommendation*", and a non-technical term, "*consumer fairness*" or "*User Fairness*". We then considered three selection criteria. First, the paper should focus on a top-n personalized recommendation task. Although other recommendation tasks exist (e.g., rating prediction), in many commercial systems, the users are presented with a personalized ranking of items, but the predicted rating values are not. Given our practical focus on personalization, we therefore targeted consumer unfairness mitigation on recommendation models aimed to find a few specific items supposed to be most appealing for the user, based on their interests. Second, one core key contribution should be a countermeasure against consumer unfairness. We therefore discarded papers providing a mere exploratory analysis to characterize an unfairness issue, which would not allow us to directly make practical action against such issue in the recommendation model. Third, the mitigation procedure should aim to reduce unfairness from the perspective of the end users that consume the recommendations. To this end, papers addressing unfairness against stakeholders other

Table 1
Consumer unfairness mitigation procedures reproduced in our study under a top-*n* recommendation task.

Paper	Year	Mitigation			Evaluation		
		Notion [1]	Groups [2]	Type [3]	Data Sets	Utility Metrics	Fairness Metrics
Burke et al. (2018)	2018	EQ	G	IN	ML1M	NDCG	CES
Ekstrand, Tian, Azpiazu, Ekstrand, Anuyah, McNeill and Pera (2018)	2018	EQ	G	PRE	ML1M-LFM1K	NDCG-MRR	DP
Tsintzou, Pitoura and Tsaparas (2019)	2019	EQ	G	POST	ML1M-SY	-	BD
Frisch, Léger and Grandvalet (2021)	2021	IND	G-A	IN	ML1M	NDCG	EPS-CHI
Li et al. (2021a)	2021	EQ	B	POST	AM	NDCG-F1	DP
Li, Chen, Xu, Ge and Zhang (2021b)	2021	IND	G-A-O-MS	IN	ML1M-INS	NDCG-HIT	AUC
Wu, Ma, Mitra, Diaz and Liu (2021b)	2021	EQ	G-A	IN	ML1M-LFM1K	RECALL-NDCG	DP
Wu, Wu, Wang, Huang and Xie (2021a)	2021	IND	G	IN	NEWS	AUC-NDCG-MRR	AUC-F1

[1] **Notion**: Equity (EQ), Independence (IND); [2] **Groups**: Gender (G), Age (A), Occupation (O), Country (C), Marital Status (MS).
[3] **Type**: Pre-Processing (PRE), In-Processing (IN), Post-Processing (POST).
[4] **Data Sets**: MovieLens 1M (ML1M), LastFM 1K (LFM1K), Amazon (AM), Synthetics (SY), Wu, Wu, Qi, Huang and Xie (2019) (NEWS).

than consumers were left out from our analysis. As a last step, we reproduced the mitigation procedure proposed in the filtered collection of papers, relying on the framework shared by Boratto et al. (2022) when possible[1].

2.2. Fairness Definition Perception

Fairness depends on context and application (Singh and Joachims (2018)), and consumer fairness follows the same principle. There is however no consensus on how to perceive unfairness from a consumer perspective in recommendation. Studies often consider different viewpoints to analyze, mitigate, and evaluate unfairness. Table 1 collects prior studies focused on consumer unfairness mitigation in a top-*n* recommendation task. The considered works were built upon a certain notion of fairness according to the representation of each demographic group in the training set (e.g., Ekstrand et al. (2018)), the way sensitive attributes were encoded inside the user and item latent representations (e.g., Li et al. (2021b); Wu et al. (2021a)), the error made by the recommender system on each demographic group during training (e.g., Burke et al. (2018); Frisch et al. (2021); Wu et al. (2021b)), or the quality of the list of items recommended to each user (e.g., Li et al. (2021a); Tsintzou et al. (2019)). Regardless of the pipeline step the fairness notion was situated and operationalized, such notions explored in prior studies mainly address the following two principles: equity of certain metric scores (e.g., the recommendation utility scores) between demographic groups (`EQ`); independence of a certain outcome (e.g., the predicted relevance scores or recommended lists) from the sensitive attribute (`IND`). Although Table 1 reports only the fairness notion the respective study focused on, we will analyze the reproduced works with respect to both fairness principles, to uncover possible interdependencies between them.

2.3. Unfairness Mitigation Perception

To lay the foundations for the perception of consumer fairness in the recommender system literature, our study considers those papers that proposed a mitigation procedure and not only an exploratory analysis focused on unfairness. Indeed, in addition to the considered fairness notions, prior works can also be categorized based on the envisioned unfairness mitigation. Mitigation procedures proposed so far fall into three main categories (Mehrabi, Morstatter, Saxena, Lerman and Galstyan (2021)): pre-processing (`PRE`), in-processing (`IN`), and post-processing (`POST`). The mitigation category a certain work focused on is reported in Table 1 under the column *Type*. Conversely, the strategies used to group consumers depending on a certain attribute (either protected or non-protected) are reported under the column *Groups* and mainly cover gender (`G`), age (`A`), occupation (`O`), country (`C`), marital status (`MS`), activeness (`B`).

Usually, the considered studies proposed mitigation procedures coherently built on top of the fairness notion they tackled. On one hand, we found that studies focusing on fairness from an `EQ` perspective usually perform a mitigation by balancing the representation of groups in the training set (e.g., Ekstrand et al. (2018)), optimizing the model such that it makes a comparable error across groups (e.g., Burke et al. (2018); Wu et al. (2021b)) or re-ranking items such that both recommendation utility and fairness are improved (e.g., Tsintzou et al. (2019); Li et al. (2021a)). On the other hand, studies that envisioned fairness from an `IND` perspective usually counter unfairness by decoupling the user and item latent representations from sensitive attribute information (e.g., Li et al. (2021b); Wu et al. (2021a)) or introducing independence guarantees between the sensitive attribute and the predicted relevance score (e.g., Frisch et al. (2021)). Therefore, a certain variability on the type of mitigation exists based on the underlying fairness principle.

[1] The Li, Chen, Fu, Ge and Zhang (2021a)'s mitigation procedure we reproduced included the technical corrections highlighted and then applied by Rahmani, Naghiaei, Dehghan and Aliannejadi (2022) to the original procedure.

2.4. Fairness Assessment Perception

From an evaluation perspective, prior works listed in Table 1 capitalized on a wide range of data sets, such as MovieLens 1M (`ML1M`), Last.FM 1K (`LFM1K`), insurance (`INS`), news articles (`NEWS`), synthetic interactions (`SY`), and Amazon (`AM`). These data sets were used to first train recommender systems and then compute both recommendation utility evaluation metrics and fairness evaluation metrics. Evaluation metrics for recommendation utility included Normalized Discounted Cumulative Gain (`NDCG`), F1 Score (`F1`), Area Under Curve (`AUC`), Mean Reciprocal Rank (`MRR`), and Hit Ratio (`HIT`). Conversely, fairness evaluation metrics included ϵ-fairness (`EPS`), Chi-Square Test (`CHI`), Demographic Parity (`DP`), Category Equity Score (`CES`), and Bias Disparity (`BD`). Except for Tsintzou et al. (2019), all studies adopted NDCG as a recommendation utility metric. Whereas, recommender system fairness was assessed in terms of parity (Ekstrand et al. (2018); Wu et al. (2021b); Li et al. (2021a)), the accuracy of a sensitive attribute predictor (Li et al. (2021b); Wu et al. (2021a)), the distribution of item categories in the recommended lists across groups (Burke et al. (2018)), the consistency ratio of bias in data and outcomes (Tsintzou et al. (2019)), the proportion of users with the same preference over any items pair across groups (Frisch et al. (2021)). The high variability directly comes from the often different definitions and mitigation procedures.

3. Research Methodology

In this section, we describe the experimental data sets, the considered sensitive attributes, the adopted recommendation models, and the unified evaluation protocol built upon the framework proposed in Boratto et al. (2022).

3.1. Experimental Data Sets

Exploring consumer fairness properties is still an open issue in the recommender system community, being the number of public data sets with consumer feedback and *protected attributes* limited. In order to be consistent with prior work in this field, we took into consideration a data set in case it provided at least one sensitive attribute for the consumers, included enough consumer feedback for recommender system training and testing ($\geq$ 200,000 interactions), and was used for experimentation in at least one reproduced paper. Based on these criteria, we selected the two public data sets reported in Table 2, namely ML1M (movies) and LFM1K (music)[2]. Given that unfairness mitigation procedures were often devised for binary labels, our study in this paper considered binarized protected attribute labels. Gender labels were already binary in both data sets. Conversely, we binarized age labels, originally reported in seven age ranges (ML1M) or as a continuous variable (LFM1K), such that the two groups included consecutive age ranges and had the most similar representation possible. Interactions of a user with the same artist in LFM1K were aggregated, using the number of plays of songs from that artist by that user as a proxy of the rating.

