On the Sustainability of Fairness and Bias Interventions in Recommender Systems: A Simulation-based Analysis

Marlene Holzleitner[1,*], Stephan Leitner[1] and Dietmar Jannach[1]

[1]University of Klagenfurt, Universitätsstraße 65-67, 9020 Klagenfurt am Wörthersee, Austria

Abstract

The evaluation of recommendation models in the academic literature mostly adopts a myopic perspective, where both recommendation accuracy and other desirable quality factors, such as fairness or broad catalog coverage, are solely assessed at isolated points in time. In reality, however, the use of recommender systems may lead to undesired, self-reinforcing longitudinal effects, for example, in terms of increased popularity bias or decreased diversity. Such effects can stem from feedback loops that emerge when deploying recommender systems in practice. In this work, we analyze such longitudinal effects, leveraging user modeling, through a simulation-based approach. The main focus of our simulation lies on quality factors of potential societal impact, such as fairness and popularity bias, and we in particular aim to study if existing fairness-enhancing and bias-mitigating intervention strategies have a lasting positive effect. Our longitudinal analyses across various algorithms and datasets reveal that such strategies, when configured appropriately, can indeed be effective in a sustainable manner. Furthermore, our investigations show that more complex models are often not better than simpler approaches in terms of accuracy but can exhibit sustained favorable properties with respect to factors beyond accuracy. Overall, our findings may offer practical insights for selecting recommender models to achieve sustainable success in terms of such societally relevant quality factors.

Keywords

Recommender Systems, Long-term Effects, Simulation, Multi-Metric Evaluation

1. Introduction

Recommender systems have become a ubiquitous part of today's online user experience, serving as information filters and search aids in an often overwhelming digital world [1]. Early models were mainly optimized for accuracy [2]. However, it soon became clear that "accuracy is not enough" [3], and researchers have begun exploring aspects beyond accuracy, including diversity or coverage [4]. More recently, research has focused on undesired effects of recommender systems that may entail negative societal impacts, such as biases and limited fairness [5]. Accordingly, various technical approaches were proposed to enhance algorithmic fairness and mitigate biases, see [6, 7, 8, 9, 10] for related survey works.

However, "beyond-accuracy" approaches that try to counteract biases [11, 12, 13, 14] and ensure fairness [15, 16] are usually assessed in myopic offline evaluation setups. These setups use random or temporal data splits and involve different computational metrics to gauge the accuracy, diversity, popularity or fairness of the generated recommendations at a single point in time. In real-world settings, however, the data that is used by recommender systems evolves over time. Importantly, deploying recommender systems in practice can create self-reinforcing feedback loops where the observed user interactions are at least partially influenced by what the system recommends [17, 18]. With our work, we continue the line of research investigating the topic of beyond-accuracy considerations and go beyond common myopic evaluation perspectives.

Studying such longitudinal effects requires alternative evaluation approaches. *Simulation-based* techniques have emerged as a promising direction. Based on assumptions about how recommendations influence user behavior and how the observed user behavior in turn impacts the system in a feedback

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author.

marlene.holzleitner@aau.at (M. Holzleitner); stephan.leitner@aau.at (S. Leitner); dietmar.jannach@aau.at (D. Jannach)

0009-0005-0715-922X (M. Holzleitner); 0000-0001-6790-4651 (S. Leitner); 0000-0002-4698-8507 (D. Jannach)

published 2026-05-28

loop, such techniques allow us to study various longitudinal phenomena of recommender systems [19, 20, 21, 22]. However, existing research on longitudinal effects rarely addresses biases and fairness to a large extent, and does also not study to what extent algorithmic interventions have a lasting effect.

Thus, the need to study popularity bias and fairness ('4*Good*') [23] aspects over time is evident, since potential reinforcement effects could make these unfair conditions even stronger, and niche content can vanish without an initial opportunity [24, 25, 26]. For example, movies [27] or songs by newcomers [28] may receive less visibility than more popular ones in the initial state, possibly perpetuating inequalities in the long run. A goal actively requested by artists in the music industry [29] is to improve visibility for niche items to overcome such social unfairness. Research into effective intervention strategies for achieving this goal in the long term however remains limited.

With this present work, we aim to narrow this research gap. Specifically, our work aims to provide important insights into the potential harm or undesirable long-term effects of recommendation algorithms before they are deployed in practical settings. In particular, we address the following two research questions:

RQ1: What are the longitudinal effects of continuous user feedback on recommender algorithms' accuracy, beyond-accuracy, and 4*Good* metrics across datasets and domains?

RQ2: To what extent do intervention strategies improve 4*Good* aspects over time, and are these benefits lasting?

Specifically, building on [20], we conducted comprehensive simulations, involving multiple algorithms and datasets, which allowed us to perform a multi-dimensional analysis of algorithmic effects over time. Unlike earlier works, we include recent 4*Good* metrics (RQ1). In addition, going beyond previous simulation-based approaches, we consider two established intervention strategies to mitigate biases and ensure fairness in our simulations (RQ2).

Indeed, a main finding of our work is that the intervention strategies from the literature, when configured appropriately, can have a lasting effect. Modern neural models can sometimes underperform in accuracy compared to earlier techniques, but they perform well in catalog coverage, popularity bias, and fairness. On a more general level, we confirm previous findings that some recommendation techniques can actually lead to an undesired decrease in diversity and an increase in popularity bias over time. Overall, our findings offer guidance for selecting and configuring recommendation algorithms in practice.

The paper is structured as follows. In Section 2, we review prior works on long-term simulation approaches. Section 3 presents our simulation approach and experimental setup, including the applied algorithms, datasets, and metrics. In Section 4, we present and discuss the results and key insights from the simulation. Finally, Section 5 discusses the relevance of the results as well as their limitations, and provides directions for future work.

2. Related Work

While studies on recommender systems that rely on purely computational experiments dominate in the literature, works on their longitudinal effects are sparse. Below, we review selected studies that are most relevant to our work.

Jannach et al. [20] present one of the earliest works in this area that employ simulation-based approaches. The authors studied popularity bias and concentration effects arising from feedback loops of recommender systems by comparing 12 algorithms on movie and music datasets. A key insight of their work is that different recommender algorithms with quite similar accuracy can lead to largely diverging bias patterns. Our present study builds on the simulation methodology established in this work[1]. Building on [20], we extend this line of research by (a) involving more recent recommendation

[1]Similar simulation setups were later used, e.g., in [30] or [19], yet none of these studies considers fairness metrics as a target variable of the simulation.

techniques, including neural models, and (b) placing a particular emphasis on fairness aspects, which have not been addressed extensively in previous simulation-based approaches.

Zhang et al. [19] use a related simulation approach with a different focus. They employ agent-based modeling (ABM) [31] to capture user-level dynamics. Specifically, they study longitudinal effects of collaborative filtering under varying levels of user reliance on individual recommendations, and evaluate accuracy, diversity, and relevance. They find that greater reliance on recommendations decreases consumption diversity, i.e., the Gini Index increases. They also observed what they call a "performance paradox": while stronger reliance on recommendations improves the relevance of consumed items over time, it also decreases consumption diversity and leads to diminishing accuracy improvements. While both the work in [19] and our study employ simulation-based approaches, we focus on a system-level analysis and [19] study individual user-level dynamics.

