User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms

Ningxia Wang
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
nxwang@comp.hkbu.edu.hk

Li Chen
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

ABSTRACT

There are various biases in recommender systems. Recognizing biases, as well as unfairness caused by problematic biases, is the first step of system optimization. Related studies on algorithmic biases are mainly from the perspective of either items or users. For the latter (we call it "algorithmic user bias"), existing works have considered algorithms' accuracy performances measured by accuracy metrics like RMSE. However, algorithmic user biases in beyond-accuracy measurements have rarely been studied, even though beyond-accuracy oriented recommendation algorithms have been increasingly investigated, with the purpose of breaking through the personalization limits of traditional accuracy-oriented algorithms (such as the typical "filter bubble" phenomenon). To fill in the research gap, in this work, we employ a large-scale survey dataset collected from a commercial platform, in which more than 11,000 users' ratings on the recommendation's 5 performance objectives (i.e., relevance, diversity, novelty, unexpectedness, and serendipity) and 8 kinds of user characteristics (i.e., gender, age, big-5 personality traits, and curiosity) are available. We study user biases of four algorithms (i.e., HOT, Rel-CF, Nov-CF, and Ser-CF) in terms of those five measurements between user groups of the eight user characteristics. We further look into users' behavior patterns like the preference of using more positive ratings, in order to interpret the observed biases. Finally, based on the observed algorithmic user bias and users' behavior patterns, we analyze the possible factors leading to the biases and recognize problematic biases that may lead to unfairness.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; Personalization; • **Applied computing** → Psychology.

KEYWORDS

Recommender systems, algorithmic bias, user bias, fairness, beyond-accuracy objectives, serendipity, personality, curiosity

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8458-2/21/09…$15.00
https://doi.org/10.1145/3460231.3474244

ACM Reference Format:
Ningxia Wang and Li Chen. 2021. User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3460231.3474244

1 INTRODUCTION

In the context of the vigorous development of recommendation algorithms, more and more researchers today, when evaluating recommendation algorithms, not only consider the performance improvement, but also consider the possible bias of the algorithm and unfairness issues. Liebig's law of the minimum [46] suggests that a bucket's capacity is determined by its shortest stave. For the given recommendation algorithm, its "shortest stave" lurks in all possible biases, which will limit the practicality of the recommendation algorithm. For example, it has been observed that conventional collaborative filtering approaches have the tendency of recommending popular items, which may limit users to explore novel items and/or cause slow sales of long-tail items [33, 45].

Related studies on algorithmic biases are mainly from the perspective of either items (like the popularity bias) [1] or users [12, 36, 40]. In this work, we call the latter type of bias as "algorithmic user bias", so as to distinguish from that mainly from the perceptive of items. Specifically, identifying algorithmic user bias is to analyze one algorithm's performances among different user groups, for which there are two major factors that normally need to be considered: *performance objective* and *user characteristic* used to group users.

However, existing works on algorithmic user bias [12, 36, 40] have mainly focused on accuracy performance of algorithms, as measured by well-known accuracy metrics such as Root Mean Squared Error (RMSE) and Normalized Discounter Cumulative Gain (NDCG). The algorithmic user biases in beyond-accuracy measurements have rarely been studied, even though beyond-accuracy oriented recommendation algorithms have been increasingly investigated [15, 20], with the purpose of breaking through the personalization limits of traditional accuracy-oriented algorithms (such as the typical "filter bubble" phenomenon).. Among them, diversity, novelty, and serendipity have widely been discussed [20].

Regarding user characteristics, previous works mainly considered users' demographic information (e.g., gender and age) [12, 40], but few have taken into account their psychological characteristics (such as personality traits). In a recent work [36], it was found that the studied algorithms produce more accurate recommendations for people with low *Openness*, low *Conscientiousness*, low *Extraversion*,

high *Agreeableness*, or high *Neuroticism* (note that they refer to five personality traits according to the Big-5 Factor Model [34]).

In our work, we not only analyze algorithmic user bias in several beyond-accuracy measurements, but also consider users' personality traits and curiosity given that they may affect users' appetite for novelty and serendipity [10, 21, 48]. There are two research questions we have been engaged in answering:

RQ1 : *Do algorithms have significantly different performances among different user groups, in terms of beyond-accuracy objectives?*
RQ2 : *If so, to what extent may the biased performance lead to the unfairness to users?*

We concretely employed a large-scale survey dataset collected from a commercial platform [10], which includes more than 11,000 users' ratings on 5 performance objectives (i.e., relevance, diversity, novelty, unexpectedness, and serendipity) of a recommendation, as randomly returned by one of four algorithms respectively oriented to recommending popular (HOT), accurate (Rel-CF), novel (Nov-CF), and serendipitous (Ser-CF) recommendations; and 8 kinds of user characteristics (i.e., gender, age, big-5 personality traits, and curiosity). We grouped users by each user characteristic into two groups (e.g., *low* or *high* curiosity) and tested whether each algorithm is significantly biased as for each performance objective or not. Then, we analyzed users' behavior patterns according to users' logs and ratings, in order to further interpret the results of between-group performance comparison. Combining the results of algorithmic user bias and users' behavior patterns, we figure out several problematic biases in the studied algorithms, i.e., biases that may lead to the algorithm's unfair treatments of different users.

In short, our contributions are mainly two-fold:

(1) We observe algorithmic user bias in not only accuracy but also beyond-accuracy measurements (i.e., diversity, novelty, unexpectedness, and serendipity), in terms of 8 kinds of user characteristics (i.e., gender, age, big-5 personality traits, and curiosity).
(2) Taking into consideration both the user biases of algorithms and user behavior patterns, we find to which degree the observed user biases may lead to unfairness and what problematic biases could be prioritized by algorithm developers.

The remainder is organized as follows. We first introduce the related work on algorithmic bias in recommender systems (Section 2). Then, we introduce the employed dataset (Section 3), followed by results of measuring algorithmic user bias (Section 4) and analyzing users' behavior patterns (Section 5). Finally, we discuss the major findings (Section 6) and conclude this work (Section 7).

2 RELATED WORK

Algorithmic bias refers to one type of bias that is not present in the input data, but instead, is added primarily by the algorithm's mechanism [35]. It may affect the recommender system's performances even if there are no biases from users and/or developers nor biases from the process of collecting and processing data [5]. In this section, we review related studies on algorithmic bias and organize them according to the object (items or users) that they mainly take into consideration: *algorithmic recommendation bias* that considers the bias of algorithms from the perspective of recommended items, and *algorithmic user bias* that considers the bias of algorithms from the perspective of target users.

Algorithmic Recommendation Bias. At present, the most studied algorithmic recommendation bias is the *popularity bias*, which refers to the fact that popular items are more frequently recommended to certain groups or all the users than less popular items [1]. Correspondingly, algorithms with popularity bias will exacerbate the long tail phenomenon, which can be unfair to product suppliers as well as users who desire diverse and personalized recommendations.