3.2. Recommendation Models

The range of recommendation models evaluated in prior work in terms of consumer unfairness was heterogeneous, since the authors often made arbitrary choices and no common and standardized protocol existed. Our study in this paper again focuses on recommendation models considered in at least one prior work. Based on this assumption, we specifically included different families of recommendation models, such as memory-based (`ItemKNN` from Ekstrand et al. (2018); `UserKNN` from Ekstrand et al. (2018)), matrix factorization (`BiasedMF` from Li et al. (2021a); `PMF` from Li et al. (2021a); `FunkSVD` from Ekstrand et al. (2018)), learning-to-rank (`NCF` from Li et al. (2021a), `LBM` from Frisch et al. (2021), `SLIM-U` from Burke et al. (2018)), and session-based (`STAMP` from Li et al. (2021a)). These recommendation models will be leveraged to investigate important consumer fairness properties, such as the consistency of a mitigation procedure applied to a recommendation model across different data sets and the transferability with respect to the mitigation procedure's ability of reducing consumer unfairness across different recommendation models.

3.3. Evaluation Protocol

For each data set, we first split consumer interactions in a training set and a testing set, including 20% of the interactions (the most recent ones if a timestamp was available, randomly selected otherwise) in the testing set and the remaining ones in the training set. Further 10% of the training set was put apart for validation. Given a data set, a protected attribute, a recommendation model, and a mitigation procedure, we ran a grid search for each recommendation model on the validation set, according to a pre-defined hyper-parameters grid (detailed in our source

[2]The data sets used in our experiments were chosen based on the predominant practice in prior works, since the goal of our study is to provide insights on the state-of-the-art mitigation procedures in commonly used evaluation setups (although, in our case, the setup is unified).

Table 2
The data sets with consumer's sensitive attributes included in our study.

Data Set	#Users	#Items	#Ratings	Sensitive Attributes
ML1M (Harper and Konstan, 2016)	6,040	3,952	1,000,209	Gender (M : 71.7%; F : 28.3%) Age (<35 : 56.6%; ≥ 35 : 43.4%)
LFM1K (Celma, 2010)	268	51,609	200,586	Gender (M : 57.8%; F : 42.2%) Age (<25 : 57.8%; ≥ 25 : 42.2%)

code repository), with and without that mitigation procedure. For each setup, we obtained the predicted relevance scores and used them for generating the recommendations following a *test items* methodology (Bellogín, Castells and Cantador, 2011). Specifically, the list of possible items to recommend included, for all users, all the items having a test rating by some user and no training rating by the target user. This setting unified the test conditions for all users. We then monitored the utility of top-n recommendations (using binary relevances) through the Normalized Discounted Cumulative Gain (NDCG[3]), selected due to its popularity in prior work (see Table 1). Unfairness between consumer groups was monitored from an equity (`EQ`) perspective in terms of NDCG Demographic Parity (DP), computed as the difference in NDCG between groups according to Mehrabi et al. (2021), and from an independence (`IND`) perspective by means of a Kolmogorov-Smirnov test (KS) on the predicted relevance scores, as also proposed by Kamishima, Akaho, Asoh and Sakuma (2018) for example. Our experiments specifically considered DP as the difference in NDCG between the majority group (male users and younger users in ML1M and LFM1K) and the minority group (female users and older users in ML1M and LFM1K) for each data set. Majority and minority groups were defined with respect to their representation in the data set. Hence, the closer to 0 DP is, the fairer the recommender system is from an `EQ` perspective. Similarly, the closer to 0 KS is, the more independent the outcomes are from the sensitive attribute (so the recommender system is fairer from an `IND` perspective). Experiments ran on a Ryzen7 machine, 32 GB RAM.

4. Mitigation Procedures Benchmark

The number of mitigation procedures against consumer unfairness under a top-n recommendation is increasing. However, our discussion in the previous section has shown that there is no common consensus on which properties a mitigation procedure should meet and be evaluated on to support the mitigation selection process. In this section, we therefore propose eight key properties to consider while evaluating a mitigation procedure offline, before moving it into practice (e.g., user studies or online experiments). Being recommender systems often powered by machine learning, these properties would be indeed applicable also to mitigation procedures for any machine-learning model, from a conceptual perspective. This generality would indeed allow to contextualize mitigation procedures for recommender systems with respect to those for other machine-learning models in the future, giving a high-level overview on unfairness mitigation across machine-learning applications. It is worth noticing that the specificity of these properties would be represented by the way they were operationalized and monitored practically in the recommendation scenario (experimental protocols and results reporting) in our study. For instance, the property concerning data robustness was operationalized by considering data imbalances pertaining to the popularity of items according to the user's interactions with them, which is peculiar to a recommendation scenario. The design process for these properties was based on the adopted practices in the current academic literature about mitigation procedures for recommender systems (when possible) or those for general machine-learning models for completeness. The following subsections introduce each property in detail, including motivation, definition, and experimental results on the reproduced works.

4.1. Applicability

A notable property a mitigation procedure would need to be evaluated on to support scientists' in their selection process is represented by the extent to which it can be technically applied to many recommendation models. As in the more general machine-learning field, mitigation procedures can be applied in pre-processing, by transforming the input data, in-processing, by constraining the training process of state-of-the-art models, and post-processing, by re-arranging the original predictions. Post-processing mitigation procedures can be applied to a wider range of models, being model-agnostic per definition. Although an in-processing approach could be extremely effective, it could be limitedly applicable to another model, especially from a different family. For instance, a mitigation procedure involving

[3]Considering Ferrante, Ferro and Fuhr (2021), even although NDCG is not on an interval scale, it is sufficient to evaluate the goodness of each mitigation. Findings based on this metric have been identical in similar experiments where F1 score, reported as an interval metric by Ferrante et al. (2021), has been used, when the length of the run is fixed.

knowledge about users' neighbors might be seamlessly applied to different neighborhood-based recommendation models, but not to deep-learning models not relying on such neighborhood concept. On the other hand, an upsampling strategy can be used to manipulate the training data to be fed to any model, and a re-ranking procedure would be able to act on the original predictions provided by any model. One specificity of this property with respect to the recommendation task is that even pre-processing procedures might not always be model-agnostic, since for instance an in-processing tailored for a pair-wise model would need to act on pairs of items (not single items). Similar considerations can be made for post-processing approaches, which might rely on assumptions about the original rankings (e.g., in knowledge-aware recommendation) a model was able to produce. The applicability property is defined as follows:

Property 1. *Applicability indicates the extent to which a mitigation procedure can be technically run on a wide range of different recommendation models without requiring any substantial change to the fundamental steps it is based on.*

Fairness is an objective that should be tackled regardless of the implementation of a recommendation model. Models used in the literature often belong to similar families, e.g., model- or memory-based, making it easier the development of mitigation procedures *applicable* to different models. The models' family is particularly important in the context of in-processing approaches, where aspects related to the implementation are relevant with respect to the range of models the mitigation is *applicable* to. Furthermore, the possible flexibility of an in-processing procedure could allow a broader range of models to benefit from it. This perspective can be envisioned for Burke et al. (2018)'s work, where the authors added a regularization term to the loss function of SLIM, a hybrid model, to balance the neighborhoods. Such term can be also used on a simpler memory-based model, e.g., UserKNN, by regularizing the weights of the neighborhoods. In addition, in-processing procedures like Frisch et al. (2021); Li et al. (2021b); Wu et al. (2021b,a) can be extended to different architectures by modifying some aspects that pertain to several families, e.g., multi-objective optimization and feature-independent user embeddings.

Other mitigation procedures are instead designed without considering in advance the recommendation model they will be applied to. Pre-processing approaches, such as data transformation or user representation balancing (Ekstrand et al., 2018), are examples that potentially have a very high *applicability* for top-*n* recommendation. Conversely, the *applicability* of post-processing approaches (Li et al., 2021a; Tsintzou et al., 2019) could depend on other aspects related to the adopted fairness notion and how this notion needs to be satisfied. Tsintzou et al. (2019) proposed an approach to reduce the amplification of input data biases over specific categories of items by a recommender system; without data sets including category labels, this mitigation could not be applicable to a model.

4.2. Coherence

Recent literature, in both the machine-learning field and the recommendation field more specifically, has proven that models can result in unfair outcomes for the users belonging to a certain demographic group. In some cases, mitigating unfairness for that demographic group might result in reversing the disparity towards the other group(s) (instead of merely reducing it for the originally disadvantaged groups). For instance, Boratto, Fenu and Marras (2021) showed that applying upsampling techniques to the training data and regularizing the training process, along with their combination, often leads to introducing disparities for the group the recommender system was originally performing the best on. A mitigation procedure would need to (at least) reduce the unfairness towards the originally disadvantaged group, without introducing any unfairness towards another group. This property is defined as follows:

Property 2. *Coherence indicates the extent to which a mitigation procedure tends to reduce the biased outcomes for the originally disadvantaged group, without reversing the disparate outcome towards the other group(s).*

To investigate this phenomenon, Tables 3-4 reported the demographic parity score and the Kolmogorov-Smirnov test score for the recommendation models originally considered in each paper, before and after applying the respective mitigation procedure. We were particularly interested in observing whether the sign of the fairness metric score changed between the two (reversed unfairness, so low coherence) or was the same (the same group was disadvantaged, so high coherence). The results show that some configurations reported a DP opposite in sign when a mitigation procedure was used, compared to the original recommendation model. A value of DP less than zero exhibits a bias towards the minority group, while the majority group is advantaged when DP is higher than 0.