Ghanem et al. [21] applied a related ABM approach to study recommender systems from an economic perspective. They analyzed the longitudinal effects of different strategies that *providers* of a recommendation service may adopt. Specifically, the strategies varied in how consumer and provider values are balanced. The main outcome variables were the development of consumer trust in the recommendation system over time and the cumulative profit for the providers. Hybrid strategies that focus on consumer value but still account for profitability lead to the most favorable long-term outcomes. While methodologically related, their focus was economic sustainability, whereas we consider aspects of potential societal relevance more directly.

Similarly to [21], Buhayh et al. [32] consider a multi-stakeholder perspective (consumer, provider, and system) and simulate the effects of users' ability to switch between recommender algorithms to select the recommender which works best for them. Consumers can choose between a "generic" content-based recommender and a niche content-based recommender [33], which focuses on a single niche genre. The authors modeled users' selection based on the perceived utility of the recommendations via two different decision models. Niche-item consumers and their providers benefit from this non-monolithic setting. While their work simulates the complexity of multi-stakeholder interactions, our study focuses on consumers and examines intervention methods applied to the algorithms rather than active algorithm selection by users.

Hazrati and Ricci [34, 35] shift focus to consumers and analyze how recommendations affect their choices. In their simulation framework in [34], each user is characterized by a *choice model* that is the basis for repeated item selections. The recommender system adapts its subsequent recommendations to these choices, thereby creating a dynamic feedback loop. They focus mainly on choice diversity and item relevance, and how personalization affects them. In a subsequent work [35], the authors explore three alternative choice strategies, in which users tend to primarily select popular, recent, or highly-rated items. Beyond this, Ungruh et al. [36] go a step further and examine user choice models, uniquely for each user, to reflect realistic and diverse consumption patterns. Differently from their focus on the user-level, we aim to assess the properties of the system's aggregate recommendations over time.

While many simulation studies address e-commerce or media streaming, there are also works that study longitudinal effects of recommendations in other areas. Akpinar et al. [37], for example, study fairness dynamics via simulations and focus on (user-to-user) *connection recommendations* that are common on platforms like LinkedIn. They find a rich-get-richer effect that favors the majority (male) user group. Although fairness interventions improve overall fairness, minority groups remain disadvantaged in the long run. Their focus on fairness aligns with ours. We extend this line of research by a multi-metric analysis that also considers further indicators, including prediction accuracy, popularity bias, and coverage.

Vandeputte et al. [38] focus on longitudinal effects of nutrition recommender systems, where the system acts as a *coach* designed to support healthier food choices; whenever a user makes a food choice, the recommender system suggests an appropriate modification of the selection. After each interaction, users can adapt their preferences, and the coaching system can adapt the strategy. The simulation models the interaction as iterated two-player game. The results suggest that non-myopic strategies are more effective in stimulating behavioral changes. Our work is related to [38] in its focus on societal

value (fairness vs. healthier eating habits). Unlike their work, we employ simulation to understand emerging system behavior rather than user preference change.

Finally, other simulation forms include reinforcement learning frameworks ("gyms") [39, 40, 41] and, recent Generative AI-based user simulations, e.g., [42, 43, 44]. The relation of these studies to our work is however limited.

3. Methodology

In this section we describe our methodological approach, which is based both on the reviewed literature and on further theoretical considerations.

Simulation Model. Our general research approach relies on simulation principles, inspired by [20, 21, 35]. The main idea is to repeatedly learn, recommend, and add new interactions from recommendations. The core of the simulation lies in how we model the behavior of the users, when they are provided with recommendations by the system. Specifically, the user's choice model is constructed as follows. The selection of recommended items is modeled via a uniform distribution, meaning that every presented item has the same probability of being chosen by a user, i.e., we do not assume any specific presentation-dependent positional bias[2]. Once users select items, they are assumed to provide ratings, which are fed into the underlying rating database. The rating values for the chosen items are drawn from a normal distribution with $\mu = 4$ and $\sigma = 0.3$. We set these parameters to model a tendency to give feedback rather to items the user liked, which reflects a specific case of selection bias [8]. For each user and iteration, a list of top-30 recommendations is generated and evaluated, where in every iteration, users pick and rate exactly one item from their personalized list. We chose a number of 30 recommended items per user to reflect real-world dynamics, where users do not necessarily restrict their choices to the top-10 list, but explore the space of alternatives in slightly more depth [45, 46].

Recommendation and Intervention Methods. To answer our two RQs, we look at eight algorithms from different families, and we consider two widely-used intervention strategies to ensure fairness and mitigate popularity bias.

In terms of algorithm selection, we adopt the experimental design from [47, 48], considering two algorithms from each of the following families. As neighborhood-based and simple graph-based models, we consider UserKNN [49] and RP$^3\beta$ [50]. EASE^R [51] and SLIM [52] represent the category of linear models. To cover matrix factorization models, we choose BPRMF [53] and iALS [54]. As neural models, NeuMF [55] and Mult-VAE [56] are selected. As a non-personalized baseline we use MostPop.[3] Optimal hyperparameter settings were determined for each algorithm and dataset at the beginning of the simulation based on the procedure from [47].

As intervention methods, we consider xQuAD [11] and Calibration [15]. xQuAD [11] is a method that re-ranks a given accuracy-optimized recommendation list under additional consideration of popularity bias. For the re-ranking process, the available items are divided into *short head* and *long tail* items. As in [11], the short head is defined as the smallest set of the most popular items that together account for 80 % of all transactions. The remaining items form the long tail.

As a second re-ranking strategy, we use Calibration [15]. We note that the term calibration is frequently applied in the context of algorithmic fairness [57, 15], which fits with one of our main goals of understanding fairness effects from a longitudinal perspective. Specifically, in the implemented calibration approach, the popularity distributions of users' historical interactions and recommendations are compared via the Kullback-Leibler (KL) divergence (the lower the more similar the distributions,

[2]In real-world applications (such as YouTube), the order of recommendations is commonly shuffled slightly to avoid the impression of monotone recommendations.

[3]Additional information and descriptions of the algorithms and datasets are provided in the online material at https://github.com/MarleneHlr/Longitudinal_Simulation.git.

which is desired). In the original Calibration approach, genre distributions are taken into account, while in our experimental setup, we categorize the items into popularity bins.

In each of the two intervention methods, a weight parameter w that is set between 0 and 1 controls the strength of the re-ranking, i.e., the trade-off between accurate and diverse (xQuAD) or popularity-calibrated (Calibration) recommendations. In our experiments, we explore different weight parameters $w = [0.2, 0.4, 0.6, 0.8]$ for exploration. A higher value of w corresponds to a greater intervention influence.

Main Simulation Loop. Algorithm 1 shows the core simulation procedure, implemented in Python using the Elliot framework [58]. We share all our code and data online. Further simulation parameters, the optimal hyperparameters for each algorithm and datasets, and our complete set of simulation results are provided in the online material as well.

