Strategies for Mitigating Artist Gender Bias in Music Recommendation: A Simulation Study

Christine Bauer
christine.bauer@plus.ac.at
Paris Lodron University Salzburg
Salzburg, Austria

Andres Ferraro
andres.ferraro@siriusxm.com
Pandora-Sirius XM
Oakland, CA, USA

ABSTRACT

As recommender systems are prone to various biases, bias mitigation approaches are needed to counteract those. In the music sector, gender imbalance is a particular topical subject. Earlier work has shown that the gender imbalance in the sector translates to the output of music recommender systems. Several works emphasize that items representing women should be given more exposure in music recommendations. This work presents an exploratory analysis of several bias mitigation strategies. Using a simulation approach, we explore the effects of different pre- and post-processing strategies for bias mitigation. We provide an in-depth analysis using state-of-the-art performance measures concerning gender fairness. The results indicate that the different strategies can help to mitigate gender bias in the long term in particular ways: Some strategies render improvement in the exposure of women in the top ranks; other approaches help recommend more variety of items representing women.

KEYWORDS

Recommender systems, artists, music, gender balance, fairness, bias

1 INTRODUCTION AND RELATED WORK

The impact of algorithmic decision-making on people's daily life is growing [10]. Thereby, predictions and recommendations made by algorithms have the power to shape an entire ecosystem. For instance, in the music domain, streaming platforms have become one of the main sources of music consumption [19]. Typically, such platforms integrate music recommender systems (MRS) that learn from large-scale user behavior and music features [25] to recommend music items tailored to a specific user [7]. The digital music value chain embraces a wide set of stakeholders, who have different goals and interests regarding the music being recommended [2, 5]. As a result, what an MRS recommends highly impacts users' listening experience [22] and considerably impacts artists regarding, for instance, exposure and resulting royalty payments [14].

While users frequently perceive algorithmic decisions as objective [16], many factors make such systems prone to biases, often resulting in unfair outcomes [7]. In the music domain, research has addressed a wide scale of biases: for instance, popularity bias (e.g., [3, 21]), cultural biases (e.g., [1]), age biases (e.g., [8, 23]), and gender biases (e.g., [8, 9, 13]). For an overview of research addressing biases related to fairness in MRS, see Dinnissen and Bauer [5].

In this work, we specifically address biases related to artist gender. Gender imbalance is a highly topical subject in the music sector (e.g., [17, 27, 28]). From interviews with artists, Ferraro et al. [14] learn that artists care about the gender imbalance in the music industry, a finding partially supported by interviews in Dinnissen and Bauer [6], too. As artists consider MRS a potential solution to promote content by women to reach a gender balance in what users consume [14], Ferraro et al. [13] analyze MRS approaches regarding gender bias and propose bias mitigation strategies to counteract the gender imbalance. In a simulation study, they demonstrate breaking bias amplification in the long term through gradually increasing exposure of minority genders.

In this work, we build on the findings of Ferraro et al. [13]. Taking a similar simulation approach, this work aims to explore the effects of different pre- and post-processing strategies for bias mitigation. While our work focuses specifically on gender imbalance in the music industry, our contribution is relevant beyond the gender attribute and the music domain: minority attributes and bias mitigation are highly relevant in a wide scale of domains.

This paper is structured as follows: Next, we present the adopted methods, including the proposed mitigation approaches, the used datasets, and the employed metrics (Section 2). After presenting the results (Section 3), we discuss the findings in Section 4. We conclude this work with an outlook on future work (Section 5).

2 METHODS

We run a simulation on a subset [11] of the *Last.fm 360K* dataset [3]. We use an Alternating Least Square (*ALS*) algorithm [18] to generate recommendations with different pre-processing and post-processing variations as bias mitigation strategies.

First, we describe the proposed mitigation strategies. After that, we present the dataset and the employed metrics, and outline the simulation approach.

2.1 Approaches

We chose *ALS* as the basis for our analysis because it is a well-known algorithm for collaborative filtering in the music domain. Each mitigation strategy (S) builds on *ALS*, where we use *pre*-processing and *post*-processing techniques. All these techniques are based on the assumption that changing the order in which the recommendations are presented to the users—increasing the exposure that items representing women receive—would positively affect reaching gender balance.

Two strategies post-process the output of the original *ALS* algorithm:

- $S_{post}^{MoveFirst}$: Move the first recommendation of a woman to the first position.
- $S_{post}^{Balanced}$: From the recommended list generated by *ALS* take interleaved items of each group of genders.