A notable example that showed low coherence was SLIM-U when the mitigation of Burke et al. (2018) was applied on gender groups. When the original model was treated with the mitigation procedure, the outcomes advantaged male

Table 3
[Consistency - Gender Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Orig*) and after mitigating (*Mit*) for gender groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.084	^0.022	^0.028	^0.032	^0.115	0.320	0.301	*-0.060	^0.072	^0.006	^0.142
Frisch et al.	LBM	0.044	0.021	^0.006	^0.004	^0.013	^0.025	0.144	0.212	*-0.035	*-0.058	^0.120	^0.126
Li et al. (a)	BiasedMF	0.112	0.051	^0.017	**^0.001**	^0.035	**^0.006**	0.287	0.114	^-0.095	^-0.060	^0.012	**^0.001**
	NCF	0.117	0.057	^0.016	**^-0.001**	^0.022	**^0.006**	0.250	0.138	*-0.073	-0.026	^0.033	**^0.001**
	PMF	0.119	0.056	*0.013	^-0.002	^0.023	**^0.006**	0.200	0.071	*-0.062	-0.027	^0.010	**^0.001**
	STAMP	0.022	0.020	***0.003**	^0.003	**^0.006**	**^0.006**	0.160	0.113	-0.021	**0.002**	**^0.001**	**^0.001**
Ekstrand et al.	FunkSVD	0.018	0.015	^0.004	0.002	^0.027	^0.018	0.010	0.013	**-0.006**	-0.003	^0.107	^0.119
	ItemKNN	**0.140**	**0.134**	^0.038	^0.030	^0.030	^0.031	0.287	0.286	^-0.127	*-0.116	^0.019	^0.022
	UserKNN	0.137	0.131	^0.031	^0.024	^0.074	^0.052	**0.406**	**0.411**	^-0.110	^-0.106	^0.067	^0.067

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

users instead of female users. This suggests that biases are not always caused by the characteristics of the data set, and the way the model learns from the data is also relevant. In particular, due to the neighborhood balancing, the model learnt patterns on the basis of a different viewpoint on the data, which resulted in a higher recommendation utility for male users. Several experiments with Li et al. (2021a)'s procedure were characterized by this phenomenon as well. On ML1M, the mitigation procedure advantaged the minority group to a small extent when applied to NCF and PMF, while on LFM1K the bias slightly shifted towards male users for STAMP and towards younger users for BiasedMF and PMF. A low coherence was also reported for FunkSVD while considering age groups. However, the bias towards the demographic groups was not substantial, neither for the original model (bias towards younger user) nor for the mitigated one (bias towards older users).

Coherence was not satisfied in ML1M and in LFM1K by two and five models out of ten respectively, which raises the question whether the bias in the outcomes can be reduced by just re-arranging the training data. The recommendation model and mitigation procedure should be thoroughly examined to understand the causes that led to unfairness towards one or more demographic groups. We therefore encourage researchers to take into account this property, so as to devise mitigation procedures with a high coherence in future works addressing consumer unfairness.

4.3. Consistency

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

Property 3. *Consistency indicates the ability of a mitigation procedure to substantially reduce the model's unfairness according to the pursued fairness notion, given any data set and any consumer grouping method.*

Unified Evaluation. Tables 3-4 report the performance of the recommender systems, before (`Orig`) and after (`Mit`) unfairness was mitigated, in terms of recommendation utility (NDCG@10) and fairness (Demographic Parity and Kolmogorov-Smirnov test). We assumed that a mitigation procedure was consistent on a recommendation model when there was a significant decrease of unfairness compared to that of the original recommendation model.

Our results on ML1M show that the mitigation procedures were able to improve DP for all the models, except for SLIM-U on gender groups. On the other hand, KS was decreased only by `Li et al. (2021a)`[4] and `Ekstrand et al.`'s mitigation procedures (except for ItemKNN, where we observed a minimal increment). In LFM1K, unfairness

[4]Li et al. (2021a) mitigation extracts only the top 10 items that maximize utility and fairness, setting the other prediction scores to 0. KS is lower for this approach because most of the prediction scores are equal to 0.

Table 4
[Consistency - Age Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of recommendation models before (*Orig*) and after mitigating (*Mit*) for age groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.048	^0.022	^0.014	^0.009	^0.095	0.320	0.207	-0.026	^-0.145	^0.017	^0.082
Frisch et al.	LBM	0.044	0.042	^0.005	^0.006	^0.021	^0.027	0.144	0.213	-0.011	-0.021	^0.125	^0.152
Li et al. (a)	BiasedMF	0.112	0.051	^0.015	**^0.000**	^0.042	**^0.006**	0.287	0.111	*-0.079	0.010	^0.019	**^0.005**
	NCF	0.117	0.057	^0.018	^0.002	^0.029	**^0.006**	0.250	0.137	*-0.067	-0.015	^0.055	**^0.005**
	PMF	0.119	0.056	^0.020	^0.004	^0.027	**^0.006**	0.200	0.071	*-0.046	**0.004**	^0.008	**^0.005**
	STAMP	0.022	0.020	**0.000**	**^0.000**	**^0.006**	**^0.006**	0.160	0.113	-0.031	-0.008	**^0.005**	**^0.005**
Ekstrand et al.	FunkSVD	0.018	0.016	^0.008	^0.006	^0.029	^0.021	0.010	0.016	**0.002**	**-0.004**	^0.054	^0.047
	ItemKNN	**0.140**	**0.138**	^0.027	^0.024	^0.029	^0.033	0.287	0.269	0.010	0.020	^0.133	^0.118
	UserKNN	0.137	0.137	^0.028	^0.023	^0.060	^0.051	**0.406**	**0.397**	-0.023	-0.031	^0.036	^0.031

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

was mitigated between gender groups in terms of DP by all the mitigation procedures, with the exception of LBM and SLIM-U. Only `Li et al. (2021a)`'s method improved DP for age groups. KS was reduced by both `Li et al. (2021a)` and `Ekstrand et al.`'s mitigation procedures, the former for both sensitive attributes, while the latter only for age groups. Overall, `Li et al. (2021a)` was the only consistent mitigation procedure across data sets and sensitive attributes.

Original Evaluation. Even though Tables 3-4 present a complete overview of the *consistency* achieved by the considered mitigation procedures, NDCG DP and KS were often not monitored in the original papers. The two metrics we adopted monitor unfairness from two viewpoints not necessarily corresponding to the fairness notions addressed by the authors of the original papers. We were therefore interested in investigating the extent to which the consistency patterns observed under the fairness metrics adopted in our study were confirmed (or reversed) in case we monitored fairness through the metrics proposed in the original papers. To this end, we experimented with the three mitigation procedures whose original paper differs substantially from this perspective, with our framework.

Specifically, Li et al. (2021a)'s paper originally monitored demographic parity with respect to the F1 Score (F1@10) as a fairness evaluation metric. We therefore computed this metric under our evaluation protocol and reported the results in Table 5. It can be observed that no recommendation model, after mitigating through the above mitigation procedure, reported a lower unfairness estimate for all the settings. Therefore, in contrast to the patterns observed in the unified protocol above, the method was not *consistent* on any of the models used in the original work in terms of F1 DP. Conversely, Frisch et al. (2021)'s work original assessed fairness with the ϵ-fairness score. This score is closer to 0 if, for any two items[5], the proportion of users with the same preference is approximately the same in all the demographic groups. The results in Table 6 show a decrease in unfairness only on ML1M. No improvement was reported on LFM1K, making `Frisch et al.`'s mitigation procedure not *consistent* on LBM.

Finally, Burke et al. (2018)'s work monitored consumer unfairness using the category equity score (CES) across item categories as a proxy. For any item category, when the items distribution for that category in the recommendation lists is equal across demographic groups, CES is equal to 1 (the recommender system is considered fair). ML1M originally included the movie categories. On the other hand, LFM1K did not contain song categories, but just the MusicBrainz[6] ID of the artist. This ID was used to retrieve the genres from the artist profile in the platform (if available), assigned by sorting the artist's genres by vote and taking the most voted genres (more than one in case they had the same number of votes). Then, for any two genres A, B, A $\neq$ B, genre B was replaced by genre A if the name of genre A was a sub-string of the name of genre B (e.g., *death metal* was replaced by *metal*). Each box plot in Figure 1 represents the averaged CES distribution on a data set for a certain attribute, before and after applying Burke et al. (2018)'s mitigation procedure. It can be observed that the CES obtained after mitigating had a higher variance and the average score was less close to 1 (and so the recommender system was less fair), compared to the original model's CES. Overall, this mitigation procedure would not be considered *consistent* on SLIM-U according to our definition.

[5] A maximum of 5000 unique combinations of items have been used to reduce execution time.
[6] `https://musicbrainz.org/`

Table 5
[Consistency] Recommendation utility (*F1* score, the higher it is, the more useful the recommendations) and equity (F1 Demographic Parity - *DP*, the closer to zero it is, the fairer the model) assessment for the recommendation models before (*Orig*) and after (*Mit*) Li et al. (2021a)'s mitigation. These two metrics were those used in the original paper.