Conventionally, popularity bias [49] is mainly measured by the average frequency of recommendations being accessed (e.g., rating, clicking, or purchasing) [1, 12, 23, 31]. For instance, [31] studied how algorithms intensify the existing bias in historical data by applying algorithms to generate recommendations for multiple iterations and then observing the change of average popularity of recommendation lists over time. Their results show that the use of the studied algorithms (such as MostPopular, BPR, and UserKNN) can amplify popularity bias. Lots of attempts have been done to mitigate popularity bias, e.g., by propagating long-tail recommendations [45], or considering beyond-accuracy metrics like recommendation diversity [4] and category coverage [39]. [2] proposed a method of matching the recommendation's popularity with users' taste of popular items in order to achieve a fairer treatment of users.

Another measure of popularity bias considers the average rating value of recommendations. In this sense, [18] compared the average rating values of 13 algorithms in two datasets, and found that 10 of the 13 algorithms tend to recommend high-rating (> 4.1 in 5-point scale) items to users. [8] defined other variants of popularity bias (such as item statistical parity) aiming at promoting fairness. [49] proposed popularity-opportunity bias, which is to consider the popularity bias of recommendations regarding the same user's different preferences, and investigated how to debias it. Besides item-level recommendation bias, [7] explored and identified category-level bias of algorithms by computing the recommendation's popularity via its category (such as "development" that is one course category).

Algorithmic User Bias. Different from algorithmic recommendation bias, algorithmic user bias mainly considers the biased performance (e.g., accuracy) of algorithms among different user groups. For instance, [40] studied the influence of users' gender, age, country, and preferred music genres, and found that users with different characteristics receive recommendations of different qualities even using the same recommendation method. So the authors appealed for employing the combination of multiple recommendation models rather than a single method. Similarly, [12] explored the between-group differences of algorithms' accuracy performance regarding users' demographic characteristics (i.e., gender and age). By re-sampling the training data to balance the demographic distribution of input data, they found that the previously identified significant between-group differences as for accuracy performance become non-significant, which indicates re-sampling and using more uniform input data can alleviate algorithmic user bias to some extent. Moreover, their experiments showed that there is an interaction effect between popularity bias and algorithmic demographic bias.

In addition to demographics, [36] compared algorithms' accuracy performance regarding users' big-5 personality traits, called

personality bias. Three algorithms were analyzed in this work, i.e., SLIM, EASE, and Mult-VAE, which are state-of-the-art accuracy-oriented algorithms for resolving sparse data and long-tail problems. The compared algorithms did not treat users in different ways nor utilize users' personality information, but significant differences among different personality groups in terms of accuracy measurement were observed. For example, it was shown that all the studied algorithms perform more accurately for highly neurotic or lowly open users.

Limitations. In summary, we can see that researches on algorithmic bias, especially from users' perspective, are still limited. *First*, existing works on algorithmic user bias have mainly focused on accuracy performance of algorithms but the algorithmic user biases in beyond-accuracy measurements have rarely been studied, even though beyond-accuracy oriented recommendation algorithms have been increasingly investigated in recent years [15]. *Second*, regarding user characteristics, previous works on algorithmic user bias mainly considered users' demographic information, but few have taken into account their psychological characteristics, even though some recommendation approaches have attempted to exploit users' personality traits to improve personalized recommendations [42]. In this view, in addition to big-5 personality traits, other trait such as curiosity deserves exploration too, since it has been regarded as an important factor affecting people's enjoyment when they interact with new things [21]. Therefore, in this work, we have been engaged in filling in these research gaps to contribute to the research area of fairness in recommender systems.

3 DATASET DESCRIPTION

To investigate algorithmic user bias, two kinds of information are important: One is the measurement of algorithms' performance and the other one is information about user characteristics. In this work, we used *Taobao Serendipity Dataset*[1], which was collected previously [10, 44] (from Dec. 21, 2017 to March 17, 2018) on the popular e-commerce platform *Mobile Taobao* in China. There are 11,383 users' valid responses in this dataset.

Concretely, four algorithms were employed in their survey [10], respectively called *HOT*, *Rel-CF*, *Nov-CF*, and *Ser-CF*, which aim at recommending popular, relevant, novel, and serendipitous products respectively. When participating in the survey through the Mobile Taobao's client interface, the user first received a recommendation generated by one of the algorithms (randomly allocated) and was asked to give multi-faceted evaluations on the recommended product. In this work, we mainly consider relevance (i.e, the user's perceived accuracy) and four common beyond-accuracy objectives (i.e., diversity, novelty, unexpectedness, and serendipity).

In this survey, each participant was also asked to fill out some demographic information (i.e., gender and age) and answer two psychological inventories (i.e., Ten-Item Personality Inventory (TIPI) [14] for the big-five personality traits, and Curiosity and Exploration Inventory-II (CEI-II) [21] for curiosity).

[1]https://github.com/greenblue96/Taobao-Serendipity-Dataset. Note that not all of the mentioned data are included in this publicly released dataset, because of confidentiality policy of the industry partner.

3.1 Performance Measurement

In addition to accuracy that has been extensively studied in related work about algorithmic user bias (see Related Work) [12, 36, 40], we particularly consider four beyond-accuracy objectives: diversity, novelty, unexpectedness, and serendipity in this work.

- ***Diversity***. It refers to the degree of the recommendation being different from the system's previous recommendations (note that we focus on user-centric diversity instead of system-centric diversity [20] in this study).
- ***Novelty***. It is to which degree the recommended item is new/unknown to the user [20].
- ***Unexpectedness***. It refers to how unexpected/surprising the recommended item is to the user [3, 19, 24].
- ***Serendipity***. It, by definition, refers to "[...] *make discoveries, by accidents and sagacity, of things which they were not in quest for* [...]" [6, 29, 43]. In recommender systems, it normally involves two aspects: *relevance* and *surprise* (or called unexpectedness) [22]. That is, a serendipitous recommendation should not only match to the target user's preferences, but also be surprising to the user.

Each performance objective was concretely measured by asking users to rate the corresponding statement (e.g., "*The item recommended to me is a pleasant surprise*" for serendipity; see Table 1) on a 5-point Likert scale from 1 ("strongly disagree") to 5 ("strongly agree") [10, 44]. The advantage of such measurement is to avoid introducing measurement bias from researchers [41] into the study, and the evaluation from users' perspective can also be more reliable than offline metrics [20].

Table 1 lists statistics of these measurements in the dataset. Results of Kolmogorov-Smirnov test show that they are not normally distributed ($p < 0.001$), so we chose to use non-parametric methods for subsequent analysis. Moreover, the inter-correlations between every pair of objectives are all below 0.80, indicating that there is no serious multicollinearity problem [13].

3.2 Algorithms

Four algorithms were used to generate recommendations in this dataset: *HOT, Rel-CF, Nov-CF,* and *Ser-CF*, which are respectively oriented to provide popular, relevant, novel, and serendipitous recommendations [10]. Except for the non-personalized algorithm HOT, the other three are variants of Collaborative Filtering (CF) algorithm. Although basic, they can be representative of typical CF methods for achieving beyond-accuracy objectives (especially novelty and serendipity) [11, 19, 28], and the similar recommendation mechanism (i.e., CF-based) allows us to see how different utilization of the same data may lead to different biases.

- ***HOT***. It recommends the most popular item, without considering the individual user's preferences.
- ***Rel-CF***. It is based on user-user collaborative filtering (i.e., user-based CF). As an improvement on traditional user-user similarity, it considers user similarity within the same item category as well as time information, under the assumption that the smaller time interval between two users' clicks on the same item would indicate higher similarity between them.