For another strategy, we pre-process the data used to train *ALS*. We split the artist items representing men into two groups (S_{pre}^{Div1}

and S_{pre}^{Div2}) and train independent *ALS* models for each group together with *all* items representing women. This way, the proportion of items representing women and men is similar in the data used to train each model. We then combine ($S_{pre}^{Combined}$) the results generated from each of the two models model (S_{pre}^{Div1} and S_{pre}^{Div2}) for each user: we do a weighted average of the scores of each model; as a result, items representing women—that are in both lists—have higher chances of getting in the final list.

ALS without any adaptations serves as the baseline (S^{ALS}).

2.2 Dataset

For our exploratory analysis, we rely on a dataset provided by Ferraro et al. [11]. It is a subset of the *Last.fm 360K* dataset [3] enriched with artist gender information collected from MusicBrainz.org (*MB*)[1] as used in Ferraro et al. [13]. This dataset contains 'solo' artists—thus, where the artist is an individual person—for which *MB* reports the gender (in *MB*: female or male). We are well aware that this binary gender classification does not reflect the multitude of gender identities [26]. However, to the best of our knowledge, there is no dataset that goes beyond this binary gender classification.[2] The dataset contains 46,469 artists, of which 10,535 are women and 35,922 are men [11]. Following the procedure of Ferraro et al. [13], we consider only users and artists with more than 30 interactions to have sufficient data for training and evaluation. Thus, we remain with 220,444 users and 12,900 artists, of whom about a third represent women. We randomly select for each user 80% of the items for training and 20% for test.

2.3 Metrics

We use several metrics to understand the system's behavior from different perspectives:

- *Accuracy.* We use *Precision* (*P@k*), *normalized discounted cumulative gain* (*nDCG@k*), and *Mean Average Precision* (*MAP@k*) to measure the accuracy of the algorithms.
- *Diversity.* With the overall *Coverage* (*Coverage@k*), we measure the number of different artists globally recommended; overall and differentiated by gender. In addition, we use the *Gini index* (*Gini@k*), which measures how concentrated the recommendations are on a few artists. A *Gini index* of 1 indicates that all recommendations are identical for all users, whereas a value of 0 means they are all different.
- *Exposure.* We particularly focus on the position in the recommendation rankings because users interact more frequently with only the top-ranked items (i.e., position bias) [4]. To this end, we average for each user the *position of the first occurrence of content by a woman* (with the highest rank on position 0) in the recommendation ranking.
- *Representation.* We use the *percentage of items representing women* in the recommendations to measure representation.

2.4 Simulation

We use a simulation to mimic feedback loops to study the long-term effect of the proposed mitigation strategies. We follow the procedure used in previous works [12, 13, 20, 30]: For each user, we take the system's recommendations and increase the counter in the original user–artist matrix, simulating that the users listened to all items recommended by the system in the top 10. We then retrain the model and compute recommendations for the next iteration. We repeat this procedure for a total of 20 iterations.

3 RESULTS

3.1 Accuracy metrics

As concerns the accuracy metrics that consider ranking (*nDCG* and *MAP*) in Fig. 1 and Fig. 2 respectively, we see that the various mitigation strategies have a significant negative impact on performance. $S_{post}^{MoveFirst}$ is the one where performance is affected least (reducing *nDCG*@10 by 13%). However, in terms of *P*@10 (Fig. 3), there is only a marginal performance loss for $S_{post}^{MoveFirst}$ and S^{ALS}.