	Gender								Age							
	ML1M				LFM1K				ML1M				LFM1K			
	F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$		F1 ↑		DP $\downarrow_0$	
Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
BiasedMF	0.057	0.024	*0.003	-0.003	0.043	0.018	^-0.014	^-0.006	0.057	0.024	*0.001	-0.002	0.043	0.017	0.002	^0.009
NCF	0.056	0.026	*0.004	-0.003	0.035	0.018	^-0.008	-0.002	0.056	0.026	^0.003	-0.001	0.035	0.018	0.003	0.004
PMF	0.058	0.026	0.000	*-0.004	0.029	0.011	^-0.007	-0.002	0.058	0.026	*0.002	-0.001	0.029	0.010	0.003	0.006
STAMP	0.008	0.008	*0.001	*0.001	0.021	0.014	-0.002	0.000	0.008	0.008	-0.001	-0.001	0.021	0.014	0.002	0.004

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

Table 6
[Consistency] Fairness assessment of recommendation models before (*Orig*) and after (*Mit*) Frisch et al. (2021)'s mitigation, on the metric used in the original paper, i.e., ϵ-fairness (the closer to zero it is, the fairer the model).

	Gender				Age			
	ML1M		LFM1K		ML1M		LFM1K	
Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
LBM	0.022	0.012	0.009	0.009	0.094	0.092	0.010	0.010

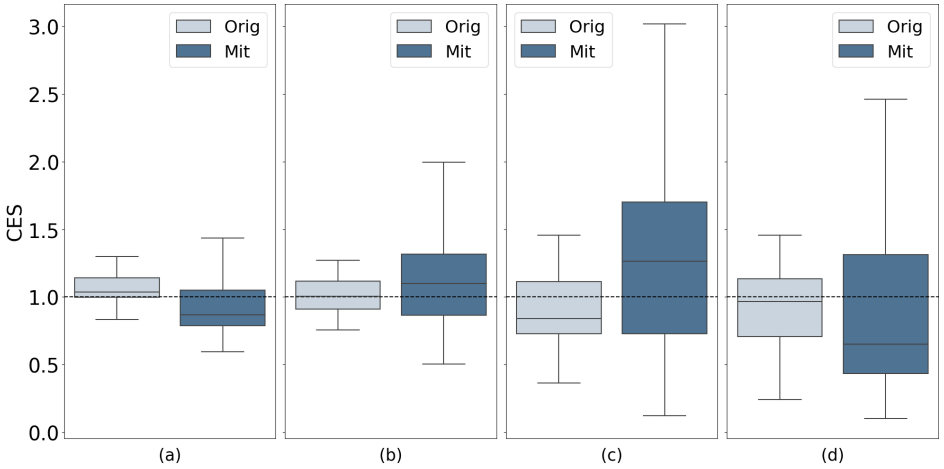


Figure 1: [Consistency] Category equity score distribution across item categories (*CES*, the closer to 1 it is, the more similar the category representation between the interactions and the recommendations) before (*Orig*) and after (*Mit*) Burke et al. (2018)'s mitigation under for the following combinations of data sets and demographic groups: (a) ML1M - Gender, (b) ML1M - Age, (c) LFM1K - Gender, (d) LFM1K - Age.

4.4. Data Robustness

User-item interaction data sets collected from online platforms are usually characterized by imbalances with respect to the collected feedback (e.g., due to the popularity bias recommender systems often emphasize) and the stakeholders influenced by the provided recommendations (e.g., under-represented provider or consumer groups). Furthermore, depending on the domain, there might be features having a causal relationship with consumer unfairness. For being practically successful, mitigation procedures should be able to deal with data characterized by (even extremely) imbalances and evident causal relationships. Inspecting robustness with respect to the data would be especially important when mitigation procedures do not report good results. Indeed, given a data set and a mitigation procedure which shows a limited consistency, it could be better to analyze the underlying data, explain the causes behind unfairness from a data perspective, and inform the mitigation procedure about them. Data robustness is therefore defined as follows:

Property 4. *Data robustness indicates the ability of a mitigation procedure to reduce unfairness also in challenging cases related to data distribution (e.g., imbalances) and relationships between unfairness and other features.*

To investigate data robustness, we analyzed the correlation between the satisfaction of the users and the distribution of popularity over items. For every row, Figure 2 shows three sub-plots, with each tick of the x-axis representing a group of items with similar popularity. The groups were formed by taking 1,000 consecutive items from an item list sorted by decreasing popularity. Considering only the interactions in the training set, each heatmap reports the difference in percentage concerning a certain metric (NDCG, NDCG DP, and KS), computed between the majority and the minority group. Each cell in sub-plots (a)-(d) represents the difference in percentage of the users who interacted with the respective group of items, (b)-(e) show the difference in percentage of users who received the items in the respective group as the top-*10* recommendations, finally (c)-(f) collect the difference in percentage of users satisfied by the items in the respective group in the top-*10* recommendations. We consider a user as satisfied from receiving an item as a recommendation if that item was included in the recommended list, and the testing set of that user included that item.

Sub-plots (a)-(b)-(c) show an analysis on gender groups by considering the predictions of UserKNN on ML1M and the respective mitigation procedure of Ekstrand et al. (2018). In particular, this method balances the consumer group representation in the training set, which leads to a more balanced interactions distribution for each demographic group over items. This aspect was highlighted by the lower differences in sub-plot (a). Sub-plot (b) reveals that also the items distribution in the recommendations was more balanced: the cells are darker for the mitigation with respect to the original model, which reported a bias towards male consumers. Ekstrand et al. (2018)'s mitigation on UserKNN successfully reduced unfairness. Sub-plot (c) reports similar satisfaction levels between male and female consumers, depicted as a lower intensity within the respective cells, compared to the ones associated with the original model. These results show that Ekstrand et al. (2018)'s procedure satisfies *data robustness* for ML1M on gender groups. The method generates fairer recommendations by taking into consideration also the popularity bias as a possible causal reason of the unfairness between gender groups.

Sub-plots (d)-(e)-(f) analyze the UserKNN predictions on LFM1K, when Ekstrand et al. (2018)'s mitigation was applied for age groups. UserKNN did not seem to be significantly affected by the mitigation method, when item popularity was considered. Instead, the balancing leads to an higher percentage of interactions for the minority group (sub-plot (d)), and so a more equal items distribution in the recommendations across demographic groups (sub-plot (e)). However, the satisfaction level switched from the majority to the minority group for the most popular items (sub-plot (f)). Interestingly, some item groups recommended by the original model satisfied more the minority group. This suggests that non-sensitive variables other than item popularity could affect model's unfairness. Hence, Ekstrand et al. (2018)'s mitigation did not satisfy *data robustness* on LFM1K for age groups.

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

4.5. Reproducibility

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

Property 5. *Reproducibility indicates the ability of taking the original source code that implements a mitigation procedure and being able to execute it under the same or a different evaluation protocol, with respect to the one used in the original paper.*

To inspect the *reproducibility* of the considered mitigation procedures, we leveraged the findings presented by Boratto et al. (2022), but focusing only on the papers that worked on a top-*n* recommendation task. The analysis reported that only 50% (4/8) of the papers could be reproduced. The other works were marked as non-reproducible if the source code was not publicly shared or additional requirements were not met.

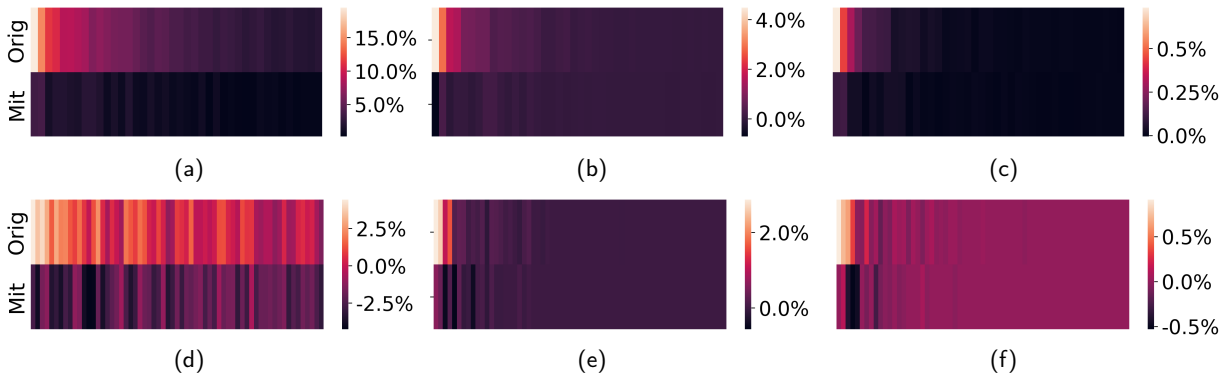


Figure 2: [Data Robustness] User interaction, recommendation, and relevant recommendations drift across item groups formed based on their popularity. The results were obtained by mitigating unfairness through Ekstrand et al. (2018)'s procedure on the original UserKNN recommendation model, on ML1M for gender groups (a, b, c) and on LFM1K for age groups (d, e, f). Each tick of the x-axis represents a group of 1,000 items with similar popularity, with groups formed by sorting items based on decreasing popularity and taking disjoint sets of 1,000 consecutive items for each group. The sub-plots (a, d) represent the percentage drift of the users who interacted with that group of items, (b, e) represent the percentage drift of users who received the items in that group within the top-*10* recommendations, and (c, f) represent the percentage drift of users who have been satisfied by receiving items of that group in the top-*10* recommendations.