- ***Nov-CF***. It is based on item-item collaborative filtering (i.e., item-based CF) for enhancing the recommendation's novelty. Being different from Rel-CF, Nov-CF primarily calculates the similarity of items from different categories, so that only candidate items from categories different from those of the considered item would get high similarity score. In this way, Nov-CF recommends items that are unlikely known by the user.
- ***Ser-CF***. It is also an item-based CF algorithm, but aims at enhancing both relevance and surprise of recommendations (i.e., serendipity-oriented). The assumption behind Ser-CF is that the more similar the trajectories of two items being accessed by different users, the higher their similarity is. More concretely, on the basis of Nov-CF that calculates pairwise similarity of two items from different categories, Ser-CF further strengthens their relevance by considering the similarities of all items that are adjacent to them in the user's accessing sequence[2].

The previous study [10] shows that in general Ser-CF outperforms other algorithms in terms of relevance, novelty, and serendipity, while HOT performs the best in terms of unexpectedness possibly due to its non-personalized nature. In our work, instead of comparing these algorithms [10], we mainly focus on identifying the potential bias of each algorithm among different user groups (i.e., the algorithmic user bias). Among those totally 1,383 surveyed users, the numbers (percentages) of users who evaluated recommendations by the four algorithms are respectively 2,819 (24.8%) for *HOT*, 2,817 (24.7%) for *Rel-CF*, 2,871 (25.2%) for *Nov-CF*, and 2,876 (25.3%) for *Ser-CF*.

3.3 User Characteristics

We have analyzed 8 kinds of user characteristics: gender, age, big-5 personality traits, and curiosity. They were also acquired from users' self-reported responses.

- ***Gender and Age***. There are 3,614 (31.75%) males and 7,769 (68.25%) females. As for the age, 3,274 (28.76%) users were under 20 years old at the time of experiment, 4,701 (41.30%) were 20-30 years old, 2,433 (21.37%) were 30-40 years old, 735 (6.46%) were 40-50 years old, 166 (1.46%) were 50-60 years old, and 74 (0.65%) were over 60 years old. To facilitate analysis, we divide all users into two age groups: younger users who were less than 30 years old (70.06%) and older users who were over 30 years old (29.94%). The demographic distributions are basically in line with the statistics reported about online shopping users in China [9, 37], which suggest that little population bias [38] exists in this dataset, and thus the biases observed in subsequent analysis can represent the actual situation of users' evaluation of the algorithms.
- ***Big-5 Personality Traits***. The Big-5 Factor Model is a commonly used model for representing an individual's *personality* from five dimensions: [14]: *Openness to Experience* (shorted as *Openness*), higher score of which indicates that the person is more open to new experiences and less conventional; *Conscientiousness*, higher score of which indicates that the person is more self-disciplined and less disorganized; *Extraversion*, higher score of which indicates that the person is more enthusiastic and less reserved; *Agreeableness*, higher score of which indicates that the person is more sympathetic and less critical; and *Neuroticism*, higher score of which indicates that the person is less emotionally stable and more easily upset. The Ten-Item Personality Inventory (TIPI) [14] was adopted to acquire users' personality values during the process of data collection [10, 44], by which each trait was assessed via two items (e.g., "*open to new experiences, complex*" and "*conventional, uncreative*" for *Openness*; each on a 7-point Likert scale). The distributions of those users' personality traits are shown in Figure 1(a) - 1(e). The means (medians) of their openness, conscientiousness, extraversion, agreeableness, and neuroticism scores are respectively 4.63 (4.50), 4.56 (4.50), 4.17 (4.00), 4.97 (5.00), and 4.26 (4.00). Kolmogorov-Smirnov tests show that all of the five personality traits are not normally distributed ($p < 0.001$).
- ***Curiosity***. Curiosity, as a personal attribute, refers to the desire to know or learn something new [27]. Previous studies [10, 44] found that users with high curiosity are more concerned about the novelty of recommended items and are more satisfied with serendipitous recommendations. In this dataset [10, 44], Curiosity and Exploration Inventory-II (CEI-II) [21] was used to measure users' curiosity, concretely about their *"motivation to seek out knowledge and new experiences"* and *"willingness to embrace the novel, uncertain, and unpredictable nature of everyday life"*. The distribution w.r.t. curiosity is shown in Figure 1(f), with the mean (median) of 3.14 (3.1). It is neither normally distributed according to Kolmogorov-Smirnov test ($p < 0.001$).

4 ALGORITHMIC USER BIAS

In order to investigate whether those four recommending algorithms (see Section 3.2) perform differently between user groups, especially regarding the beyond-accuracy objectives, we first split users into two groups as for each type of user characteristic. Median split method [17] was used to group users (except gender and age). Then, we used Mann-Whitney U test[3] to see whether users' evaluations on one objective are significantly different between the two groups (e.g., *low* and *high* curiosity). Results are reported in Table 2, where the bold numbers indicate significant performance differences of the algorithm between two compared user groups in terms of the corresponding objective. We organize those results by the five performance objectives (i.e., relevance, diversity, novelty, unexpectedness, and serendipity) and describe the identified algorithmic user biases.

4.1 Relevance

All of the four studied algorithms have significantly different performances ($p < 0.05$) between the user groups w.r.t. Age, two personality traits (Conscientiousness and Neuroticism), and Curiosity. To be more specific, these algorithms all produce more relevant recommendations for older users (mean = 3.47/3.02 in older/younger user group), more conscientious users (mean = 3.30/3.04 in high/low

[2]Interested readers can refer to [10] for detailed description of each algorithm.

[3]Mann-Whitney U test is a non-parametric test of whether one of two random variables is larger than the other, and can work with unequal sample sizes [30].

Table 1: Statistics of the five objective measurements

Objective &point Likert scale)	Statement for user evaluation (responded on a 5-	Mean (Std.)	Median	K-S test
Relevance	*"The item recommended to me matches my interests."*	3.15 (1.452)	4.00	0.251***
Diversity	*"The item recommended to me is similar to the system's prior recommendations." (reversed)*	3.04 (1.320)	3.00	0.192***
Novelty	*"The item recommended to me is novel."*	2.94 (1.426)	3.00	0.229***
Unexpectedness	*"The item recommended to me is unexpected."*	3.15 (1.456)	3.00	0.208***
Serendipity	*"The item recommended to me is a pleasant surprise."*	2.65 (1.454)	2.00	0.200***

Note: $^{***}p < 0.001$ for Kolmogorov-Smirnov (K-S) test.