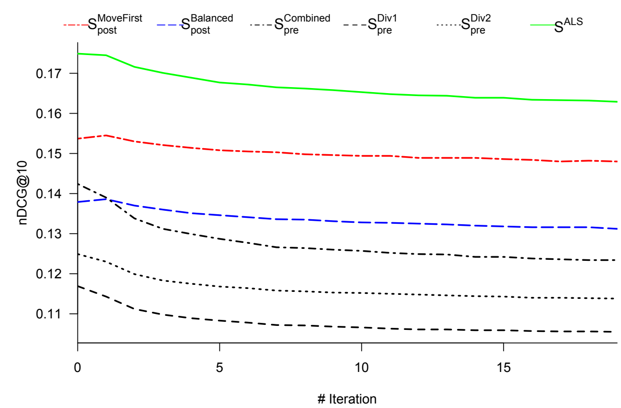

Figure 1: ***nDCG*@10 for all artists**

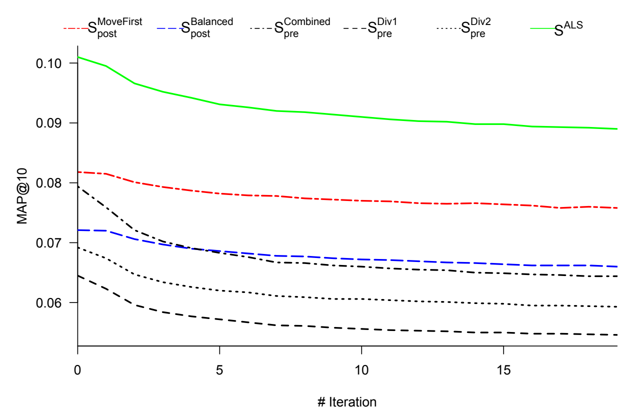

Figure 2: ***MAP*@10 for all artists**

3.2 Diversity

As concerns the overall *Coverage*@10 of recommendations for all artists (Fig. 4), the performance is roughly similar across all models ($S_{post}^{MoveFirst}$, $S_{post}^{Balanced}$, $S_{pre}^{Combined}$, and S^{ALS}). From Fig. 5, we see that *Coverage* of items representing women in the top 100

[1] http://musicbrainz.org
[2] For an overview of current practices concerning the use of gender in research on information access, see Pinney et al. [24].

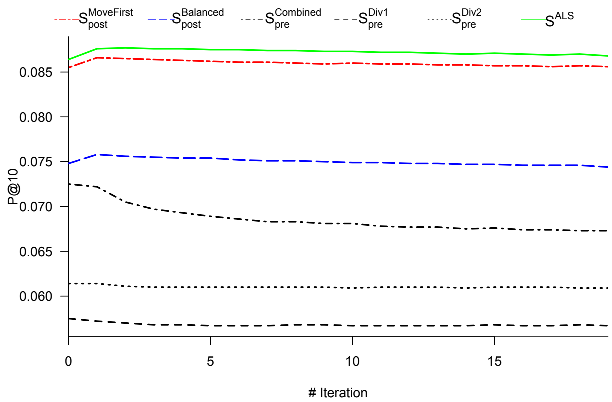

Figure 3: $P@10$ **for all artists**

($Coverage@100_{women}$) is far higher for the models offering a higher proportion of items representing women ($S_{post}^{Balanced}$, $S_{pre}^{Combined}$, and the pre-models S_{pre}^{Div1} and S_{pre}^{Div2}) compared to the baseline S^{ALS} and $S_{post}^{MoveFirst}$.

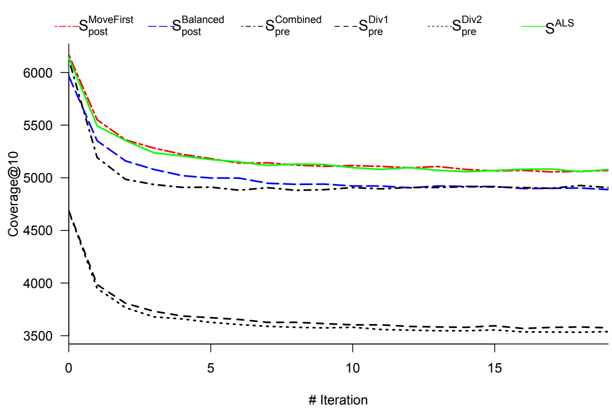

Figure 4: $Coverage@10$ **for all users and artists**

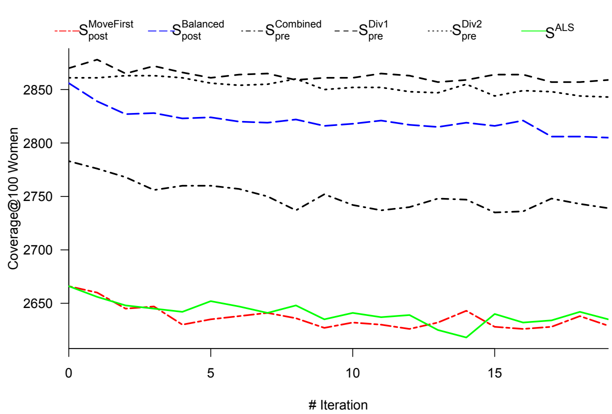

Figure 5: $Coverage@100_{women}$ **for items representing women**

Fig. 6 shows that the models $S_{post}^{Balanced}$, S_{pre}^{Div1}, and S_{pre}^{Div2} improve the distribution of different artists recommended by a large margin compared to S^{ALS}.