Algorithm 1 Longitudinal Simulation (one repetition)

```
procedure SIMULATION(input: dataset, algorithms, number of iterations, distribution parameters
for new ratings, intervention strategies with weights w)
    for number of iterations do
        Retrain algorithms and create recommendation lists
        for all users
        if intervention strategies ≠ ∅ then
            for all strategy ∈ intervention strategies do
                Rerank recommendation lists using strategy
                with weight w
            end for
        end if
        for all user ∈ users do
            Select and rate recommended item according to
            distribution parameters
            Update dataset with new rating
        end for
    end for
end procedure
```

Datasets and Prefiltering. We consider four publicly available datasets from the music, movie, and e-commerce domains to study a broader set of use cases in which personalization can have a societally relevant impact on user behavior. Three datasets are chosen based on [47]: MovieLens-1M[4], with over 1,000,000 ratings from 6,040 users on 3,706 movies, Amazon Music[5], with 1,584,082 ratings from 840,372 users on 456,992 songs and Epinions[6], with 300,548 ratings from 8,514 users on 8,510 items using only the ratings from "trustworthy" consumers. We include the smaller MovieLens-100k[7] dataset in addition, as this dataset is frequently used for evaluating computationally complex models in the literature. It comprises 100,000 ratings from 943 users on 1,682 movies.

As usually done in the related body of literature, we applied a preprocessing step in which we considered ratings ≥ 4 as positive implicit signals and removed all ratings below that global threshold. This procedure was applied to MovieLens-100K, MovieLens-1M, and Amazon Digital Music, as they comprise ratings on a 1-5 scale. For Epinions, a binary dataset, we simply retained the positive interactions. In a second pre-filtering step, users and items with less than 10 interactions were removed

[4] https://grouplens.org/datasets/movielens/1m/
[5] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/
[6] http://www.trustlet.org/downloaded_epinions.html
[7] https://grouplens.org/datasets/movielens/100k/

from MovieLens-100K and MovieLens-1M. In Amazon Digital Music, users with at least 5 interactions and items with at least 5 interactions were kept. In Epinions, the minimum number of interactions regarding users and items was 2. The minimum number of interactions chosen varied due to the density and size of the datasets. A lower number was used for sparser datasets. We then followed the commonly used strategy on rating datasets and binarized the three not yet binary ones. We also removed possible duplicates in all four datasets [47]. Table 1 summarizes the characteristics of the datasets after pre-processing and filtering as we used them at the beginning of the simulation.

Data Splitting Our simulation process involves repeated model training and evaluation after the new ratings of the previous step were added, see Algorithm 1. For this process, we apply a common random train-test split ratio of $80/20$.[8]

Table 1
Dataset characteristics at the beginning of the simulation (i.e., after preprocessing).

Dataset	#Users	#Items	Ratings	Ratings/User	Sparsity
MovieLens-100k	887	822	52,764	59.49	0.927
MovieLens-1M	5,949	2,810	571,531	96.07	0.965
Amazon Music	10,631	8,594	104,546	9.83	0.998
Epinions	8,485	8,462	300,017	35.36	0.995

Metrics. We use accuracy, beyond-accuracy, and 4*Good* metrics to assess algorithm performance at every timestep, resulting in a multi-metric evaluation over time; Table 2 provides an overview of the used metrics. Generally, we repeated the simulations 5 times, and we report average results in Section 4.

Table 2
Overview of metrics used in the experiments.

Accuracy	
NDCG	*Normalized Discounted Cumulative Gain* [47], a commonly used accuracy metric.
Beyond Accuracy	
Coverage	*Coverage* reflects the total number of unique recommended items across all users, see [20].
Diversity	*Gini Index*: Quantifies the distribution inequality in the dataset. Lower values indicate lower inequality and are desirable [59].
4Good	
Popularity Bias	*ARP* (Average Recommendation Popularity): Measures the popularity bias of the recommendations based on the general popularity of each item in the dataset [60]. Lower values are considered better. *ACLT* (Average Coverage of Long-tail Items): Quantifies the fraction of non-popular items the recommender algorithm has covered. We recall that we rely on the definition from [11] to distinguish long-tail from short-head items. Higher values are desirable.
Fairness	*PopREO* (Popularity-based Ranking-based Equal Opportunity): Assesses whether algorithm performance (measured via *Recall*)) is equal across groups of popular and non-popular items [24]. Lower values indicate higher fairness.

[8]We note that this procedure intentionally does not reflect temporal consumption developments. Instead, the random splitting procedure allows us to preserve the overall user and item distribution during the simulation and leads to a consistent evaluation approach as the dataset grows.

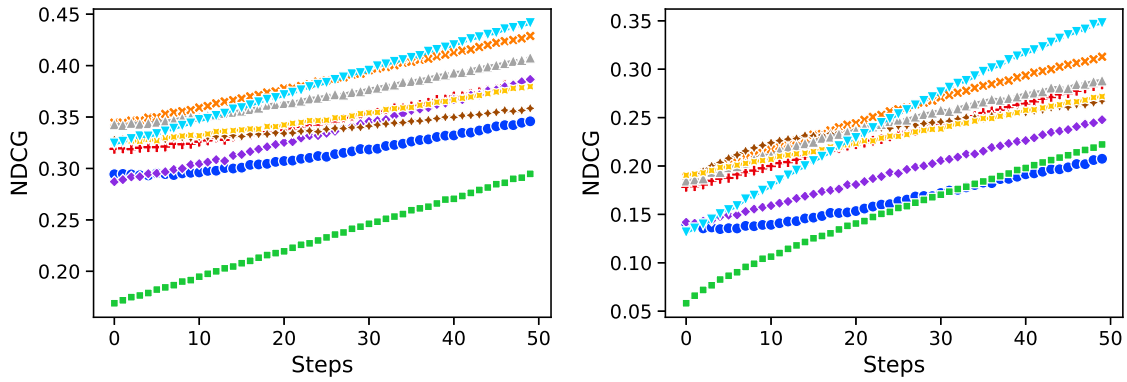

Figure 1: NDCG over time for MovieLens-1M (left) and Epinions (right).

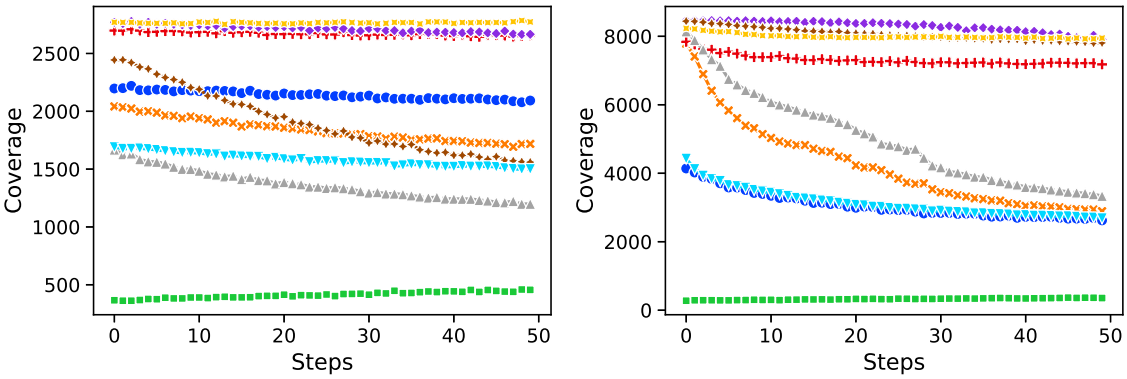

Figure 2: Coverage over time for MovieLens-1M (left) and Epinions (right).