We did not investigate whether any new paper that proposed a mitigation procedure against consumer fairness appeared. Instead, we checked whether the source code of the eight papers addressing a top-*n* recommendation task was still not available. Without considering the additional requirements of Boratto et al. (2022), we consider papers as reproducible if we found evidence of source code publicly available. From our investigation, for the 75% (6/8) of the papers, a source code repository was linked in the paper or the code was found on the Web. Our analysis showed that the *reproducibility* level was higher than the 50% reported by Boratto et al. (2022), but it remarks in any case the need to sharing the source code. In particular, no link to a public repository was reported in Tsintzou et al. (2019) and Wu et al. (2021b), and no trace was found querying on the web through a search engine. Indeed, the source code used by the authors of these two publications is not available anymore or it has not been made public yet.

4.6. Scalability

The increasing number of users and items in online platforms that leverage recommender systems is demanding a high amount of computational resources in terms of memory and execution time. Requirements pertaining to these resources should not be ignored, especially when a mitigation procedure against consumer unfairness should be run within or on top of the original recommendation model. As shown by our perception of the state of the art in Section 2, existing mitigation procedures were originally evaluated on barely small data sets, questioning the extent to which such procedures could be applied in real-world online platforms that involve thousands or even millions of users, items, interactions, and so on. We therefore define the scalability property as follows:

Property 6. *Scalability indicates the ability of a mitigation procedure to scale well when the number of interactions, users, items, and sensitive attributes, and other relevant features increases consistently.*

Table 7 shows the estimated requirements of time and memory for the recommendation models treated with their respective mitigation procedures. SLIM-U was the model with the highest time requirement for ML1M, even though the amount of iterations was overall the lowest among the considered approaches. This high time requirement was caused by the in-processing mitigation procedure itself, since the original recommendation model could complete the training in less than a third of that time. In LFM1K, the model was faster probably due to the lower number of users compared to ML1M (although the number of items in LFM1K is higher than ML1M). The highest memory demand among mitigation procedures was required by `Li et al. (2021a)`, due to the mathematical optimization solver, which stored a variable for any interaction in the predictions file[7]. `Ekstrand et al.`'s mitigation was the most

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

Table 7
[Scalability] Number of iterations (*Iterations*), execution time (*Time*) and computational resources (*Memory*) required by the considered mitigations for reducing unfairness on the recommendation models considered in the original papers.

Model	Iterations	ML1M		LFM1K	
		Time	Memory	Time	Memory
SLIM-U Burke et al. (2018)	10	~10h	~10 GB	~2m	~2 GB
LBM Frisch et al. (2021)	300	~2h	~500 MB	~3h	~400 MB
PMF Li et al. (2021a)	100	~5-10m	~20 GB	~5-10m	~20 GB
BiasedMF Li et al. (2021a)	100	~5-10m	~20 GB	~5-10m	~20 GB
NCF Li et al. (2021a)	100	~5-10m	~20 GB	~5-10m	~20 GB
STAMP Li et al. (2021a)	100	~5-10m	~20 GB	~5-10m	~20 GB
FunkSVD Ekstrand et al. (2018)	150	0.01s	-	0.01s	-
UserKNN Ekstrand et al. (2018)	-	0.01s	-	0.01s	-
ItemKNN Ekstrand et al. (2018)	-	0.01s	-	0.01s	-

scalable according to our framework. This pre-processing approach did not need additional computational memory to operate and the group balancing was the fastest among mitigation procedures.

Scalability was evaluated also for the studies whose source code was not available in the framework of Boratto et al. (2022). The assessment was performed on the basis of a theoretical analysis from a time complexity viewpoint. In particular, we analyzed the structure of the mitigation procedures and their scalability with regard to the increment of the number of sensitive attributes, users, items, and interactions. Specifically, Li et al. (2021b); Wu et al. (2021a,b) include components that can be largely affected by the number of the sensitive attributes in their in-processing approaches, e.g., discriminators to generate bias-free embeddings and pairwise fairness losses. Tsintzou et al. (2019)'s procedure swaps items within the recommendation lists on the basis of a fairness notion, requiring knowledge of demographic groups and items categories. Nevertheless, it is scalable because it is not affected by any of the considered factors and depends on the manually specified number of swaps.

In general, applying a mitigation procedure is an additional step in a recommendation pipeline aimed to reduce unfairness according to the pursued fairness notion. Approaches that excessively affect recommendation models in terms of time and memory would not be well suited for this purpose. On data sets with a higher number of users or interactions, the Burke et al. (2018)'s in-processing mitigation would make SLIM-U much slower, while `Li et al. (2021a)`'s mitigation procedure would lead to unmanageable memory requirements. Since the primary task of a recommender system is suggesting items of potential interest to users, researchers should focus on designing mitigation procedures that account as much as possible for *scalability*.

4.7. Trade-off Management

In the fair machine learning literature, it has been often highlighted a trade-off between accuracy/utility and fairness. This phenomenon particularly characterized the recommendation domain as well, e.g.., (Boratto et al., 2021). A notable observation coming from these prior studies is that such trade-off often appeared for the best performing recommendation models, while was less evident for the ones reporting lower utility. This would be expected since less accurate models tend to provide random predictions more frequently. Inspecting the trade-off between utility and fairness in the context of a mitigation procedure for consumer unfairness would therefore be more relevant in case the original model already achieved a good performance. In the same way, while applying a mitigation procedure on a high-performing model, in an ideal case, we expect that the mitigation effectively reduces unfairness while preserving as much as possible the original model's performance. The trade-off management property is defined as follows:

Property 7. *Trade-off management indicates the ability of a mitigation procedure to preserve the performance estimate achieved by the target recommendation model originally (before the mitigation was applied).*

To inspect trade-off management, Figure 3 reports, for each sensitive attribute and target metric, the variation of the latter after applying the respective mitigation. The value in the heatmap cell represents the ratio in percentage between the metric score achieved after mitigating and the one achieved by the original model. For instance, for NDCG, this means that all positive values are gains over the original model, while for fairness metrics (DP, KS) a gain in fairness is represented by a loss in the considered metrics (the lower they are, the fairer).

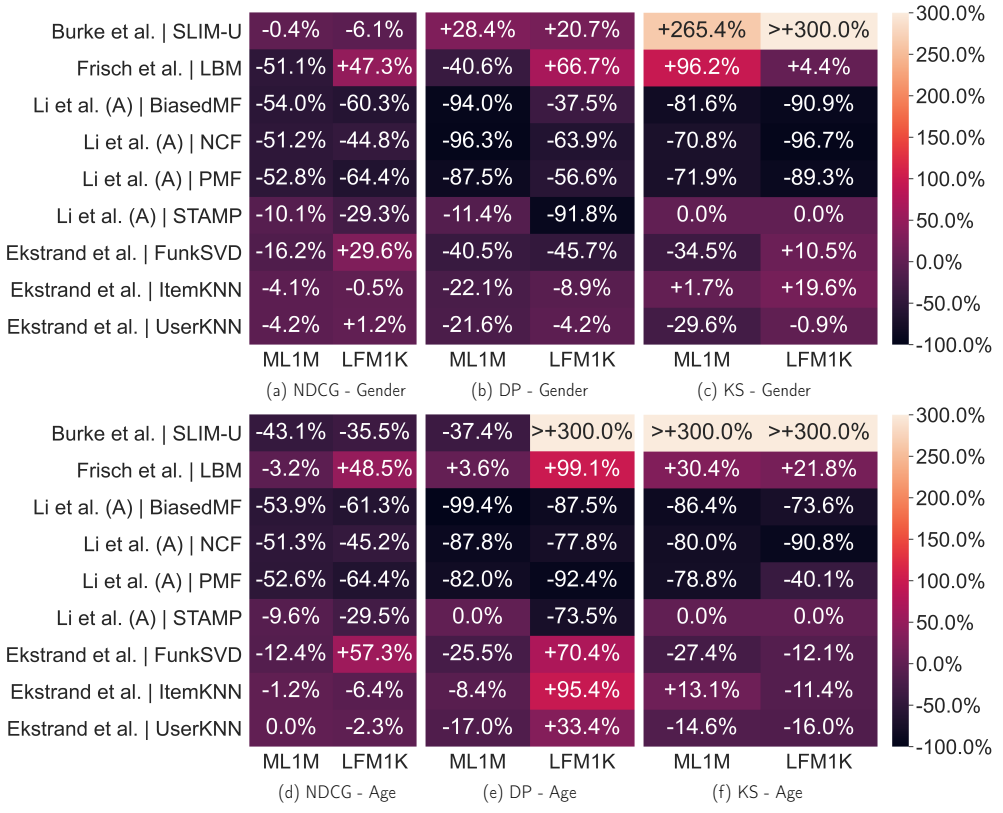


Figure 3: [Trade-off] Gain/loss in recommendation utility (*NDCG*), equity (NDCG Demographic Parity - *DP*) and independence (Kolmogorov-Smirnov - *KS*) resulted from applying the mitigations for reducing unfairness on the recommendation models used in the original papers, considering gender (a, b, c) and age (d, e, f) groups. Concerning NDCG (DP and KS), positive (negative) percentages indicate a gain in recommendation utility (fairness) after mitigating.