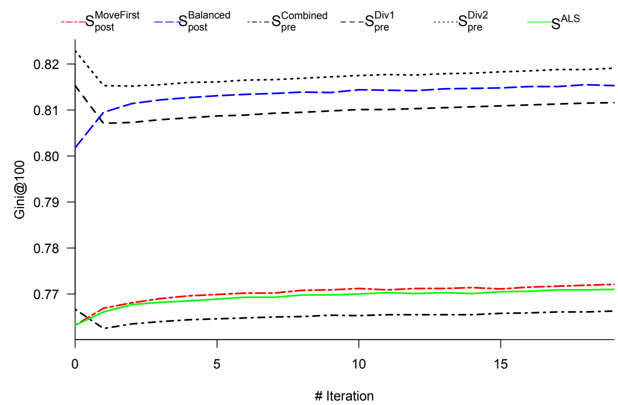

Figure 6: $Gini@100$ **for all artists**

3.3 Exposure

As concerns the average first position of items representing women (Fig. 7), $S_{post}^{MoveFirst}$ and $S_{post}^{Balanced}$ seem effective as these reach an average position of 0 for women. Comparing with the impact in the average first position of items representing men in Fig. 8, we see that $S_{post}^{MoveFirst}$ and $S_{pre}^{Combined}$ are the ones with more impact reaching 1.6 compared to S^{ALS} that gives the lowest value (0.8).

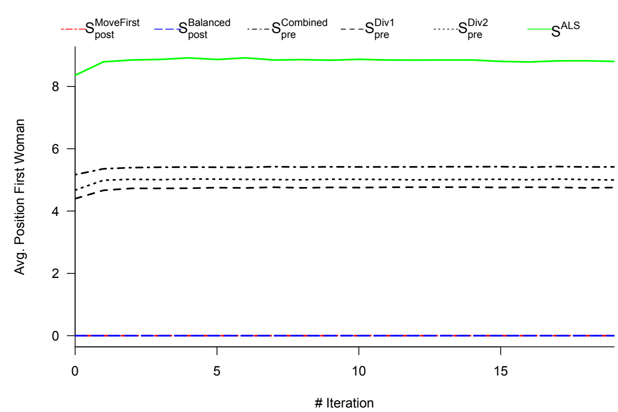

Figure 7: Average position of first recommended item representing a woman. Note, both $S_{post}^{Balanced}$ and $S_{post}^{MoveFirst}$ equally render 0 for all iterations.

3.4 Representation

Fig. 9 shows the percentages of items representing women compared to men. As expected, $S_{post}^{Balanced}$ always recommends 50% items representing women. $S_{post}^{MoveFirst}$ provides a slightly higher percentage of items representing women than S^{ALS}.