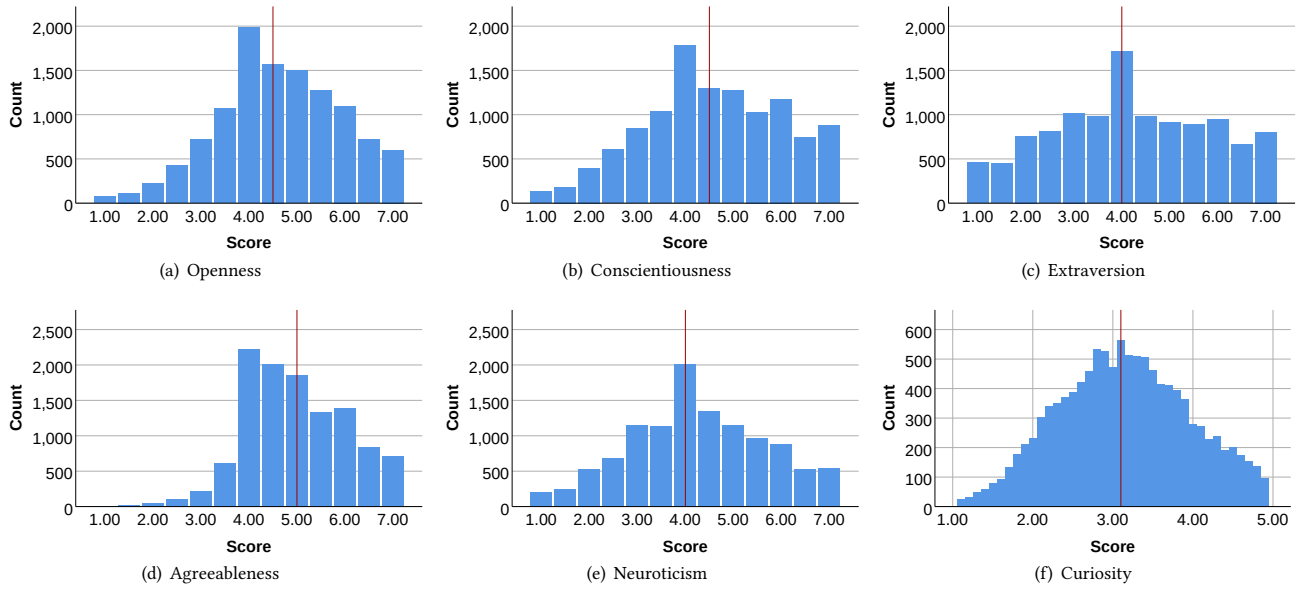

Figure 1: Distributions of the big-5 personality traits (from 1 to 5, with the granularity of 0.5) and curiosity (from 1 to 7, with the granularity of 0.1) among all users. The red line indicates the median for each trait.

conscientiousness group), more neurotic users (mean = 3.25/3.06 in high/low neuroticism group), and more curious users (mean = 3.41/2.91 in high/low curiosity group). In addition, Nov-CF and Ser-CF show relevance bias between users of high and low Openness, and HOT and Nov-CF have bias between users of high and low Agreeableness. If grouping users by Gender, only HOT is biased in terms of relevance, with means 2.84 and 2.66 respectively for male and female users; and if grouping users by Extraversion, only Rel-CF is biased in terms of relevance, with means 3.11 and 2.90 respectively for high-extraversion and low-extraversion users.

On the other hand, each algorithm has different degrees of bias for different user characteristics. Specifically, HOT biases most with respect to Age with the largest between-group difference 0.59 (i.e., $3.14 - 2.55 = 0.59$), followed by Curiosity (0.42) and Conscientiousness (0.30). Rel-CF biases most w.r.t. Curiosity (difference between high and low curiosity groups is 0.49), followed by Conscientiousness (0.42) and Age (0.40). Nov-CF biases most w.r.t. Curiosity (0.51), followed by Age (0.36) and Conscientiousness (0.25). Ser-CF biases most w.r.t. Curiosity (0.47), also followed by Age (0.37) and Conscientiousness (0.17).

Shortly, although many biases observed in terms of the recommendation's relevance, considering algorithm optimization, developers may consider debiasing algorithms' relevance between groups w.r.t. Age (for HOT) and Curiosity (for Rel-CF, Nov-CF, and Ser-CF) first.

4.2 Diversity

All of the algorithms are biased in terms of diversity between user groups regarding Age and Curiosity. Specifically, algorithmic diversity performances perceived by younger users (mean = 3.14) and lowly curious users (mean = 3.19) are higher than those perceived by older users (mean = 2.80) and highly curious users (mean = 2.89) respectively ($p < 0.001$). For personality traits, except Ser-CF, recommendations are more diverse for low Conscientiousness users than for high Conscientiousness users; while Ser-CF is biased w.r.t. Agreeableness (mean = 2.74 and 2.60 respectively for

Table 2: Results of algorithmic user bias for each objective measurement regarding different user groups by Mann-Whitney U test