4. Evaluation & Discussion

In this section, we report results for one dense dataset (MovieLens-1M) and one sparse dataset, (Epinions), while only highlighting similar or different trends in Amazon Digital Music and MovieLens-100k. As our experiments use dynamically evolving datasets based on simulated feedback loops, traditional statistical tests for comparisons between recommender algorithms cannot be applied. We therefore focus on identifying longitudinal trends. This evaluation approach is consistent with previous simulations in the literature [21, 20, 19].

Accuracy. In line with previous studies [61, 62, 51], we find that simple models, and in particular linear models, perform surprisingly well in accuracy on both datasets. This observation is stable over time, see Figure 1. In the long run, iALS achieves the highest accuracy, as model confidence grows with user-item interactions [54]. Notably, the more recent NeuMF approach does not reach similar performance levels and also does not profit from the growing dataset. Similar trends regarding accuracy can be observed for the MovieLens-100k and Amazon Digital Music datasets. We note that we expect generally increasing NDCG values across models due to a growing number of interactions, both per user and overall.

Beyond Accuracy. Figure 2 shows the *Coverage* results. The neural models (NeuMF and Mult-VAE) and UserKNN achieve the highest coverage on both datasets. At the same time, as indicated above, NeuMF's accuracy is only medium at best, while UserKNN and Mult-VAE lead to medium to high accuracy. In contrast, the models that were best in accuracy (iALS and EASE^{R}) are at the lower end of coverage, indicating a trade-off between coverage and accuracy.

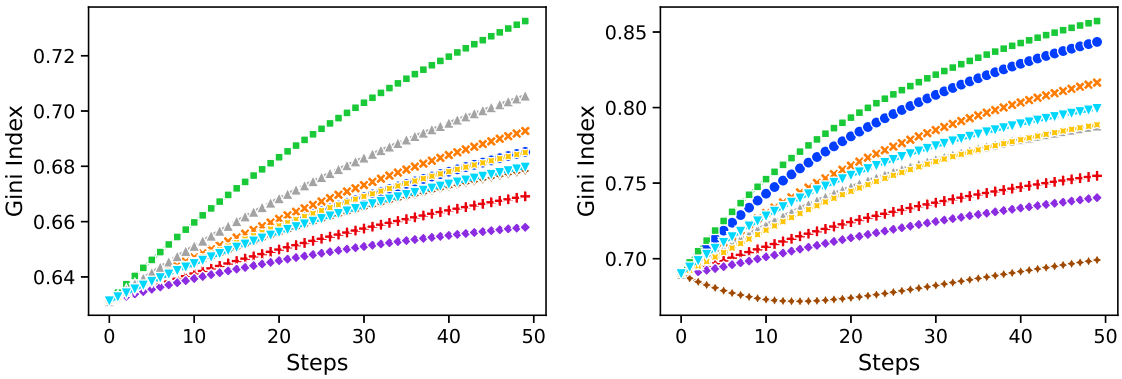

Figure 3: Gini Index over time for MovieLens-1M (left) and Epinions (right).

Furthermore, the performance in terms of coverage may depend on dataset characteristics. The RP$^3\beta$ method, for example, is among the top performers on the sparse Epinions dataset, but its coverage is mediocre and degrades fast on the dense MovieLens-1M dataset. Similarly, the coverage of BPRMF is relatively low for the sparse dataset. Generally, we find that several algorithms, including the highly-accurate linear models EASE^R and SLIM can lead to a decreased coverage over time. This indicates that these models learn to focus on a smaller set of items over time, leading to the potentially undesired effect that some items of the catalog are not recommended to anyone at all. Considering broader social implications, we note that there are various domains where it is important that all items in the catalog receive a certain level of exposure through recommendations, e.g., on job platforms [63]. Similar trends regarding coverage can be observed for the MovieLens-100k and Amazon Digital Music datasets, again with stronger coverage degradations for the sparse Amazon dataset.

Our *Diversity* results in terms of the Gini Index are shown in Figure 3. The findings align well with those for coverage in that the neural models exhibit a favorable behavior. On both datasets, their Gini Index values are on the lower end, indicating limited concentration effects and good diversity. Likewise, the linear models tend to lead to a stronger concentration effect over time, which may also be seen as a reduced level of personalization [20]. Again, dataset characteristics seem to play a role. On the sparse Epinions dataset, BPRMF exhibits a quite strong concentration tendency, whereas RP$^3\beta$ initially even leads to an increase in diversity and only a slow decrease afterwards. Overall, however, we find that all algorithms, somewhat paradoxically, induce a *decrease* of aggregate diversity, despite being designed to cater for individual preferences. This finding confirms earlier observations from [64]. An exception is NeuMF on the small MovieLens-100k dataset, where we observed that NeuMF is actually able to consistently increase diversity. Furthermore, for the Amazon Music Dataset, we found that RP$^3\beta$ initially exhibited a similar pattern as on Epinions, but diversity decreased significantly over time.

4Good Metrics. We start our discussion of 4*Good* metrics with an analysis of potential popularity bias in the algorithms, including the two intervention strategies. Figure 4 shows the outcomes for the *ARP* metric. Generally, we expect all ARP values to grow over time, given that new ratings are added to the data in each simulation step. The observed increase in popularity is however over-proportional. For a random recommender system, where each item has a uniform chance to be recommended, the popularity of each item should increase in each round by one (Epinions) or two (MovieLens-1M) items per simulation step. For Epinions, for example, we have 8,485 users and 8,462 items. Adding one new rating per user and simulation round means 8,485 new ratings would lead to an average increase in ARP per item of about 1 ($\frac{8{,}485}{8{,}462} \approx 1$). Looking at the results in Figure 4, we however see that the ARP of the recommended items increases markedly stronger, indicating an increasing popularity bias.

Generally, the differences between algorithms in terms of popularity bias can be substantial, and sometimes dependent on the dataset characteristics. iALS generally exhibits low bias on both datasets in Figure 4, whereas BPRMF, as observed also in [20], commonly has a strong tendency to recommend