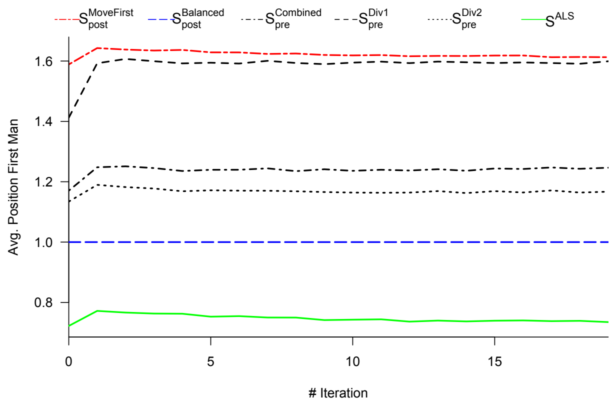

Figure 8: Average position of first recommended item representing a man

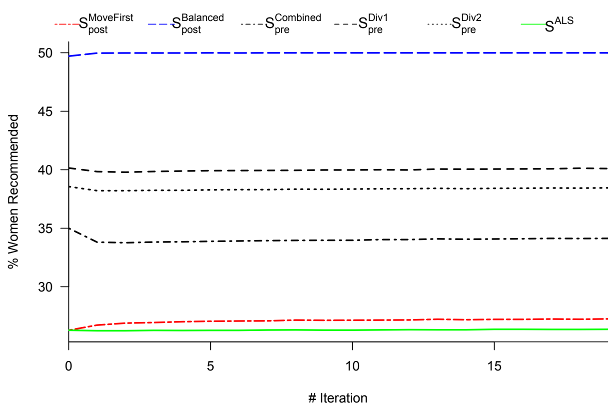

Figure 9: Proportion of items representing women in recommendations

4 DISCUSSION

Concerning accuracy, we observe that performance decreases for all models in terms of MAP@10 and $nDCG$@10. In terms of P@10, the performance remains stable for most models except for the performance loss of $S_{pre}^{Combined}$. The S^{ALS}—the algorithm without mitigation intervention—achieves better accuracy scores than the other models. Yet, $S_{post}^{MoveFirst}$ is not far below S^{ALS}; in terms of P@10, the performance of $S_{post}^{MoveFirst}$ and $S^{ALS}$is comparable. While S^{ALS} starts well in terms of MAP@10 and $nDCG$@10, the performance drops more compared to $S_{post}^{MoveFirst}$ and $S_{post}^{Balanced}$. $S_{pre}^{Combined}$ does not keep up with the other models in terms of accuracy performance in the long term; $S_{pre}^{Combined}$ starts better than $S_{post}^{Balanced}$ for MAP@10 and $nDCG$@10 but loses quickly.

$Coverage$@10 in general drops quickly in the first three iterations for all models; thereafter, all seem relatively stable. $S_{post}^{MoveFirst}$ reaches comparable $Coverage$@10 as S^{ALS} throughout all iterations. $S_{post}^{Balanced}$ and $S_{pre}^{Combined}$ are slightly lower in terms of $Coverage$@10, where $S_{pre}^{Combined}$ has a stronger drop in the first two iterations but then keeps up with $S_{post}^{Balanced}$ in the long term. As expected, the pre-models S_{pre}^{Div1} and S_{pre}^{Div2} reach low coverage only, as both are trained on smaller datasets. Compared to the other metrics, $Coverage$@100_{women} does not evolve smoothly along the iterations but shows up and downs for all models. While $Coverage$@100_{women} is similarly low for $S_{post}^{MoveFirst}$ and S^{ALS}, $S_{pre}^{Combined}$ achieves far higher values and $S_{post}^{Balanced}$ even more so. As expected, the pre-models S_{pre}^{Div1} and S_{pre}^{Div2} have high values for $Coverage$@100_{women} because both pre-models are trained on datasets with a roughly balanced proportion of items representing men and women. Comparing the overall $Coverage$@10 with $Coverage$@100_{women} across all models, we observe that the coverage of women remains relatively stable while the general coverage drops; hence, a larger proportion in coverage refers to women. $Gini$@100 increases steadily for all models, yet slightly only. Still, there is a discrepancy between $S_{post}^{Balanced}$ and the pre-models S_{pre}^{Div1} and S_{pre}^{Div2}, which all score high on $Gini$@100, and the other models (S^{ALS}, $S_{post}^{MoveFirst}$, and $S_{pre}^{Combined}$) scoring comparably lower. Again, $S_{post}^{MoveFirst}$ scores similar to S^{ALS} throughout all iterations.

Regarding exposure, all models are stable throughout the iterations but at different levels. All proposed mitigation strategies show improvements over S^{ALS} regarding the average position of the first recommended item representing a woman. $S_{post}^{MoveFirst}$ and $S_{post}^{Balanced}$ are radical in that regard, as these strategies put one woman always first (position 0). $S_{post}^{Balanced}$, in turn, puts the first man on position 1; with $S_{post}^{Balanced}$, the first man appears on average on position 1.6. In general, for all models, we observe a feedback loop. None of the proposed strategies can break the loop in this regard. A similar picture is observed for representation, which is stable throughout the iterations for all models—but at different levels, with S^{ALS} delivering the lowest representation of women. $S_{post}^{MoveFirst}$ does only render slightly higher scores.

5 CONCLUSION

With the goal of reaching gender balance in users' consumption of music items, we investigated different bias mitigation strategies that either pre-process the input or post-process the output of the recommendations. These strategies include affecting only the position of a single element (as in $S_{post}^{MoveFirst}$, where the highest-ranked item representing a woman is put on rank 0) or a more substantial re-organization in the ranking of the recommendations.