Group	Algo.	Relevance			Diversity			Novelty			Unexpectedness			Serendipity		
		All	High	Low	All	High	Low	All	High	Low	All	High	Low	All	High	Low
Gender	HOT	2.71	**2.84****	2.66	3.45	3.34	**3.49****	2.70	**2.88*****	2.62	3.36	**3.49****	3.31	2.31	**2.47*****	2.24
	Rel-CF	3.00	3.03	2.98	3.16	3.03	**3.21****	2.80	**2.96*****	2.73	3.16	**3.33*****	3.09	2.50	**2.63****	2.44
	Nov-CF	3.29	3.30	3.28	2.89	2.86	2.91	3.06	**3.20*****	2.99	3.06	**3.25*****	2.96	2.80	**2.92****	2.74
	Ser-CF	3.61	3.64	3.59	2.67	2.62	**2.70***	3.19	**3.34*****	2.11	3.03	**3.19*****	2.96	2.99	**3.10****	2.94
	All	3.15	**3.22*****	3.12	3.04	2.95	**3.08*****	2.94	**3.10*****	2.86	3.15	**3.31*****	3.08	2.65	**2.79*****	2.59
Age	HOT	2.71	**3.14*****	2.55	3.45	3.11	**3.58*****	2.70	**3.15*****	2.52	3.36	**3.58*****	3.28	2.31	**2.77*****	2.13
	Rel-CF	3.00	**3.28*****	2.88	3.16	2.93	**3.25*****	2.80	**3.17*****	2.65	3.16	**3.47*****	3.03	2.50	**2.86*****	2.35
	Nov-CF	3.29	**3.54*****	3.18	2.89	2.72	**2.97*****	3.06	**3.37*****	2.92	3.06	**3.29*****	2.95	2.80	**3.13*****	2.65
	Ser-CF	3.61	**3.86*****	3.49	2.67	2.50	**2.76*****	3.19	**3.55*****	3.02	3.03	**3.41*****	2.86	2.99	**3.41*****	2.80
	All	3.15	**3.47*****	3.02	3.04	2.80	**3.14*****	2.94	**3.32*****	2.78	3.15	**3.43*****	3.03	2.65	**3.06*****	2.48
Ope.	HOT	2.71	2.77	2.67	3.45	3.48	3.42	2.70	2.70	2.70	3.36	3.38	3.34	2.31	**2.39***	2.24
	Rel-CF	3.00	3.00	3.00	3.16	3.17	3.15	2.80	2.75	2.84	3.16	3.16	3.16	2.50	2.53	2.47
	Nov-CF	3.29	**3.35****	3.23	2.89	2.90	2.88	3.06	3.10	3.03	3.06	**3.12***	3.01	2.80	**2.88****	2.73
	Ser-CF	3.61	**3.68*****	3.55	2.67	2.69	2.68	3.19	3.19	3.18	3.03	3.06	3.02	2.99	**3.05***	2.95
	All	3.15	**3.20*****	3.11	3.04	3.05	3.03	2.94	2.94	2.94	3.15	3.18	3.13	2.65	**2.72*****	2.60
Con.	HOT	2.71	**2.89*****	2.59	3.45	3.39	**3.49***	2.70	**2.82*****	2.61	3.36	**3.45****	3.29	2.31	**2.53*****	2.15
	Rel-CF	3.00	**3.31*****	2.89	3.16	3.08	**3.22****	2.80	**2.95*****	2.68	3.16	**3.24****	3.10	2.50	**2.63*****	2.40
	Nov-CF	3.29	**3.42*****	3.17	2.89	2.85	**2.93***	3.06	**3.20*****	2.93	3.06	**3.20*****	2.93	2.80	**2.97*****	2.65
	Ser-CF	3.61	**3.70*****	3.53	2.67	2.66	2.69	3.19	**3.31*****	3.08	3.03	**3.15*****	2.93	2.99	**3.14*****	2.87
	All	3.15	**3.30*****	3.04	3.04	2.98	**3.09*****	2.94	**3.08*****	2.82	3.15	**3.26*****	3.07	2.65	**2.83*****	2.51
Ext.	HOT	2.71	2.77	2.67	3.45	3.45	3.45	2.70	2.75	2.65	3.36	3.39	3.34	2.31	**2.38***	2.25
	Rel-CF	3.00	**3.11*****	2.90	3.16	3.13	3.18	2.80	**2.89****	2.72	3.16	3.20	3.13	2.50	**2.59****	2.43
	Nov-CF	3.29	3.34	3.25	2.89	2.91	2.88	3.06	3.10	3.03	3.06	3.09	3.03	2.80	**2.86***	2.75
	Ser-CF	3.61	3.64	3.58	2.67	2.66	2.69	3.19	3.22	3.16	3.03	3.07	3.01	2.99	3.04	2.95
	All	3.15	**3.22*****	3.10	3.04	3.03	3.05	2.94	**3.00*****	2.89	3.15	**3.18***	3.13	2.65	**2.72*****	2.59
Agr.	HOT	2.71	2.65	**2.79***	3.45	3.48	3.41	2.70	2.66	2.74	3.36	3.29	**3.45***	2.31	2.23	**2.40****
	Rel-CF	3.00	2.97	3.02	3.16	3.2	3.11	2.80	2.73	**2.88***	3.16	3.16	3.17	2.50	2.47	2.54
	Nov-CF	3.29	3.21	**3.38****	2.89	2.92	2.87	3.06	3.00	**3.13*****	3.06	3.02	3.11	2.80	2.74*	**2.87***
	Ser-CF	3.61	3.57	3.65	2.67	**2.74****	2.60	3.19	3.16	**3.23***	3.03	2.98	**3.09***	2.99	2.92	**3.09****
	All	3.15	3.10	**3.21*****	3.04	**3.08*****	3.00	2.94	2.89	**3.00*****	3.15	3.11	**3.20****	2.65	2.59	**2.72*****
Neu.	HOT	2.71	**2.81****	2.63	3.45	3.42	3.47	2.70	**2.78****	2.62	3.36	3.36	3.36	2.31	**2.40****	2.22
	Rel-CF	3.00	**3.14*****	2.88	3.16	3.12	3.19	2.80	**2.91*****	2.71	3.16	**3.24****	3.09	2.50	**2.66*****	2.36
	Nov-CF	3.29	**3.34***	3.24	2.89	2.88	2.90	3.06	**3.14****	2.99	3.06	**3.15****	2.97	2.80	**2.87***	2.74
	Ser-CF	3.61	**3.69*****	3.52	2.67	2.68	2.67	3.19	**3.30*****	3.08	3.03	**3.12****	2.96	2.99	**3.10*****	2.89
	All	3.15	**3.25*****	3.06	3.04	3.02	**3.06*****	2.94	**3.04*****	2.85	3.15	3.22	3.09	2.65	**2.76*****	2.55
Cur.	HOT	2.71	**2.94*****	2.52	3.45	3.27	**3.60*****	2.70	**2.90*****	2.53	3.36	**3.48*****	3.26	2.31	**2.62*****	2.04
	Rel-CF	3.00	**3.25*****	2.76	3.16	3.00	**3.30*****	2.80	**3.02*****	2.60	3.16	**3.28*****	3.06	2.50	**2.80*****	2.22
	Nov-CF	3.29	**3.54*****	3.03	2.89	2.78	**3.00*****	3.06	**3.31*****	2.81	3.06	**3.30*****	2.81	2.80	**3.15*****	2.45
	Ser-CF	3.61	**3.83*****	3.36	2.67	2.56	**2.80*****	3.19	**3.43*****	2.93	3.03	**3.26*****	2.80	2.99	**3.34*****	2.63
	All	3.15	**3.41*****	2.91	3.04	2.89	**3.19*****	2.94	**3.18*****	2.71	3.15	**3.33*****	2.99	2.65	**2.99*****	2.33

Note: 1). The "All" row indicates the means regardless of the algorithm. The "All" column indicates the means regardless of the user group. The columns "High" and "Low" give the average evaluation ratings in the male/older/high-personality/high-curiosity user group and female/younger/low-personality/low-curiosity user group, respectively. 2). The group having significantly higher evaluation rating is marked in bold ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, by Mann-Whitney U test).

high and low agreeableness groups). As for Gender, except Nov-CF, recommendations are more diverse for females than for males.

Regarding the bias degree, all the studied algorithms have the largest bias w.r.t. Age, followed by Curiosity. For instance, the difference in terms of HOT's diversity performance between Age groups (younger and older users) is 0.47, followed by Curiosity (0.33).

Therefore, to improve the algorithm's diversity performance, researchers could consider the desired diversity by users of different ages as well as considering their curiosity levels. For instance, they may aim to increase diversity for younger and low-curiosity users, which may potentially optimize the overall diversity performance of the target algorithm.

4.3 Novelty

All of the studied algorithms produce more novel recommendations regarding five user characteristics, i.e., Gender (mean = 3.10/2.86 in the male/female group), Age (mean = 3.32/2.78 in the older/younger user group), Conscientiousness (mean = 3.08/2.82 in the high/low conscientiousness group), Neuroticism (mean = 3.04/2.85 in the high/low neuroticism group), and Curiosity (mean = 3.18/2.71 in the high/low curiosity group). Moreover, for Agreeableness groups,

except HOT, all algorithms produce more novel recommendations for low agreeableness users. For Extraversion groups, only Rel-CF is biased, which produces more novel items to high extraversion users (mean = 2.89/2.72 in the high/low extraversion group).

As for each algorithm, HOT and Rel-CF bias most w.r.t Age, Nov-CF biases most w.r.t. Curiosity, and Ser-CF biases most w.r.t. Gender. Specifically, the largest between-group difference of HOT's novelty performance is 0.63 w.r.t. Age, followed by 0.37 w.r.t. Curiosity. For Rel-CF, the difference is 0.52 w.r.t. Age, followed by 0.42 w.r.t. Curiosity. For Nov-CF, it is 0.50 w.r.t. Curiosity, followed by 0.25 w.r.t. Age. For Ser-CF, it is 1.23 w.r.t. Gender, followed by 0.53 w.r.t. Age and 0.50 w.r.t. Curiosity.