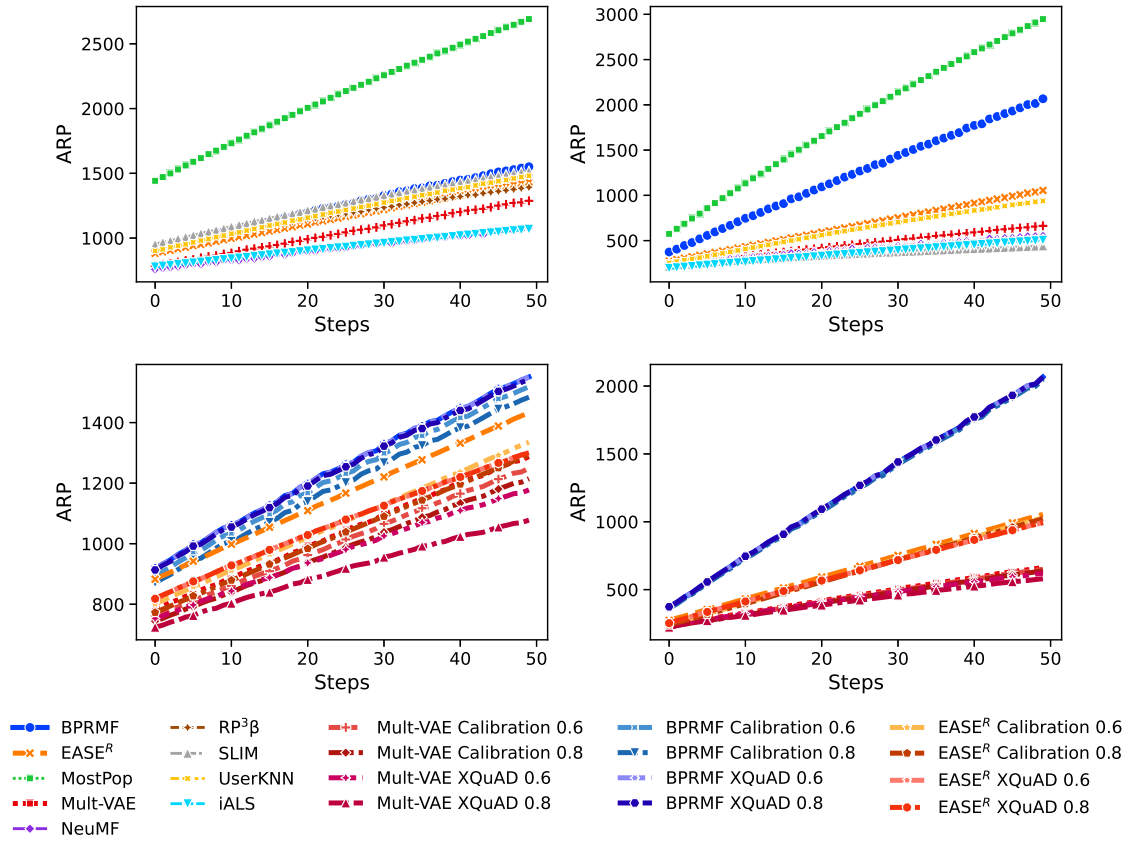

Figure 4: ARP over time for MovieLens-1M (left) and Epinions (right): Plain algorithm results shown in top row, results after interventions in bottom row.

popular items. The performance of other methods like SLIM is varying across the datasets. On the sparse Epinions dataset, the popularity reinforcement of BPRMF is particularly striking. We found similar popularity bias results for the MovieLens-100k and Amazon Music datasets. In particular, SLIM shows a high popularity bias for the former dataset and a relatively low bias for the latter dataset.

The effectiveness of bias mitigation can best be seen for the MovieLens-1M dataset. Focusing on the BPRMF and the EASE^R methods as baselines, we see that the effects are lasting over time, i.e., the methods effectively reduce the bias throughout the simulation horizon. We however observe that relatively high influence weights (0.6, 0.8) have to be used to achieve a notable effect. On the sparse Epinions dataset, the effectiveness of the intervention strategies can be quite limited, calling for alternative intervention strategies for sparse datasets. The intervention strategies lead very similar results for MovieLens-100k and Amazon Music, respectively.

Regarding the ACLT metric, see Figure 5, we observe that the neural models are most effective in recommending items from the long tail for both datasets. The models with highest accuracy values, e.g., iALS, SLIM and EASE^R, on the other hand, almost never recommend items with very low popularity. An interesting observation is that iALS is very good at identifying and recommending items that are not 'blockbusters' (low ARP), but at the same time is effective in avoiding items that may be too niche (low ACLT). The general patterns are similar for both datasets.[9] For the MovieLens-100k and Amazon Digital Music datasets, we found that iALS ultimately leads to ACLT results that are competitive with those of the neural models, a phenomenon we did not observe for the other datasets. Considering the intervention strategies, we again see that the interventions have a sustained effect, with the xQuAD method often being more effective than Calibration. Depending on the algorithm and dataset characteristics, the effects can however be modest, even when comparably high weight values

[9]Only RP$^3\beta$ stands out for the sparse Epinions dataset, which starts with a strong coverage of long-tail items, but after a few simulation steps increasingly moves to recommending more popular items.

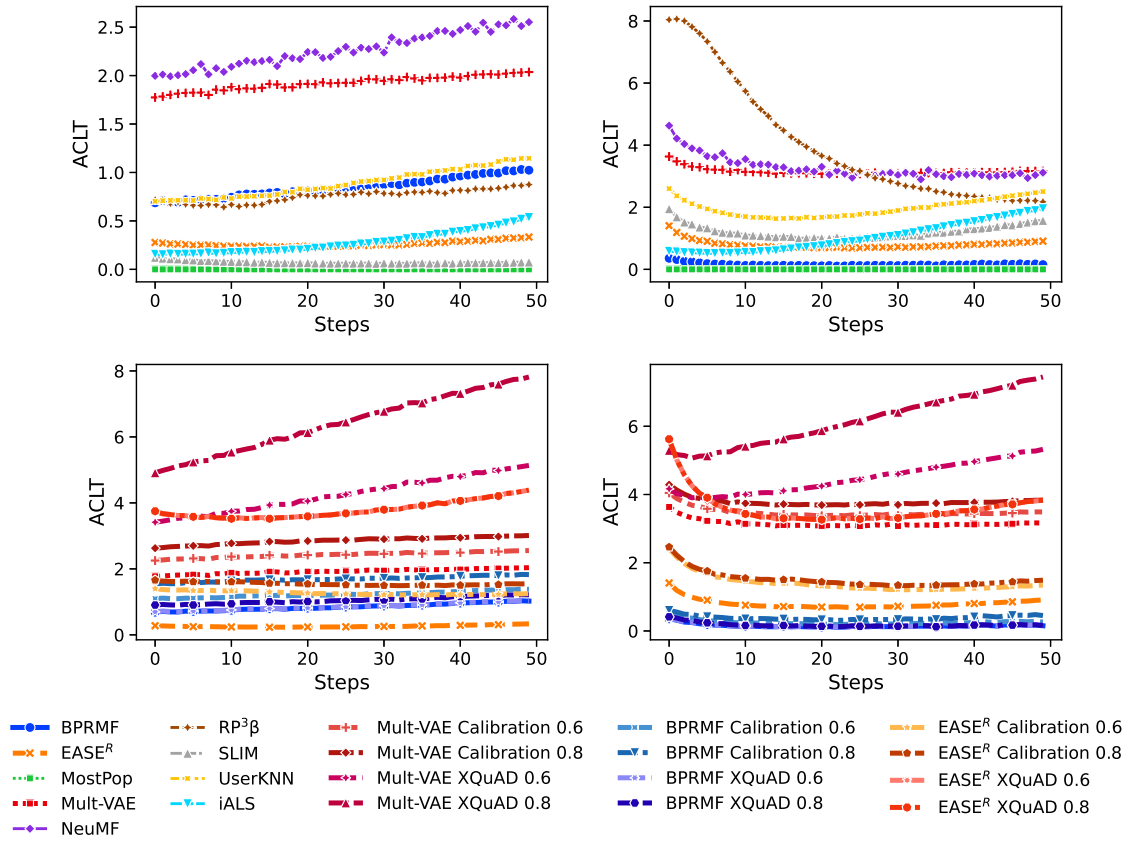

Figure 5: ACLT over time for MovieLens-1M (left) and Epinions (right): Plain algorithm results shown in top row, results after interventions in bottom row.

are chosen. For better readability, we show the effects of the interventions separately in the lower row of Figure 5. For the MovieLens-100k and Amazon Digital Music datasets, xQuAD consistently outperforms Calibration over time.