The results indicate that the different strategies can help to mitigate gender bias in the long term in particular ways. From our (preliminary) results, we see the effectiveness of the $S_{post}^{MoveFirst}$ strategy in terms of ranking (i.e., the average position of the first recommended item representing a woman) with only a small impact on precision. We also see the strong advantage of $S_{post}^{Balanced}$ and $S_{pre}^{Combined}$ of recommending more variety of items representing women, with an impact on accuracy, though.

Our work comes with some limitations. For exploratory purposes, we used a small dataset. Further, we used a random test-training split, where a temporal split might render different results [29]. We used simple choice models in our simulation, whereas people typically follow complex choice models [15].

Building on our (preliminary) results, future work will use bigger and more recent datasets and expand to a larger variety of pre- and post-processing methods. Further, it is an essential avenue to study the impact on intersectional groups of artists.

REFERENCES

[1] Christine Bauer and Markus Schedl. 2019. Global and Country-Specific Mainstreaminess Measures: Definitions, Analysis, and Usage for Improving Personalized Music Recommendation Systems. *PLOS ONE* 14, 6 (2019), 36 pages. https://doi.org/10.1371/journal.pone.0217389

[2] Christine Bauer and Eva Zangerle. 2019. Leveraging Multi-method Evaluation for Multi-stakeholder Settings. In *Proceedings of the 1st Workshop on the Impact of Recommender Systems, co-located with 13th ACM Conference on Recommender Systems (ACM RecSys 2019)* (Copenhagen, Denmark, 19 September) *(CEUR Workshop Proceedings, Vol. 2462)*, Oren Sar Shalom, Dietmar Jannach, and Ido Guy (Eds.). CEUR-WS.org, Aachen, Germany, Article 3, 3 pages. http://ceur-ws.org/Vol-2462/short3.pdf

[3] Òscar Celma. 2010. *Music recommendation and discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space.* Springer, Berlin, Heidelberg, Germany. https://doi.org/10.1007/978-3-642-13287-2

[4] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. Position Bias in Recommender Systems for Digital Libraries. In *Transforming Digital Worlds. iConference 2018*, Gobinda Chowdhury, Julie McLeod, Val Gillet, and Peter Willett (Eds.). Springer, Cham, Germany, 335–344. https://doi.org/10.1007/978-3-319-78105-1_37

[5] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in Music Recommender Systems: a Stakeholder-Centered Mini Review. *Frontiers in Big Data* 5, Article 913608 (2022), 9 pages. https://doi.org/10.3389/fdata.2022.913608

[6] Karlijn Dinnissen and Christine Bauer. 2023. Amplifying Artists' Voices: Item Provider Perspectives on Influence and Fairness of Music Streaming Platforms. In *Proceedings of the 31st Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3565472.3592960

[7] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Foundations and Trends® in Information Retrieval* 16, 1–2 (2022), 177 pages. https://doi.org/10.1561/1500000079

[8] 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, FAT* 2018 (FAT* '18, Vol. 81)*. Proceedings of Machine Learning Research, New York, NY, USA, 172–186. http://proceedings.mlr.press/v81/ekstrand18b/ekstrand18b.pdf

[9] Avriel Epps-Darling, Romain Takeo Bouyer, and Henriette Cramer. 2020. Artist Gender Representation in Music Streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference* (Montréal, Canada, October 11–16) *(ISMIR '20)*. ISMIR, 248–254. https://doi.org/10.5281/zenodo.4245416

[10] Panagiota Fatourou, Chris Hankin, and Bran Knowles. 2021. *Gender Bias in Automated Decision Making Systems.* Technical Report. Policy paper, endorsed by the ACM Europe Technology Policy Committee. https://www.acm.org/binaries/content/assets/public-policy/aigenderbiaspaper.pdf

[11] Andres Ferraro, Christine Bauer, and Xavier Serra. 2020. Last.fm Artists Gender Information. https://doi.org/10.5281/zenodo.3748787

[12] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jason Yoon. 2019. Artist And Style Exposure Bias in Collaborative Filtering Based Music Recommendations. In *Proceedings of the 1st Workshop on Designing Human-Centric Music Information Research Systems* (Delft, The Netherlands) *(wsHCMIR '19)*. 8–10. arXiv:1911.04827