Still, similar to the above observations, biases between Age/Curiosity groups could be paid more attention to by algorithm developers. It is also worth noting that the novelty of recommendations provided by Ser-CF biases seriously between males and females, which might deserve attention too.

4.4 Unexpectedness

It is observed that all the studied algorithms produce biased recommendations in terms of unexpectedness evaluation as for four user characteristics, i.e., Gender (mean = 3.31/3.08 in the male/female group), Age (mean = 3.43/3.03 in the older/younger user group), Conscientiousness (mean = 3.26/3.07 in the high/low conscientiousness group), and Curiosity (mean = 3.33/2.99 in the high/low curiosity group). Moreover, the three CF-based algorithms (i.e., Rel-CF, Nov-CF, and Ser-CF) are all biased as for Neuroticism (higher unexpectedness for more neurotic users). As for Openness, Nov-CF produces more unexpected recommendations for high-openness users (mean = 3.12/3.01 in the high/low openness group). For Agreeableness groups, HOT and Ser-CF are biased and generate more unexpected recommendations for low-agreeableness users.

Regarding the largest between-group difference of each algorithm's unexpectedness performance, three algorithms (i.e., HOT, Rel-CF, and Ser-CF) bias most between Age groups, followed by Curiosity groups; while Nov-CF biases most w.r.t. Curiosity, then Age.

4.5 Serendipity

In terms of serendipity, there are just three *unbiased* results, i.e., Rel-CF between Openness/Agreeableness groups and Ser-CF between Extraversion groups. It shows that those studied algorithms all produce significantly more serendipitous recommendations for male users (mean = 2.79 vs. 2.59 for females), older users (mean = 3.06 vs. 2.48 for younger users), high-Openness users (mean = 2.72 vs. 2.60 for low-openness users), high-Conscientiousness (mean = 2.83 vs. 2.51 for low-conscientiousness users), high-Extraversion users (mean = 2.72 vs. 2.59 for low-extraversion users), low-Agreeableness users (mean = 2.72 vs. 2.59 for high-agreeableness users), high-Neuroticism users (mean = 2.76 vs. 2.55 for low-neuroticism users), and high-Curiosity users (mean = 2.99 vs. 2.33 for low-curiosity users).

Among those biases, HOT biases most between Age groups (difference is 0.64), followed by Curiosity (0.58); and the other three CF-based algorithms bias most w.r.t. Curiosity (differences are 0.58, 0.70, and 0.71 respectively for Rel-CF, Nov-CF, and Ser-CF).

Combining all of the above observations, it suggests similar implication regarding those five objectives. That is, from the perspective of balancing the algorithm's performance between different user groups, user characteristics would better be considered during the algorithm development.

5 USERS' BEHAVIOR PATTERNS

We further look into users' historical behavior patterns, in order to interpret the results in Table 2 from another perspective. For example, to figure out why, given a particular algorithm, younger users receive recommendations with perceived higher diversity than older users. In this way, we may better understand, to what extent, the statistical differences found in the previous analyses would be caused by users' behavioral patterns or by the algorithm's mechanism.

Concretely, we measured users' historical preference patterns by three variables (i.e., *percentage of purchases*, *category coverage*, and *interest diversity*) and their behavior patterns when participating the survey [10, 44] by the *percentage of positive ratings*.

5.1 Historical Preference Patterns

We obtained users' logs in the past 3 months before the time when they took part in the survey as described in Section 3. There are in total 7,717,420 item ids, 9,085 category ids[4], and 21,405,555 user-item clicking records (4.21% of which contain purchasing records). We then calculated:

- **Percentage of purchases**. It refers to the ratio of purchase records in a user's profile that includes all items s/he has clicked. Because users' purchase behavior is binary, we performed log transformation to correct for outliers and skewness [16].
- **Category coverage**. It is calculated by the number of distinct category ids in the user's profile. Users with higher category coverage have interacted with a wider variety of item categories.
- **Interest diversity**. It is calculated by the Shannon entropy of all categories that have been visited by the user [16]. Higher interest diversity implies that the user's preference is more diverse, since her/his visits on those categories are more evenly distributed.

The results of Spearman's correlation coefficients between user characteristics and the above three behavior variables are given in Table 3, from which we can see that: 1). Males, high-openness users, high-neuroticism users, and high-curiosity users are more likely to purchase items ($p < 0.01$). 2). The results of category coverage are basically in line with those of interest diversity in terms of Gender, Age, and Openness. Specifically, we find that females, older users, and higher-openness users have not only visited more categories but also more evenly interacted with them ($p < 0.05$). 3). High-conscientiousness users and high-curiosity users have more diverse preferences across the categories they have visited (i.e., with higher interest diversity, $p < 0.05$), while high-agreeableness users and

[4]The categories of each item can be represented as a path in the hierarchical taxonomy [44] (e.g., "Clothes" → "Women's clothing" → "Suit uniform" → "Work uniform", in which case the leaf category is "Work uniform"). In this work, we mainly counted the item's leaf category id.

Table 3: Spearman's correlation coefficients between user characteristics and behavior variables ($^{*}p < 0.05$, $^{**}p < 0.01$)

	Percentage of purchases (log)	Category coverage	Interest diversity
Gender	**0.035****	**-0.194****	**-0.111****
Age	-0.006	**0.117****	**0.128****
Openness	**0.033****	**0.055****	**0.066****
Conscientiousness	-0.015	0.011	**0.022***
Extraversion	0.012	0.008	0.015
Agreeableness	-0.001	**0.024****	0.015
Neuroticism	**0.027****	**-0.019***	-0.004
Curiosity	**0.025****	-0.007	**0.019***

low-neuroticism users have visited more categories (i.e., with larger category coverage, $p < 0.05$).

5.2 Rating Patterns in the Survey

We also analyzed users' rating patterns in the survey data. We formally counted the percentage of positive ratings (>=4 on a 5-point Likert scale; see Table 1) in different user groups. Moreover, in view of the observation that users' ratings in terms of different performance objectives can be related to the difficulty of achieving each (e.g., the recommendation's relevance can reach mean rating (median) of 3.15 (4.00), while that of serendipity is relatively low, i.e., 2.65 (2.00); see Table 1), we concretely analyzed users' rating patterns in terms of each objective.

Results are reported in Table 4. The Chi-Square test was used to see whether there is a significant association between two variables (e.g., relevance and gender). If there is significant result ($p < 0.05$), it means the percentages of positive ratings (e.g., ratings on relevance) are significantly different between the two concerned user groups (e.g., Gender groups). Results show that males, older users, and high-openness/high-conscientiousness/high-extraversion/high-neuroticism/high-curiosity users tend to provide more positive ratings in terms of relevance, novelty, unexpectedness, and serendipity; but less positive ratings in terms of diversity. In the next section, we discuss how the results could be helpful for interpreting the identified biases in Table 2.