The *Fairness* results in terms of the PopREO metric are finally shown in Figure 6. Again, looking at the MovieLens-1M dataset, we can observe that high accuracy, as achieved for example by iALS or EASE^R, can stand in contrast with fairness properties of a model. Neural models like Multi-VAE can be favorable from a fairness perspective, and their level of fairness can be further increased through an appropriate fairness-enhancing intervention. In fact, combining Mult-VAE with the xQuAD strategy leads to the best fairness effects on both datasets. However, the fairness results are less consistent for the very sparse Epinions dataset. In some cases, e.g., for BPRMF, a method with high popularity bias, the fairness of the recommendations can even decrease over time, and the fairness intervention can have limited effects even when relatively high weights are chosen. Similar fairness results are also observed in Amazon Digital Music, where a combination of Mult-VAE and xQuad performs best. For MovieLens-100k, in contrast, EASE^R with xQuAD actually leads to the best fairness result. Interestingly, on this small dataset, iALS is able to balance the accuracy-fairness trade-off very well over time.

5. Implications, Limitations and Outlook

The literature shows that one can easily overlook important undesirable effects of algorithm recommendations, e.g., strong biases and low fairness, when focusing exclusively on accuracy metrics. In this work, we analyzed such effects from a longitudinal perspective. Our analyses revealed strong differences across algorithms when applying a multi-metric analysis. A main result of our simulations is that existing intervention methods like xQuAD or Calibration, when configured appropriately, can have a *sustained effect*. A societal implication of this finding thus is that properly configured interventions

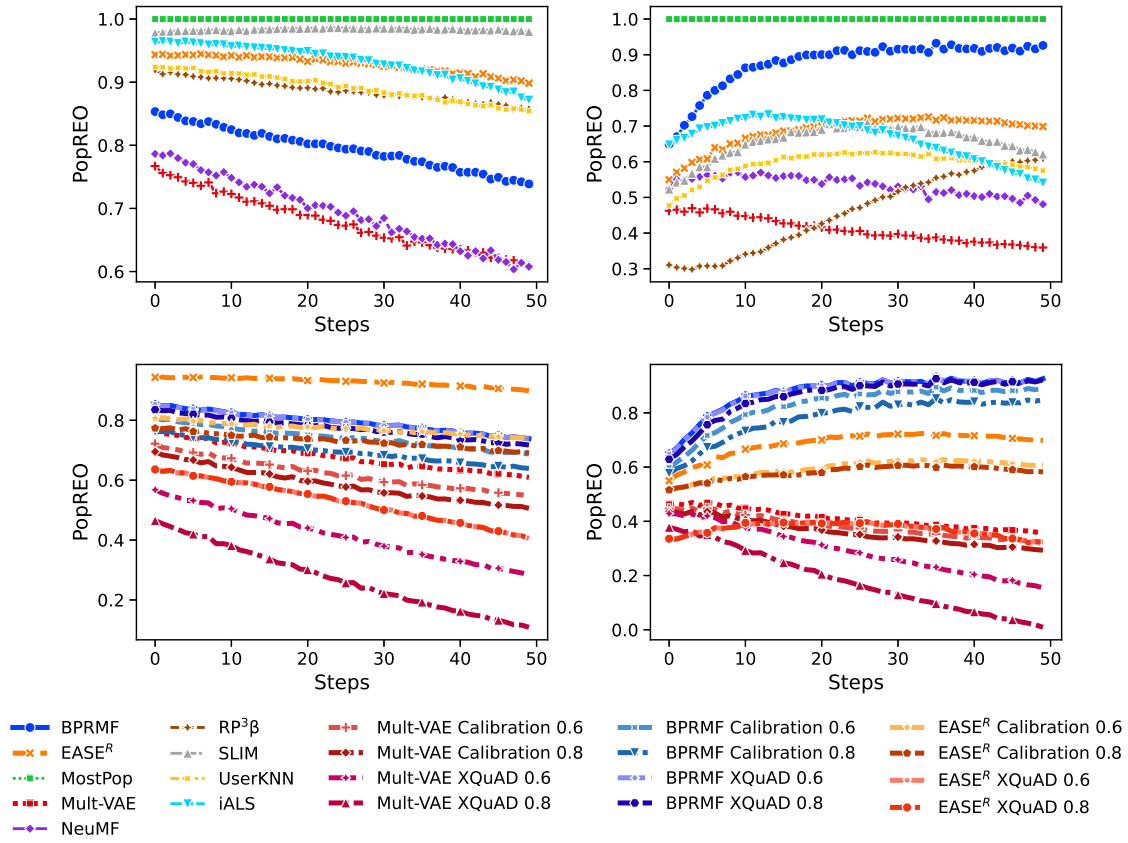

Figure 6: PopREO over time for MovieLens-1M (left) and Epinions (right): Plain algorithm results shown in top row, results after interventions in bottom row.

can lead to improved exposure opportunities for underrepresented items in the long term. In the music domain, for example, songs from emerging artists or from less popular genres, can gain and maintain greater visibility through these interventions over time.

From a practical perspective, it is generally important to understand the relative importance of different recommendation qualities in a given use case (e.g., accuracy vs. fairness). Our analysis however, also showed that the characteristics of the dataset can largely impact both the effectiveness of the models and the intervention strategies. Thus, careful analyses are required before deploying models in practice, and the work presented in our paper aims to provide practical guidance in such situations.

Our present work does not come without limitations. Generally, like previous works, our simulation approach is based on certain assumptions regarding item consumption, including simplifications of modeling user choices, i.e., we neither consider individual user behavior patterns nor different user selection and rating propensity, which may not fully capture real-world dynamics [36]. Additionally, users in our simulation cannot switch between recommender algorithms based on the utility of the recommendations they receive [32]. Besides, the observed reinforcement patterns may have a different strength in practice.

Furthermore, we are aware that fairness is a complex and multi-faceted construct, which cannot be easily captured through computational metrics alone [7, 65, 66]. Thus, our research shares limitations of other works that rely on computational metrics, which only serve as a rough proxy for fairness aspects in reality. Moreover, the widely used metrics we report are similar and likely correlated. Nonetheless, despite these limitations, we are confident that our work represents an important step towards a better understanding of longitudinal effects of recommender systems. Furthermore, our simulation-based approach to modeling user behavior provides a contribution to ongoing research on social unfairness, without facing potential ethical issues of experiments with real users.

In terms of future work, we, for example, plan to consider a dynamic catalog of users and items, and

we will consider demographic information like (e.g., age, gender) [67] to enable subgroup and more detailed individual fairness analyses.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check, Improve writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] D. Jannach, P. Pu, F. Ricci, M. Zanker, Recommender systems: Past, present, future, in: AI Magazine, volume 42, 2021, pp. 3–6. doi:https://doi.org/10.1609/aimag.v42i3.18139.