For gender groups (top row of the figure grid), it can be observed that `Li et al. (2021a)`'s mitigation was the best at reducing unfairness in terms of both DP and KS. However, the NDCG loss was very high, with a negative peak of -64.4% for PMF on LFM1K. A mitigation procedure that better met the *trade-off* management property was the one proposed by `Ekstrand et al. (2018)`. Indeed, on ML1M, both DP and KS were improved for most of the models, while the NDCG loss was negligible. On LFM1K, only UserKNN was positively affected by the mitigation, while a higher KS was reported by the other two models under `Ekstrand et al. (2018)`'s mitigation.

Concerning age groups, no mitigation procedure was able to perform well on both data sets in terms of trade-off. `Li et al. (2021a)`'s mitigation reported a very high fairness estimate, with a reduction of -99.4% of DP for BiasedMF, but again the NDCG decrease was substantially high. `Ekstrand et al. (2018)` reported the best *trade-off* on ML1M. Their mitigation procedure led to gains on both DP and KS (except ItemKNN for KS) and a minimal loss in NDCG, e.g., UserKNN. On the other hand, only the mitigation method proposed by `Li et al. (2021a)` was successful in reducing DP for LFM1K, whereas the best result reported by the other mitigation procedures was an increase of 33.4% on UserKNN. Conversely, in terms of KS, we observed gains in fairness by most of the models and mitigation procedures. Overall, `Ekstrand et al. (2018)`'s method was the one that reported the best *trade-off* across all the data sets and sensitive attributes. It accomplished the goal of reducing unfairness, while minimally affecting the recommendation utility. Researchers are encouraged to consider this property, so as to develop mitigation procedures that could be applied in practice in an effective way.

4.8. Transferability

In recent years, a large number of studies addressed consumer unfairness in recommender systems. However, these studies often examined the impact of the respective mitigation procedure on a restricted set of recommendation models or, in extreme cases, only on a single one. It therefore remains unclear the extent to which a mitigation procedure,

Table 8
[Transferability - Gender Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Orig*) and after (*Mit*) Ekstrand et al. (2018)'s mitigation for gender groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.079	^0.022	*0.021	^0.032	^0.032	**0.320**	**0.356**	*-0.060	*-0.082	^0.006	^0.006
Frisch et al.	LBM	0.044	0.020	**^0.006**	**-0.001**	^0.013	^0.012	0.144	0.034	*-0.035	**-0.008**	^0.120	^0.061
Li et al. (A)	BiasedMF	0.112	0.101	^0.017	^0.027	^0.035	^0.034	0.287	0.261	^-0.095	-0.068	^0.012	^0.015
	NCF	0.117	**0.113**	^0.016	^0.018	^0.022	^0.017	0.250	0.185	*-0.073	*-0.059	^0.033	^0.020
	PMF	**0.119**	0.101	*0.013	^0.027	^0.023	^0.030	0.200	0.194	*-0.062	-0.049	^0.010	^0.009
	STAMP	0.022	0.027	*0.003	**0.001**	**^0.006**	**^0.006**	0.160	0.150	**-0.021**	-0.027	**^0.001**	**^0.002**

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

Table 9
[Transferability - Age Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Orig*) and after (*Mit*) Ekstrand et al. (2018)'s mitigation for age groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.081	^0.022	^0.017	^0.009	^0.013	**0.320**	**0.345**	-0.026	-0.084	^0.017	^0.013
Frisch et al.	LBM	0.044	0.026	^0.005	0.007	^0.021	^0.013	0.144	0.016	**-0.011**	**-0.007**	^0.125	^0.066
Li et al. (A)	BiasedMF	0.112	0.107	^0.015	^0.039	^0.042	^0.032	0.287	0.271	*-0.079	-0.064	^0.019	^0.023
	NCF	0.117	**0.117**	^0.018	*0.018	^0.029	^0.019	0.250	0.186	*-0.067	-0.030	^0.055	^0.018
	PMF	**0.119**	0.107	^0.020	^0.039	^0.027	^0.034	0.200	0.175	*-0.046	-0.026	**^0.008**	^0.011
	STAMP	0.022	0.026	**0.000**	**0.003**	**^0.006**	**^0.007**	0.160	0.149	-0.031	*-0.051	**-0.008**	**^0.006**

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

originally evaluated and proven to be effective on a range of recommendation models, transfers well and it is equally effective (and so reduces unfairness) on another recommendation model. From a conceptual perspective, *transferability* is related to *consistency* (being effective across data sets and users' groups) and *applicability* (being applicable to a recommendation model). The main difference is that, with *transferability*, we aim to highlight the fact that, even if a mitigation procedure has been proven to be effective on a model, there are no guarantees that it will be equally effective on other models it could be applied to. Being *transferable* implies, for a mitigation procedure, that the consistency is preserved regardless of the model. The transferability property is defined as follows:

Property 8. *Transferability indicates the ability of a mitigation procedure to be effective (and not only applicable) on a wide range of recommendations models, even those it was not originally designed for or tested on.*

To inspect transferability, we applied the mitigation procedures proposed by `Ekstrand et al. (2018)` and `Li et al. (2021a)` on the recommendation models used by the other papers. We selected those two mitigation procedures, since they had the highest applicability among the considered ones. Tables 8-9 report the recommendation utility and fairness estimated for the recommendation models treated with the `Ekstrand et al. (2018)`'s procedure for gender and age groups, respectively. The column `Paper` identifies only the paper that used the respective model (under column `Model`) in their study, not the mitigation applied.

In ML1M, the performance measured for the considered mitigation procedure highly varied, decreasing DP for some models, e.g., SLIM-U, and increasing it for others, e.g., BiasedMF and PMF, for both age and gender groups. A similar behavior was reported for KS, but not always for the same models. The results were better for LFM1K, but the performance still varied, reducing DP in a substantial way on SLIM-U and STAMP for both gender and age groups. KS was improved for most of the models for both sensitive attributes, except when the mitigation strategy was applied

Table 10
[Transferability - Gender Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Orig*) and after (*Mit*) Li et al. (2021a)'s mitigation for gender groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit	Orig	Mit
Burke et al.	SLIM-U	0.084	0.046	^0.022	^0.008	^0.032	^0.007	0.320	0.172	*-0.060	^-0.033	**^0.006**	**^0.001**
Ekstrand et al.	FunkSVD	0.018	0.016	**^0.004**	**^0.003**	^0.027	^0.007	0.010	0.010	**-0.006**	***-0.005**	^0.107	**^0.001**
	ItemKNN	**0.140**	**0.079**	^0.038	^0.008	^0.030	^0.007	0.287	0.016	^-0.127	-0.006	^0.019	**^0.001**
	UserKNN	0.137	0.043	^0.031	^-0.007	^0.074	^0.007	**0.406**	**0.189**	^-0.110	*-0.036	^0.067	**^0.001**
Frisch et al.	LBM	0.044	0.037	^0.006	^0.005	**^0.013**	**^0.006**	0.144	0.118	*-0.035	*-0.032	^0.120	**^0.001**

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

Table 11
[Transferability - Age Groups] Recommendation utility (*NDCG*, the higher it is, the more useful the recommendations), equity (NDCG Demographic Parity - *DP*, the closer to zero it is, the fairer the model) and independence (Kolmogorov-Smirnov - *KS*, the closer to zero it is, the fairer the model) assessment of the recommendation models used in the other reproduced papers, before (*Orig*) and after (*Mit*) Li et al. (2021a)'s mitigation for age groups.

		ML1M						LFM1K					
		NDCG ↑		DP $\downarrow_0$		KS ↓		NDCG ↑		DP $\downarrow_0$		KS ↓	
Paper	Model	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit	Base	Mit
Burke et al.	SLIM-U	0.084	0.046	^0.022	^0.006	**^0.009**	**^0.006**	0.320	0.175	-0.026	0.024	**^0.017**	^0.006
Ekstrand et al.	FunkSVD	0.018	0.016	^0.008	^0.007	^0.029	**^0.006**	0.010	0.010	**0.002**	**0.002**	^0.054	^0.006
	ItemKNN	**0.140**	**0.079**	^0.027	**^0.004**	^0.029	**^0.006**	0.287	0.018	0.010	*0.012	^0.133	^0.006
	UserKNN	0.137	0.043	^0.028	^-0.007	^0.060	**^0.006**	**0.406**	**0.190**	-0.023	0.032	^0.036	^0.006
Frisch et al.	LBM	0.044	0.037	**^0.005**	^0.006	^0.021	**^0.006**	0.144	0.120	-0.011	**-0.002**	^0.125	**^0.005**

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

on BiasedMF, which reported a higher KS for both gender and age groups. In summary, Ekstrand et al. (2018) does not hold a good transferability level.

For gender and age respectively, Tables 10-11 report the results obtained after applying `Li et al. (2021a)`'s mitigation to the models included in the other papers. In ML1M, DP was reduced in all the settings, except for LBM on age groups. In LFM1K, the mitigation procedure improved DP by a considerable amount for gender groups, e.g., ItemKNN, whereas DP for age groups increased for two models out of five. KS cannot be analyzed for `Li et al. (2021a)` as pointed out in Footnote 4. Due to the poor mitigation performance on age groups in LFM1K, transferability is not satisfied by Li et al. (2021a). We encourage readers to consider *transferability* in their future works, enforcing the solidity of the logic of the proposed technique, as well as giving a clear demonstration on how their mitigation procedure can be applied to other models.