[13] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In *6th ACM SIGIR Conference on Human Information Interaction and Retrieval* (Canberra, ACT, Australia, 14-19 March) *(CHIIR '21)*. ACM, New York, NY, USA, 249–254. https://doi.org/10.1145/3406522.3446033

[14] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. What Is Fair? Exploring the Artists' Perspective on the Fairness of Music Streaming Platforms. In *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference* (Bari, Italy) *(INTERACT '21, Vol. 12933)*. Springer, Cham, Germany, 562–584. https://doi.org/10.1007/978-3-030-85616-8_33

[15] Naieme Hazrati and Francesco Ricci. 2023. Choice Models and Recommender Systems Effects On Users' Choices. *User Modeling and User-Adapted Interaction* (2023), 37 pages. https://doi.org/10.1007/s11257-023-09366-x

[16] Natali Helberger, Theo Araujo, and Claes H. de Vreese. 2020. Who is The Fairest of Them All? PUblic Attitudes and Expectations Regarding Automated Decision-Making. *Computer Law & Security Review* 39, Article 105456 (2020), 16 pages. https://doi.org/10.1016/j.clsr.2020.105456

[17] Karla Hernandez, Stacy L. Smith, Marc Choueiti, and Katherine Pieper. 2022. *Inclusion in the Recording Studio?: Gender and Race/Ethnicity of Artists, Songwriters & Producers across 1,000 Popular Songs from 2012–2021.* Technical Report. Annenberg Inclusion Initiative. https://assets.uscannenberg.org/docs/aii-inclusion-recording-studio-20220331.pdf

[18] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining* (Pisa, Italy) *(ICDM 2008)*. IEEE, New York, NY, USA, 263–272. https://doi.org/10.1109/ICDM.2008.22

[19] IFPI. 2020. *Global Music Report 2020: The Industry in 2019.* Technical Report. IFPI, London, UK. https://www.ifpi.org/wp-content/uploads/2020/07/Global_Music_Report-the_Industry_in_2019-en.pdf

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

[21] Dominik Kowald, Peter Müllner, Eva Zangerle, Christine Bauer, Markus Schedl, and Elisabeth Lex. 2021. Support the Underground: Characteristics of Beyond-Mainstream Music Listeners. *EPJ Data Science* 10, 1, Article 14 (2021), 28 pages. https://doi.org/10.1140/epjds/s13688-021-00268-9

[22] Jin Ha Lee, Liz Pritchard, and Chris Hubbles. 2019. Can We Listen To It Together? Factors Influencing Reception of Music Recommendations and Post-Recommendation Behavior. In *Proceedings of the 20th International Society for Music Information Retrieval Conference* (Delft, The Netherlands) *(ISMIR 2019)*. ISMIR, 663–669. https://doi.org/10.5281/zenodo.3527896

[23] Nicola Neophytou, Bhaskar Mitra, and Catherine Stinson. 2022. Revisiting Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Advances in Information Retrieval* (Stavanger, Norway) *(ECIR '22)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, Germany, 641–654. https://doi.org/10.1007/978-3-030-99736-6_43

[24] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 269–279. https://doi.org/10.1145/3576840.3578316

[25] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. Music Recommendation Systems: Techniques, Use Cases, and Challenges. In *Recommender Systems Handbook* (3rd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, USA, 927–971. https://doi.org/10.1007/978-1-0716-2197-4_24

[26] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26, 4 (June 2019), 62–65. https://doi.org/10.1145/3338283

[27] Yixue Wang and Emőke-Ágnes Horvát. 2019. Gender Differences in the Global Music Industry: Evidence from MusicBrainz and The Echo Nest. In *Proceedings of the 13th International AAAI Conference on Web and Social Media* (Munich, Germany) *(ICWSM '19)*. AAAI Press, Palo Alto, CA, USA, 517–526. https://doi.org/10.1609/icwsm.v13i01.3249

[28] Jada Watson. 2020. Programming Inequality: Gender Representation on Canadian Country Radio (2005–2019). In *Proceedings of the 21st International Society for Music Information Retrieval Conference* (Montréal, Canada) *(ISMIR 2020)*. ISMIR, 392–399. https://doi.org/10.5281/zenodo.4245452

[29] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* 55, 8, Article 170 (2022), 38 pages. https://doi.org/10.1145/3556536

[30] Jingjng Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems via an Agent-Based Simulation Framework. *Information Systems Research* 31, 1 (2020), 76–101. https://doi.org/10.1287/isre.2019.0876