[2] U. Shardanand, P. Maes, Social Information Filtering: Algorithms for Automating "Word of Mouth", in: CHI '95, 1995, pp. 210–217. URL: https://doi.org/10.1145/223904.223931.

[3] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: How accuracy metrics have hurt recommender systems, in: CHI '06 Extended Abstracts on Human Factors in Computing Systems, CHI EA '06, 2006, pp. 1097–1101. doi:10.1145/1125451.1125659.

[4] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems (TOIS) 22 (2004) 5–53. doi:10.1145/963770.963772.

[5] K. Dinnissen, C. Bauer, Fairness in music recommender systems: A stakeholder-centered mini review, Frontiers in Big Data Volume 5 - 2022 (2022). URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.913608. doi:10.3389/fdata.2022.913608.

[6] A. Klimashevskaia, D. Jannach, M. Elahi, C. Trattner, A survey on popularity bias in recommender systems, User Modeling and User-Adapted Interaction 34 (2024) 1777–1834. doi:10.1007/s11257-024-09406-0.

[7] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo, D. Zanzonelli, Fairness in recommender systems: Research landscape and future directions, User Modeling and User-Adapted Interaction 34 (2023) 59–108. doi:10.1007/s11257-023-09364-z.

[8] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, ACM Trans. Inf. Syst. 41 (2023). doi:10.1145/3564284.

[9] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. 41 (2023). doi:10.1145/3547333.

[10] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness in information access systems, Found. Trends Inf. Retr. 16 (2022) 1–177. doi:10.1561/1500000079.

[11] H. Abdollahpouri, R. Burke, B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking, in: The Florida AI Research Society, 2019. URL: https://api.semanticscholar.org/CorpusID:59158829.

[12] G. Adomavicius, Y. Kwon, Improving aggregate recommendation diversity using ranking-based techniques, IEEE Transactions on Knowledge and Data Engineering 24 (2012) 896–911. doi:10.1109/TKDE.2011.15.

[13] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, J. Caverlee, Popularity-opportunity bias in collaborative filtering, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 85–93. URL: https://doi.org/10.1145/3437963.3441820. doi:10.1145/3437963.3441820.

[14] A. Bellogín, P. Castells, I. Cantador, Statistical biases in information retrieval metrics for recommender systems, Information Retrieval Journal 20 (2017) 606–634. URL: https://doi.org/10.1007/s10791-017-9312-z. doi:10.1007/s10791-017-9312-z.

[15] H. Steck, Calibrated recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, 2018, p. 154–162. doi:10.1145/3240323.3240372.

[16] N. Sonboli, F. Eskandanian, R. Burke, W. Liu, B. Mobasher, Opportunistic multi-aspect fairness through personalized re-ranking, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 239–247. URL: https://doi.org/10.1145/3340631.3394846. doi:10.1145/3340631.3394846.

[17] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, Feedback loop and bias amplification in recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, 2020, p. 2145–2148. doi:10.1145/3340531.3412152.

[18] R. Jiang, S. Chiappa, T. Lattimore, A. György, P. Kohli, Degenerate feedback loops in recommender systems, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19, 2019, p. 383–390. doi:10.1145/3306618.3314288.

[19] J. Zhang, G. Adomavicius, A. Gupta, W. Ketter, Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework, Info. Sys. Research 31 (2020) 76–101. doi:10.1287/isre.2019.0876.

[20] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, User Modeling and User-Adapted Interaction 25 (2015) 427–491. doi:10.1007/s11257-015-9165-3.

[21] N. Ghanem, S. Leitner, D. Jannach, Balancing consumer and business value of recommender systems: A simulation-based analysis, Electronic Commerce Research and Applications 55 (2022). doi:https://doi.org/10.1016/j.elerap.2022.101195.

[22] M. Zhou, J. Zhang, G. Adomavicius, Longitudinal impact of preference biases on recommender systems' performance, Information Systems Research 35 (2023) 1634–1656. doi:10.2139/ssrn.3799525.

[23] D. Jannach, A. Said, M. Tkalcic, M. Zanker, Recommender Systems for Good (RS4Good): Survey of Use Cases and a Call to Action for Research that Matters, ACM Trans. Recomm. Syst. (2025). doi:10.1145/3746648.

[24] Z. Zhu, J. Wang, J. Caverlee, Measuring and mitigating item under-recommendation bias in personalized ranking systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, 2020, p. 449–458. doi:10.1145/3397271.3401177.

[25] Y. Zhao, Y. Wang, Y. Liu, X. Cheng, C. C. Aggarwal, T. Derr, Fairness and diversity in recommender systems: A survey, ACM Trans. Intell. Syst. Technol. 16 (2025). doi:10.1145/3664928.

[26] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, Information Processing & Management 58 (2021) 102666. URL: https://www.sciencedirect.com/science/article/pii/S0306457321001540. doi:https://doi.org/10.1016/j.ipm.2021.102666.

[27] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The unfairness of popularity bias in recommendation, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019, volume 2440 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2440/paper4.pdf.

[28] A. Ferraro, X. Serra, C. Bauer, What is fair? exploring the artists' perspective on the fairness of music streaming platforms, in: Human-Computer Interaction – INTERACT 2021, Springer International Publishing, Cham, 2021, pp. 562–584. doi:https://doi.org/10.1007/978-3-030-85616-8_33.

[29] K. Dinnissen, C. Bauer, Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms, in: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, UMAP '23, 2023, p. 238–249. doi:10.1145/3565472.3592960.

[30] A. Ferraro, D. Jannach, X. Serra, Exploring longitudinal effects of session-based recommendations, in: Proceedings of the 2020 ACM Conference on Recommender Systems (RecSys '20), 2020.