4.9. Discussion

In this section, we connect the findings coming from the individual experiments and present their implications for future research practices in the field. As a summary, Table 12 provides a categorization of the considered mitigation procedures with respect to the proposed properties. For each property and mitigation procedure, we assigned one of the two following labels: `Higher` when the corresponding work was better than the others on average for the selected property, `Lower` otherwise. A blank entry was left for studies that could not be evaluated in terms of the corresponding property. For papers whose source code was not available, the corresponding mitigation procedure was only analyzed in terms of *applicability* and *scalability*. The first two rows refer to the works that reported the highest number of above-average properties. These two studies performed better on most of the properties, e.g., on *trade-off management* and *data robustness*. `Burke et al.`'s mitigation procedure did not perform well on several properties, according to our taxonomy and framework. Our findings in this work are expected to represent a guideline for researchers working on mitigating consumer unfairness. However, some properties could not always be evaluated. For instance, evaluating

Table 12
Summary categorization of the considered mitigation procedures with respect to the proposed properties.

Paper	Applicability	Coherence	Consistency	Data Robustness	Reproducibility	Scalability	Trade-off	Transferability
Ekstrand et al. (2018)	Higher	Higher	Higher	Higher	Higher	Higher	Higher	Lower
Li et al. (2021a)	Higher	Lower	Higher	Higher	Higher	Lower	Higher	Higher
Frisch et al. (2021)	Lower	Higher	Higher	Lower	Higher	Lower	Lower	-
Burke et al. (2018)	Lower	Lower	Lower	Lower	Higher	Lower	Lower	-
Tsintzou et al. (2019)	Higher	-	-	-	Lower	Higher	-	-
Li et al. (2021b)	Higher	-	-	-	Higher	Lower	-	-
Wu et al. (2021a)	Higher	-	-	-	Higher	Lower	-	-
Wu et al. (2021b)	Higher	-	-	-	Lower	Lower	-	-

transferability would require extensive changes of the original recommender system for some mitigation procedures. Similarly, *coherence* could be affected by the performed data split or by the considered fairness metric. Furthermore, *data robustness* would be better suitable for mitigation procedures that consider causal relations with unfairness and use sensitive attributes only for assessment. Nevertheless, future works should acknowledge these circumstances.

5. Conclusions and Future Work

In this paper, we proposed to build a common framework for assessing mitigation procedures against consumer unfairness in recommender systems. First, we collected relevant papers addressing consumer unfairness mitigation and categorized them according to the definition, mitigation, and assessment strategies for consumer (un)fairness they proposed. Second, we reproduced the mitigation procedure introduced in each paper and defined a unified experimental protocol for their evaluation, including eight technical properties (applicability, coherence, consistency, data robustness, reproducibility, scalability, trade-off management, transferability) a mitigation procedure should meet. Third, we evaluated the reproduced mitigation procedures on the basis of the defined evaluation properties on two public data sets, summarizing the main trends and open issues in the field. In what follows, we present some limitations of our study and key future research directions.

Our study assessed consumer unfairness mitigation procedures in two real-world data sets (ML1M and LFM1K), which contain both users' ratings and sensitive attributes. Despite coming from different domains, having a different degree of sparsity, and varying in the number of users and items, the number and the size of the considered data sets did not allow us to provide evidently generalizable conclusions for the considered properties. Nevertheless, collecting and finding large data sets including also consumer demographic information is still an open issue for the research community. For instance, several data sets including consumer demographic attributes, such as Sushi (Kamishima, 2003), Insurance[8], RentTheRunway (Misra, Wan and McAuley, 2018), and Steam (Pathak, Gupta and McAuley, 2017) are too small even in comparison to the data sets in our study. On the other hand, larger interaction data sets like MovieLens25M (Harper and Konstan, 2016) do not report any demographic attributes, while those like BookCrossing (Ziegler, McNee, Konstan and Lausen, 2005), COCO (Gómez, Zhang, Boratto, Salamó and Marras, 2021; Gómez, Zhang, Boratto, Salamó and Ramos, 2022; Marras, Boratto, Ramos and Fenu, 2021), and LastFM-2b (Melchiorre, Rekabsaz, Parada-Cabaleiro, Brandl, Lesota and Schedl, 2021) provide disjoint sets of consumer demographic attributes (often also not publicly available), thus not enabling comparison of the same set of attributes across data sets. These latter data sets are also far larger than the others, making results less comparable across data sets. Moreover, they need novel methodological solutions for addressing the efficiency issues emerged in our study for the mitigation procedures, usually tested on small data sets in the original papers, going beyond the scope of our reproducibility study. Future work will extend our analyses on more (larger) data sets and devise more efficient mitigation procedures for them.

Connected with the above limitation, having only two data sets may not support us in concluding about the actual consistency of the mitigation procedure depending on the underlying data set. While having many different random samples from the two considered data sets could help to strengthen the generalizability of our work, it also brings several limitations. First, the two considered data sets are small and creating even smaller random samples would influence the training and mitigation outcomes more, especially for those mitigation methods that are somehow influenced by the size of the underlying data set. Second, whenever present (e.g., for ML1M), the training-test split was performed

[8] https://www.kaggle.com/mrmorj/insurance-recommendation

on a temporal basis, by sorting user's items and then taking the least recent for the training and the more recent ones for testing. A random sampling would artificially change the observed interaction behavior of users across time and therefore biasedly influence the outcome of our study with respect to what the real world provided us with. Third, our experimental protocol includes a very large number of experiments and hyper-parameters tuning (with a separate validation set for this purpose), which already allows us to partially mitigate the abovementioned issue and collect evaluation metric scores across various subsets of the data sets. Nevertheless, further analyses on more data sets or under time-aware specific sampling strategies will be conducted to make our assessment stronger.

Another main limitation of our current study calls for the assessment of the impact of consumer unfairness mitigation on other relevant properties pertaining to the provided recommendations. So far, we relied on a range of properties strictly related to the mitigation itself (e.g., its applicability and consistency) which reveal the extent to which the considered mitigation procedure can be successfully applied to a certain applicative scenario for reducing consumer unfairness (and collaterally preserving recommendation utility). However, countering such unfairness might also impact on several properties, including but not limited to novelty, diversity, catalog coverage, and serendipity. Further work is needed to account for mitigation procedures from this viewpoint, e.g., the effect of consumer unfairness on providers' fairness (Rahmani et al., 2022).

Our study in this paper focused on a top-n task for recommender systems. We therefore decided to monitor consumer unfairness according to measures pertaining to the ranked list of items (e.g., NDCG demographic parity). Although many commercial applications of these systems usually involve a top-n recommendation, there might be other recommendation tasks (e.g., rating prediction) suffering from consumer unfairness. Monitoring unfairness from a pure rating prediction perspective while having the above focus, for instance by measuring the difference in the averaged error made for two groups in predicting their ratings, might not have provided a real estimate of how this unfairness would impact users. For instance, the error disparity between groups might be due to predicted ratings on items not really relevant for the users, which would be ranked at the very bottom of the item list. Nevertheless, our framework, including formulation, design, and evaluation decisions, could be also used in other recommendation tasks in the future.

With our work, we provided a range of properties to be considered while evaluating mitigation procedures against consumer unfairness. Reporting evidence and meeting the proposed properties would better support scientists' in selecting the most suitable mitigation procedure. Although being operationalized and experimented in a recommendation scenario, which is the specificity of our study, our properties were consciously general enough for being taken into consideration for mitigation procedures targeting any other machine-learning model. Future work will be also devoted to contextualizing our findings on the considered properties in the recommendation domain with those in other machine-learning domains (e.g., speech processing and natural language processing). Indeed, these different machine-learning models are often used in the same application, such as by voice assistants to offer personalized suggestions to users. Therefore, our set of properties would enable a comparable analysis on the current state of mitigation procedures (and therefore on the extent to which unfairness is present and can be reduced the most / the least) through the pipeline steps adopted by real-world human-machine personalized interaction systems that include recommendation inside.

Inspecting mitigation procedures from a wide range of different perspectives in recommendation, as aimed by our study, is receiving more and more attention by the community. For instance, a recent study by Rahmani et al. (2022) conducted several experiments on Li et al. (2021a)'s mitigation procedure (included in our study as well), considering seven data sets (two of which with demographic attributes, ML1M and LFM), groups based on popularity item consumption and activity level, six underlying recommendation models, and the impact on provider unfairness and popularity bias. Given our focus on consumer unfairness, we limited the groups (and data sets) to those based on (providing) protected attributes regulated by the law[9]. We covered a wider range of mitigation procedures. According to our framework, concerning consistency, we similarly found that the effectiveness of the mitigation depends on the domain, with a significantly higher improvement in consumer fairness on the movie domain than in music in our case. In terms of trade-off management, differently from Rahmani et al. (2022), we often observed a decrease in recommendation utility. We conjecture that this is due to the different consumer grouping assumption, which was an influential factor in their work. Finally, we both observed that the gain in consumer fairness significantly varies across models, with a higher transferability to memory-based models than deep learning models.

[9]Please see Art. 21 of the EU Charter of Fundamental Rights, Art. 14 of European Convention on Human Rights, and Art. 18-25 of the Treaty on the Functioning of the European Union.

Our work allows to have a better understanding of the aspects that could increase the mitigation effectiveness and what can be done to avoid the phenomena outlined by our experiments. Future work will consider novel mitigation procedures able to satisfy all the properties introduced in our paper.