6 DISCUSSION

6.1 RQ1: *Do algorithms have significantly different performance among different user groups, in terms of beyond-accuracy objectives?*

We first investigated algorithmic user bias in terms of various performance objectives. Based on results in Section 4, we have three major observations: 1). Algorithms are more biased in terms of ***serendipity*** compared with other objectives, as almost all of the results in terms of serendipity are significantly biased between concerned user groups, and the differences between user groups' mean ratings on serendipity are mostly larger than those for the other objectives. This might be because achieving serendipity is more challenging as it accommodates both relevance and surprise to be balanced. Previous study [12] showed that correcting for one bias may lead to another bias, so with two objectives considered together, obtaining unbiased serendipity between user groups would become a tougher task. 2). Algorithms are more subject to produce distinct biased results between user groups w.r.t. **Age and Curiosity**. 3). For a specific kind of algorithmic user bias, algorithms that exhibit significant bias are all biased in the similar manner, being consistent with the finding in [36]. For example, those algorithms, which are significantly biased between Curiosity groups in terms of serendipity (i.e., HOT, Rel-CF, Nov-CF, and Ser-CF), all provide more serendipitous recommendations for highly curious users.

We further looked into users' behavior patterns and found some interesting phenomena that may explain the above observations, especially about Age and Curiosity. For instance, older users have larger catalog coverage and higher interest diversity than younger users, which indicate that older users are more likely to visit a wider variety of categories and have more even preference distribution. This phenomenon may explain why, when evaluating recommendations by an algorithm, older users were easier to feel the recommendation relevant (since the probability of a given category being relevant to older users would be higher), but were inclined to give lower diversity ratings compared to younger users, since with more categories already being interacted with and known, older users would be less likely to perceive the recommendation different from previously visited ones.

Another interesting finding is about Curiosity. It shows that high-curiosity users have higher interest diversity according to their historical records, but there is no significant correlation with category coverage, which implies that high curiosity may not necessarily promote users to explore more categories, but would stimulate them to explore a category in depth. It may also explain why high-curiosity users were easier to give higher ratings on recommendation serendipity, which is in line with the previous findings that more curious users are more likely to perceive a novel item as serendipitous and be satisfied with a serendipitous recommendation [10].

6.2 RQ2: *To what extent may the biased performance lead to the unfairness to users?*

As discussed before, because the studied algorithms all bias with regard to Age/Curiosity groups, it implies that such bias would better be prioritized by algorithm developers.

Furthermore, combining algorithmic user bias (Table 2) with users' survey rating patterns (Table 4), we notice two potentially problematic biases that may lead to the unfairness to users: 1). **High-agreeableness users** are more likely to give positive ratings in term of ***novelty*** as shown in Table 4, but their average ratings on the novelty of recommendations by Rel-CF/Nov-CF/Ser-CF are significantly lower compared to low-agreeableness users in Table 2, which infers that the recommendations generated by these three algorithms are less novel to high-agreeableness users. 2). **Low-neuroticism users** are more likely to give positive ratings in term of ***relevance*** as shown in Table 4, but their average ratings on the relevance of recommendations by HOT/Rel-CF/Nov-CF/Ser-CF are all lower compared to high-neuroticism users in Table 2, which infers that these algorithms' produced recommendations might be less relevant to low-neuroticism users.

Table 4: The percentage (%) of positive survey ratings (>=4) in user groups

Objective	Gender		Age		Ope.		Con.		Ext.		Agr.		Neu.		Cur.	
	M	F	>30	18-30	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Relevance	**55.7**	52.0	**63.5**	48.7	**54.9**	52.6	**58.5**	48.8	**55.5**	51.2	52.0	**54.5**	46.4	**50.2**	**61.3**	45.4
Diversity	36.1	**38.7**	30.4	**41.1**	**39.3**	36.7	37.0	38.6	37.9	37.9	38.5	37.1	37.4	38.3	34.5	**41.1**
Novelty	**51.8**	43.2	**59.0**	40.0	46.2	45.2	**51.2**	41.2	**48.0**	43.8	**54.0**	47.7	**48.8**	42.8	**54.1**	37.7
Unexpectedness	**54.4**	46.2	**58.7**	44.6	49.7	48.0	**53.0**	45.4	**50.3**	47.5	47.3	**50.5**	**51.4**	46.4	**55.1**	42.8
Serendipity	**40.9**	32.8	**48.6**	29.7	**37.6**	33.5	**41.7**	30.2	**37.6**	33.5	33.7	**37.3**	**39.3**	31.8	**46.6**	24.7

Note: The Chi-Square test was run to identify whether there is significant difference in the percentage of positive ratings between two user groups (e.g., Males (M) and Females (F) regarding Gender). The number in bold is significantly larger than the counterpart ($p < 0.05$).

7 CONCLUSION

To the best of our knowledge, this is the first work that investigates recommendation algorithms' biases regarding beyond-accuracy objectives (i.e., diversity, novelty, unexpectedness, and serendipity). It is also one of few works that investigate the algorithm bias from users' psychological characteristics (such as personality traits and curiosity). Our method that takes into consideration both algorithmic user biases and users' behavior patterns might also be suggestive for identifying some problematic biases that could be prioritized by algorithm developers. In particular, we have three major observations: 1). Regarding performance objectives, all the studied algorithms bias more obviously in terms of *serendipity* relative to other objectives. 2). As for user characteristics, the degrees of the observed biases are more distinct between user groups w.r.t. Age and Curiosity. 3). For a certain kind of bias, the studied algorithms are all biased (if any) towards the same group. With increasing studies on the impact of user characteristics on improving beyond-accuracy recommendations [29, 42, 47, 48], we believe our findings could promote the development of more unbiased and fairer recommendation methods.

However, as restricted by the used dataset, we did not investigate biases of other state-of-the-art algorithms [24–26, 47, 48], and some of other behavior patterns (like the uniformity of user ratings [32]). In the future work, we will attempt to not only address these limitations, but also study how to develop more unbiased recommendation algorithms based on the observations.

ACKNOWLEDGMENTS

This work was supported by Hong Kong Research Grants Council (RGC) (project RGC/HKBU12201620). We are also thankful for Yonghua Yang, Keping Yang, and Quan Yuan who helped collect the data in the previous work. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the collaborators and sponsor.

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. arXiv:1907.13286 [cs.IR]

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-Centered Evaluation of Popularity Bias in Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 11 pages. https://doi.org/10.1145/3450613.3456821

[3] Panagiotis Adamopoulos and Alexander Tuzhilin. 2015. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Transactions on Intelligent Systems and Technology* 5, 4 (Jan. 2015), 1–32. https://doi.org/10.1145/2559952

[4] Gediminas Adomavicius and YoungOk Kwon. 2012. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (May 2012), 896–911. https://doi.org/10.1109/TKDE.2011.15

[5] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[6] Toine Bogers and Lennart Björneborn. 2013. Micro-serendipity: Meaningful Coincidences in Everyday Life Shared on Twitter. In *Proceedings of the iConference 2013.* 196–208. https://doi.org/10.9776/13175

[7] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2019. The Effect of Algorithmic Bias on Recommender Systems for Massive Open Online Courses. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer International Publishing, Cham, 457–472. https://doi.org/10.1007/978-3-030-15712-8_30

[8] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Connecting User and Item Perspectives in Popularity Debiasing for Collaborative Recommendation. *Information Processing & Management* 58, 1 (Jan. 2021), 102387. https://doi.org/10.1016/j.ipm.2020.102387

[9] China Internet Network Information Center. 2018. *The 41st Statistical Report on Internet Development in China.* http://www.cac.gov.cn/files/pdf/cnnic/CNNIC41.pdf

[10] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 240–250. https://doi.org/10.1145/3308558.3313469

[11] Anup A. Deshmukh, Pratheeksha Nair, and Shrisha Rao. 2018. A Scalable Clustering Algorithm for Serendipity in Recommender Systems. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. 1279–1288. https://doi.org/10.1109/ICDMW.2018.00182

[12] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 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*. PMLR, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[13] Donald E. Farrar and Robert R. Glauber. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics* 49, 1 (1967), 92–107.