doi:10.1145/3383313.3412213.
[31] E. Bonabeau, Agent-based modeling: Methods and techniques for simulating human systems, Proceedings of the National Academy of Sciences 99 (2002) 7280–7287. doi:10.1073/pnas.082080899.
[32] A. Buhayh, E. McKinnie, C. Canel, R. Burke, Simulating the algorithm store: Multistakeholder impacts of recommender choice, in: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 274–279. URL: https://doi.org/10.1145/3708319.3733705. doi:10.1145/3708319.3733705.
[33] A. Buhayh, E. McKinnie, R. Burke, Decoupled recommender systems: Exploring alternative recommender ecosystem designs, in: RecSoGood@RecSys, 2025. URL: https://api.semanticscholar.org/CorpusID:276782494.
[34] N. Hazrati, F. Ricci, Recommender systems effect on the evolution of users' choices distribution, Information Processing & Management 59 (2022) 102766. doi:https://doi.org/10.1016/j.ipm.2021.102766.
[35] N. Hazrati, F. Ricci, Choice models and recommender systems effects on users' choices, User Modeling and User-Adapted Interaction 34 (2023) 109–145. doi:10.1007/s11257-023-09366-x.
[36] R. Ungruh, A. Bellogín, M. S. Pera, From monolith to mosaic: Uncovering behavioral differences for choice models in recommender systems simulations, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 2717–2722. URL: https://doi.org/10.1145/3726302.3730199. doi:10.1145/3726302.3730199.
[37] N.-J. Akpinar, C. DiCiccio, P. Nandy, K. Basu, Long-term dynamics of fairness intervention in connection recommender systems, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22, 2022, p. 22–35. doi:10.1145/3514094.3534173.
[38] J. Vandeputte, A. Cornuéjols, N. Darcel, F. Delaere, C. Martin, Coaching Agent: Making Recommendations for Behavior Change. A Case Study on Improving Eating Habits, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS '22, Richland, SC, 2022, p. 1292–1300.
[39] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, A. Karatzoglou, RecoGym: A reinforcement learning environment for the problem of product recommendation in online advertising, arXiv preprint arXiv:1808.00720 (2018).
[40] B. Shi, M. G. Ozsoy, N. Hurley, B. Smyth, E. Z. Tragos, J. Geraci, A. Lawlor, PyRecGym: A reinforcement learning gym for recommender systems, in: Proceedings RecSys '19, 2019, pp. 491–495. doi:10.1145/3298689.3346981.
[41] M. Mladenov, C. Hsu, E. I. Vihan Jain, C. Colby, N. Mayoraz, H. Pham, D. Tran, I. Vendrov, C. Boutilier, RecSim NG: Toward Principled Uncertainty Modeling for Recommender Ecosystems, arXiv preprint arXiv:2103.08057 (2021).
[42] A. Zhang, Y. Chen, L. Sheng, X. Wang, T.-S. Chua, On generative agents in recommendation, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, 2024, p. 1807–1817. doi:10.1145/3626772.3657844.
[43] L. Chen, Q. Dai, Z. Zhang, X. Feng, M. Zhang, P. Tang, X. Chen, Y. Zhu, Z. Dong, Recusersim: A realistic and diverse user simulator for evaluating conversational recommender systems, in: Companion Proceedings of the ACM on Web Conference 2025, WWW '25, 2025, p. 133–142. doi:10.1145/3701716.3715258.
[44] Z. Zhang, S. Liu, Z. Liu, R. Zhong, Q. Cai, X. Zhao, C. Zhang, Q. Liu, P. Jiang, Llm-powered user simulator for recommender system, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 13339–13347. doi:10.1609/aaai.v39i12.33456.
[45] A. Chuklin, P. Serdyukov, M. de Rijke, Modeling clicks beyond the first result page, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 1217–1220. URL: https://doi.org/10.1145/2505515.2507859. doi:10.1145/2505515.2507859.

[46] A. Cockburn, B. McKenzie, What do web users do? an empirical analysis of web use, International Journal of Human-Computer Studies 54 (2001) 903–922. URL: https://www.sciencedirect.com/science/article/pii/S1071581901904598. doi:https://doi.org/10.1006/ijhc.2001.0459.

[47] V. W. Anelli, A. Bellogín, T. Di Noia, D. Jannach, C. Pomo, Top-n recommendation algorithms: A quest for the state-of-the-art, in: Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22, 2022, p. 121–131. doi:10.1145/3503252.3531292.

[48] M. Ferrari Dacrema, S. Boglio, P. Cremonesi, D. Jannach, A troubling analysis of reproducibility and progress in recommender systems research, ACM Trans. Inf. Syst. 39 (2021). URL: https://doi.org/10.1145/3434185. doi:10.1145/3434185.

[49] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, Grouplens: an open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94, Association for Computing Machinery, New York, NY, USA, 1994, p. 175–186. URL: https://doi.org/10.1145/192844.192905. doi:10.1145/192844.192905.

[50] B. Paudel, F. Christoffel, C. Newell, A. Bernstein, Updatable, accurate, diverse, and scalable recommendations for interactive applications, ACM Transactions on Interactive Intelligent Systems 7 (2016) 1–34. doi:10.1145/2955101.

[51] H. Steck, Embarrassingly shallow autoencoders for sparse data, in: The World Wide Web Conference, WWW '19, 2019, p. 3251–3257. doi:10.1145/3308558.3313710.

[52] X. Ning, G. Karypis, Slim: Sparse linear methods for top-n recommender systems, in: 2011 IEEE 11th International Conference on Data Mining, 2011, pp. 497–506. doi:10.1109/ICDM.2011.134.

[53] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.

[54] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 263–272. doi:10.1109/ICDM.2008.22.

[55] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 173–182. URL: https://doi.org/10.1145/3038912.3052569. doi:10.1145/3038912.3052569.

[56] D. Liang, R. G. Krishnan, M. D. Hoffman, T. Jebara, Variational autoencoders for collaborative filtering, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 689–698. URL: https://doi.org/10.1145/3178876.3186150. doi:10.1145/3178876.3186150.

[57] G. Alves, D. Jannach, R. Ferrari De Souza, M. G. Manzato, User perception of fairness-calibrated recommendations, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 78–88. URL: https://doi.org/10.1145/3627043.3659558. doi:10.1145/3627043.3659558.

[58] V. W. Anelli, A. Bellogin, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. Di Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, 2021, p. 2405–2414. doi:10.1145/3404835.3463245.

[59] C. Gini, Measurement of inequality of incomes, The Economic Journal 31 (1921) 124–126. URL: http://www.jstor.org/stable/2223319.

[60] H. Yin, B. Cui, J. Li, J. Yao, C. Chen, Challenging the long tail recommendation, Proc. VLDB Endow. 5 (2012) 896–907. URL: https://doi.org/10.14778/2311906.2311916. doi:10.14778/2311906.2311916.

[61] M. Ferrari Dacrema, S. Boglio, P. Cremonesi, D. Jannach, A troubling analysis of reproducibility and progress in recommender systems research, ACM Trans. Inf. Syst. 39 (2021). URL: https://doi.org/10.1145/3434185. doi:10.1145/3434185.

[62] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, C. Geng, Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison, in: Pro-

ceedings of the 14th ACM Conference on Recommender Systems, RecSys '20, 2020, p. 23–32. doi:10.1145/3383313.3412489.

[63] F. Abel, Y. Deldjoo, M. Elahi, D. Kohlsdorf, Recsys challenge 2017: Offline and online evaluation, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, 2017, p. 372–373. doi:10.1145/3109859.3109954.

[64] D. M. Fleder, K. Hosanagar, Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity, Management Science 55 (2009) 697–712. doi:10.1287/mnsc.1080.0974.

[65] A. Olteanu, C. Castillo, F. Diaz, E. Kıcıman, Social data: Biases, methodological pitfalls, and ethical boundaries, Frontiers in Big Data Volume 2 (2019). URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2019.00013. doi:10.3389/fdata.2019.00013.

[66] M. Schedl, N. Rekabsaz, E. Lex, T. Grosz, E. Greif, Multiperspective and multidisciplinary treatment of fairness in recommender systems research, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22 Adjunct, Association for Computing Machinery, New York, NY, USA, 2022, p. 90–94. URL: https://doi.org/10.1145/3511047.3536400. doi:10.1145/3511047.3536400.

[67] C. Bauer, A. Said, E. Zangerle, Evaluation Perspectives of Recommender Systems: Driving Research and Education (Dagstuhl Seminar 24211), Dagstuhl Reports 14 (2024) 58–172. doi:10.4230/DagRep.14.5.58.