Acknowledgements. Giacomo Medda gratefully acknowledges the University of Cagliari for the financial support of his PhD scholarship.

Declaration of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contribution.

Ludovico Boratto: Conceptualization, Investigation, Supervision, Project administration, Validation, Writing: original draft, Writing: review & editing, Methodology, Data curation, Formal Analysis.

Gianni Fenu: Funding acquisition, Supervision, Project administration, Validation, Resources, Formal Analysis, Writing: review & editing.

Mirko Marras: Conceptualization, Investigation, Supervision, Project administration, Validation, Writing: original draft, Writing: review & editing, Methodology, Data curation, Formal Analysis.

Giacomo Medda: Conceptualization, Investigation, Software, Visualization, Writing: original draft, Writing: review & editing, Methodology, Data curation.

References

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., Pizzato, L.A., 2020. Multistakeholder recommendation: Survey and research directions. User Model. User Adapt. Interact. 30, 127–158. URL: https://doi.org/10.1007/s11257-019-09256-1, doi:10.1007/s11257-019-09256-1.

Armentano, M.G., Monteserin, A., Berdun, F., Bongiorno, E., Coussirat, L.M., 2018. User recommendation in low degree networks with a learning-based approach, in: Mexican International Conference on Artificial Intelligence, Springer. pp. 286–298.

Bellogín, A., Castells, P., Cantador, I., 2011. Precision-oriented evaluation of recommender systems: an algorithmic comparison, in: Mobasher, B., Burke, R.D., Jannach, D., Adomavicius, G. (Eds.), Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011, ACM. pp. 333–336. URL: https://doi.org/10.1145/2043932.2043996, doi:10.1145/2043932.2043996.

Boratto, L., Fenu, G., Marras, M., 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. User Model. User Adapt. Interact. 31, 421–455. URL: https://doi.org/10.1007/s11257-021-09294-8, doi:10.1007/s11257-021-09294-8.

Boratto, L., Fenu, G., Marras, M., Medda, G., 2022. Consumer fairness in recommender systems: Contextualizing definitions and mitigations, in: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V. (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham. pp. 552–566.

Burke, R., Sonboli, N., Ordonez-Gauger, A., 2018. Balanced neighborhoods for multi-sided fairness in recommendation, in: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, PMLR. pp. 202–214. URL: http://proceedings.mlr.press/v81/burke18a.html.

Celma, Ò., 2010. Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer. doi:10.1007/978-3-642-13287-2.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X., 2020. Bias and debias in recommender system: A survey and future directions. CoRR abs/2010.03240. URL: https://arxiv.org/abs/2010.03240, arXiv:2010.03240.

Dinnissen, K., Bauer, C., . Fairness in music recommender systems: A stakeholder-centered mini review. Frontiers in Big Data , 63.

Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D., Pera, M.S., 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness, in: Conference on Fairness, Accountability and Transparency, FAT 2018, PMLR. pp. 172–186. URL: http://proceedings.mlr.press/v81/ekstrand18b.html.

Ferrante, M., Ferro, N., Fuhr, N., 2021. Towards meaningful statements in IR evaluation: Mapping evaluation measures to interval scales. IEEE Access 9, 136182–136216. URL: https://doi.org/10.1109/ACCESS.2021.3116857, doi:10.1109/ACCESS.2021.3116857.

Frisch, G., Léger, J., Grandvalet, Y., 2021. Stereotype-aware collaborative filtering, in: Ganzha, M., Maciaszek, L.A., Paprzycki, M., Slezak, D. (Eds.), Proceedings of the 16th Conference on Computer Science and Intelligence Systems, Online, September 2-5, 2021, pp. 69–79. URL: https://doi.org/10.15439/2021F117, doi:10.15439/2021F117.

Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Marras, M., 2021. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems, in: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM. pp. 1808–1812. URL: https://doi.org/10.1145/3404835.3463235, doi:10.1145/3404835.3463235.

Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Ramos, G., 2022. Enabling cross-continent provider fairness in educational recommender systems. Future Gener. Comput. Syst. 127, 435–447. URL: https://doi.org/10.1016/j.future.2021.08.025, doi:10.1016/j.future.2021.08.025.

Harper, F.M., Konstan, J.A., 2016. The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst. 5, 19:1–19:19. URL: https://doi.org/10.1145/2827872, doi:10.1145/2827872.

Kamishima, T., 2003. Nantonac collaborative filtering: recommendation based on order responses, in: Getoor, L., Senator, T.E., Domingos, P.M., Faloutsos, C. (Eds.), Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003, ACM. pp. 583–588. URL: https://doi.org/10.1145/956750.956823, doi:10.1145/956750.956823.

Kamishima, T., Akaho, S., Asoh, H., Sakuma, J., 2018. Recommendation independence, in: Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA, PMLR. pp. 187–201. URL: http://proceedings.mlr.press/v81/kamishima18a.html.

Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., Schedl, M., 2021. Analyzing item popularity bias of music recommender systems: Are different genders equally affected?, in: Fifteenth ACM Conference on Recommender Systems, pp. 601–606.

Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y., 2021a. User-oriented fairness in recommendation, in: WWW '21: The Web Conference 2021, ACM / IW3C2. pp. 624–632. URL: https://doi.org/10.1145/3442381.3449866, doi:10.1145/3442381.3449866.

Li, Y., Chen, H., Xu, S., Ge, Y., Zhang, Y., 2021b. Towards personalized fairness based on causal notion, in: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM. pp. 1054–1063. URL: https://doi.org/10.1145/3404835.3462966, doi:10.1145/3404835.3462966.

Marras, M., Boratto, L., Ramos, G., Fenu, G., 2021. Equality of learning opportunity via individual fairness in personalized recommendations. International Journal of Artificial Intelligence in Education , 1–49URL: https://doi.org/10.1007/s40593-021-00271-1, doi:10.1007/s40593-021-00271-1.

Mauro, N., Ardissono, L., Cocomazzi, S., Cena, F., 2022. Using consumer feedback from location-based services in poi recommender systems for people with autism. Expert Systems with Applications 199, 116972.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54, 115:1–115:35. URL: https://doi.org/10.1145/3457607, doi:10.1145/3457607.

Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M., 2021. Investigating gender fairness of recommendation algorithms in the music domain. Inf. Process. Manag. 58, 102666. URL: https://doi.org/10.1016/j.ipm.2021.102666, doi:10.1016/j.ipm.2021.102666.

Misra, R., Wan, M., McAuley, J.J., 2018. Decomposing fit semantics for product size recommendation in metric spaces, in: Pera, S., Ekstrand, M.D., Amatriain, X., O'Donovan, J. (Eds.), Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018, ACM. pp. 422–426. URL: https://doi.org/10.1145/3240323.3240398, doi:10.1145/3240323.3240398.

Neidhardt, J., Sertkan, M., 2022. Towards an approach for analyzing dynamic aspects of bias and beyond-accuracy measures, in: International Workshop on Algorithmic Bias in Search and Recommendation, Springer. pp. 35–42.

Pathak, A., Gupta, K., McAuley, J.J., 2017. Generating and personalizing bundle recommendations on *Steam*, in: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (Eds.), Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM. pp. 1073–1076. URL: https://doi.org/10.1145/3077136.3080724, doi:10.1145/3077136.3080724.

Rahmani, H.A., Naghiaei, M., Dehghan, M., Aliannejadi, M., 2022. Experiments on generalizability of user-oriented fairness in recommender systems, in: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM. pp. 2755–2764. URL: https://doi.org/10.1145/3477495.3531718, doi:10.1145/3477495.3531718.

Singh, A., Joachims, T., 2018. Fairness of exposure in rankings, in: Guo, Y., Farooq, F. (Eds.), Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018, ACM. pp. 2219–2228. URL: https://doi.org/10.1145/3219819.3220088, doi:10.1145/3219819.3220088.

Tsintzou, V., Pitoura, E., Tsaparas, P., 2019. Bias disparity in recommendation systems, in: Burke, R., Abdollahpouri, H., Malthouse, E.C., Thai, K.P., Zhang, Y. (Eds.), Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019, CEUR-WS.org. URL: http://ceur-ws.org/Vol-2440/short4.pdf.

Wu, C., Wu, F., Qi, T., Huang, Y., Xie, X., 2019. Neural gender prediction from news browsing data, in: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (Eds.), Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings, Springer. pp. 664–676. URL: https://doi.org/10.1007/978-3-030-32381-3_53, doi:10.1007/978-3-030-32381-3_53.

Wu, C., Wu, F., Wang, X., Huang, Y., Xie, X., 2021a. Fairness-aware news recommendation with decomposed adversarial learning, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press. pp. 4462–4469. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16573.

Wu, H., Ma, C., Mitra, B., Diaz, F., Liu, X., 2021b. Multi-fr: A multi-objective optimization method for achieving two-sided fairness in e-commerce recommendation. CoRR abs/2105.02951. URL: https://arxiv.org/abs/2105.02951, arXiv:2105.02951.

Ziegler, C., McNee, S.M., Konstan, J.A., Lausen, G., 2005. Improving recommendation lists through topic diversification, in: Ellis, A., Hagino, T. (Eds.), Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005, ACM. pp. 22–32. URL: https://doi.org/10.1145/1060745.1060754, doi:10.1145/1060745.1060754.