[14] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. 2003. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality* 37, 6 (Dec. 2003), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

[15] Asela Gunawardana and Guy Shani. 2015. *Evaluating Recommendation Systems.* Springer US, Boston, MA, 265–308.

[16] Rong Hu and Pearl Pu. 2014. Exploring Personality's Effect on Users' Rating Behavior. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 2599–2604. https://doi.org/10.1145/2559206.2581317

[17] Dawn Iacobucci, Steven S. Posavac, Frank R. Kardes, Matthew J. Schneider, and Deidre L. Popovich. 2015. The Median Split: Robust, Refined, and Revived. *Journal of Consumer Psychology* 25, 4 (2015), 690–704. https://doi.org/10.1016/j.jcps.2015.06.014

[18] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What Recommenders Recommend: An Analysis of Recommendation Biases and Possible Countermeasures. 25, 5 (2015), 427–491. https://doi.org/10.1007/s11257-015-9165-3

[19] Marius Kaminskas and Derek Bridge. 2014. Measuring Surprise in Recommender Systems. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*. Silicon Valley, United States, 6.

[20] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1 (Dec. 2016), 2:1–2:42. https://doi.org/10.1145/2926720

[21] Todd B. Kashdan, Matthew W. Gallagher, Paul J. Silvia, Beate P. Winterstein, William E. Breen, Daniel Terhar, and Michael F. Steger. 2009. The Curiosity and Exploration Inventory-II: Development, Factor Structure, and Psychometrics. *Journal of Research in Personality* 43, 6 (Dec. 2009), 987–998. https://doi.org/10.1016/j.jrp.2009.04.011

[22] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A Survey of Serendipity in Recommender Systems. *Knowledge-Based Systems* 111 (2016), 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

[23] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval* (Cham) *(Lecture Notes in Computer Science)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[24] Pan Li, Maofei Que, Zhichao Jiang, Yao Hu, and Alexander Tuzhilin. 2020. PURS: Personalized Unexpected Recommender System for Improving User Satisfaction. In *Proceedings of the 14th ACM Conference on Recommender Systems*. Association for Computing Machinery, 10. https://doi.org/10.1145/3383313.3412238

[25] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, and Guojun Wang. 2019. HAES: A New Hybrid Approach for Movie Recommendation with Elastic Serendipity. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1503–1512. https://doi.org/10.1145/3357384.3357868

[26] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 122–132. https://doi.org/10.1145/3366423.3380100

[27] George Loewenstein. 1994. The Psychology of Curiosity: A Review and Reinterpretation. *Psychological Bulletin* 116, 1 (1994), 75.

[28] Qiuxia Lu, Tianqi Chen, Weinan Zhang, Diyi Yang, and Yong Yu. 2012. Serendipitous Personalized Ranking for Top-N Recommendation. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE, Macau, China, 258–265. https://doi.org/10.1109/WI-IAT.2012.135

[29] Valentina Maccatrozzo, Manon Terstall, Lora Aroyo, and Guus Schreiber. 2017. SIRUP: Serendipity In Recommendations via User Perceptions. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. Association for Computing Machinery, New York, NY, USA, 35–44. https://doi.org/10.1145/3025171.3025185

[30] Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. https://doi.org/10.1214/aoms/1177730491

[31] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2145–2148. https://doi.org/10.1145/3340531.3412152

[32] Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2021. Flatter Is Better: Percentile Transformations for Recommender Systems. *ACM Transactions on Intelligent Systems and Technology* 12, 2 (March 2021), 19:1–19:16. https://doi.org/10.1145/3437910

[33] Christian Matt, Thomas Hess, Alexander Benlian, and Christian Weiß. 2014. Escaping from the Filter Bubble? The Effects of Novelty and Serendipity on Users' Evaluations of Online Recommendations. In *Proceedings of the 35th International Conference on Information Systems* (Auckland, New Zealand) *(ICIS)*.

[34] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of personality* 60, 2 (1992), 175–215.

[35] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. *A Survey on Bias and Fairness in Machine Learning*. arXiv:1908.09635 [cs] http://arxiv.org/abs/1908.09635

[36] Alessandro B. Melchiorre, Eva Zangerle, and Markus Schedl. 2020. Personality Bias of Music Recommendation Algorithms. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event Brazil). Association for Computing Machinery, New York, NY, USA, 533–538. https://doi.org/10.1145/3383313.3412223

[37] Aurora Mobile. 2018. *2017 Online Shopping App Market Research Report*. https://community.jiguang.cn/article/246360

[38] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (2019). https://doi.org/10.3389/fdata.2019.00013

[39] Shameem A. Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. 2016. A Coverage-Based Approach to Recommendation Diversity On Similarity Graph. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM Press, Boston, Massachusetts, USA, 15–22. https://doi.org/10.1145/2959100.2959149

[40] Markus Schedl, David Hauger, Katayoun Farrahi, and Marko Tkalčič. 2015. On the Influence of User Characteristics on Music Recommendation Algorithms. In *Advances in Information Retrieval* (Cham) *(Lecture Notes in Computer Science)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Springer International Publishing, 339–345. https://doi.org/10.1007/978-3-319-16354-3_37

[41] Harini Suresh and John V. Guttag. 2020. A Framework for Understanding Unintended Consequences of Machine Learning. (Feb. 2020). arXiv:1901.10002 [cs, stat]

[42] Marko Tkalcic and Li Chen. 2015. Personality and Recommender Systems. In *Recommender systems handbook*. Springer, 715–739.

[43] Horace Walpole. 1960. To Mann, Monday 18 January 1754. In *Horace Walpole's Correspondence*. Yale University Press, 407–411.

[44] Ningxia Wang, Li Chen, and Yonghua Yang. 2020. The Impacts of Item Features and User Characteristics on Users' Perceived Serendipity of Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, 266–274. https://doi.org/10.1145/3340631.3394863

[45] Shanfeng Wang, Maoguo Gong, Haoliang Li, and Junwei Yang. 2016. Multi-Objective Optimization for Long Tail Recommendation. *Knowledge-Based Systems* 104 (July 2016), 145–155. https://doi.org/10.1016/j.knosys.2016.04.018

[46] Wikipedia contributors. 2021. Liebig's Law of The Minimum — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Liebig%27s_law_of_the_minimum&oldid=1002543189 [Online; accessed 28-April-2021].

[47] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction* 28, 3 (2018), 237–276.

[48] Pengfei Zhao and Dik Lun Lee. 2016. How Much Novelty is Relevant?: It Depends on Your Curiosity. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 315–324. https://doi.org/10.1145/2911451.2911488

[49] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. 2021. Popularity-Opportunity Bias in Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 85–93. https://doi.org/10.1145/3437963.3441820