# Break the Loop: Gender Imbalance in Music Recommenders

Andres Ferraro
andres.ferraro@upf.edu
Universitat Pompeu Fabra
Barcelona, Spain

Xavier Serra
xavier.serra@upf.edu
Universitat Pompeu Fabra
Barcelona, Spain

Christine Bauer
c.bauer@uu.nl
Utrecht University
Utrecht, The Netherlands

## ABSTRACT

As recommender systems play an important role in everyday life, there is an increasing pressure that such systems are fair. Besides serving diverse groups of users, recommenders need to represent and serve item providers fairly as well. In interviews with music artists, we identified that gender fairness is one of the artists’ main concerns. They emphasized that female artists should be given more exposure in music recommendations. We analyze a widely-used collaborative filtering approach with two public datasets—enriched with gender information—to understand how this approach performs with respect to the artists’ gender. To achieve gender balance, we propose a progressive re-ranking method that is based on the insights from the interviews. For the evaluation, we rely on a simulation of feedback loops and provide an in-depth analysis using state-of-the-art performance measures and metrics concerning gender fairness.

## CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing**;

## KEYWORDS

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

**ACM Reference Format:**
Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR ’21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3406522.3446033

## 1 INTRODUCTION

Recommender systems influence the products we buy [36], the news we read [27], or how we interact with others [41]. The goal of a recommender system is to predict which items a user might like given the user’s previous ratings or interactions with items. This may lead to situations where users only see a narrow subset of the entire range of available recommendations [49]. Users will respond to those recommendations, which will then be used as input for future recommendations; with this feedback loop, the recommender system will learn to recommend increasingly similar items [11, 14, 22, 49]. However, this strategy may have negative effects on other humans involved in and affected by recommender systems: the item providers [4]. When calling for fair recommender systems [16], we need to consider all humans that are affected.

In this work, we zoom in the music domain. Music recommender systems are important drivers in the music industry, and for music platforms in particular [5]. These systems may privilege the content of a small group of artists when maximizing user satisfaction. As a consequence, this limits some artists’ chances to reach a larger audience particularly due to the feedback loop.

We reached out to artists to understand their perspective on *fair recommendations*. From the interviews, we understand that one of the major problems artists see in the music business is the gender imbalance. Gender bias in the music domain is subject of academic investigation (e.g., [43, 46, 51, 52]) and concerns about bias and discrimination are also repeatedly voiced in the media (e.g., [15, 34, 53]). One novel contribution from the interviews is that the artists believe that recommender systems could mitigate this issue.

We address this problem in two steps: First, we analyze a widely-used collaborative filtering approach concerning the artists’ gender. Second, based on the insights from the interviews and the first step of the analysis, we propose a progressive re-ranking method to achieve gender balance. For the evaluation, we rely on a simulation of feedback loops to provide an in-depth analysis of the longitudinal effects considering state-of-the-art performance measures, and metrics concerning gender fairness.

## 2 RELATED WORK

There is a wealth of research on what is fair or unfair (for an overview, see Hutchinson and Mitchell [25]). Recently, fairness received particular attention in the field of artificial intelligence [30].

In recommender systems research, unwanted bias has long been a topic of interest [26]. For example, systems being optimized for accuracy may privilege popular items [8, 38, 48], which are not necessarily more satisfying for the users [29, 32]. While fairness for various consumer groups increasingly gains attention in recommender systems research (e.g., [17]), studies on fairness considering other stakeholders are scarce (e.g., [1]). Smith et al. [45] raise the fundamental question, “what is fair in the context of recommendation—particularly when there are multiple stakeholders?” and conduct interviews with users to understand their ideas about fair treatment in recommendations, deriving common topics from the answers.

Gender bias has been of particular interest in research. Studies have shown that the design of software [6, 7, 9, 50] or websites [33] may introduce bias for users of different genders.Another thread of research investigates gender bias in algorithmic decision-making. For instance, Keyes [28] criticizes how gender is defined in extant works on automatic gender recognition because the current (implicit) understanding of gender and how it is implemented in such systems negatively affects transgender people.

*CHIIR ’21, March 14–19, 2021, Canberra, Australia.*
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8055-3/21/03.
https://doi.org/10.1145/3406522.3446033

In recommender system research, studies have shown that algorithms perform differently for different demographic user groups [17] and differ in the (book authors') gender distribution in the computed recommendation lists [18]. Pierson [39] found that women are likely to oppose the inclusion of gender as a feature in course recommendation algorithms because such algorithms are less likely to recommend science courses to female students.

In the music business, many studies show discrimination and bias related to gender. For example, female artists are less likely to reach an audience only for being female [43], they are underrepresented in charts and awards nominations [46], and less radio air time is dedicated to females [52]. This bias is also present in streaming services [3, 19]. For example, the largest proportion of female and mixed-gender artists appear in the lower levels of popularity [19].

Recent works studied the role of music recommender systems on gender bias. For example, Oliveira et al. [37] investigate gender diversity in music recommendations, where they consider the type of artist (i.e., 'band', 'orchestra', 'solo', etc.) as a gender. Epps-Darling et al. [19] compare organically generated listening events with the ones that are algorithmically induced. Shakespeare et al. [44] compare the recommendations of multiple collaborative filtering approaches in terms of gender distribution and conclude that these methods propagate the gender bias that is present in the dataset.

Our work distinguishes from extant work in these aspects: (i) We incorporate the opinion of those people concerned (i.e., artists) in finding a solution for gender bias on music platforms. (ii) We propose a way to gradually mitigate this bias following the insights from the interviews with artists. (iii) We use a simulation approach considering feedback loops to understand the longitudinal effects.

## 3 INSIGHTS FROM INTERVIEWS

### 3.1 Participants, Material, and Focus

We conducted semi-structured interviews (with 11 guiding questions) with 9 artists that we consider diverse in the kind of music they perform (including folk, pop, punk-rock, dubstep, jazz, flamenco, progressive rock, hip-hop, and reggae), their popularity (2 globally known, 4 known within their country, 3 regionally popular), experience (4 to 25 years active in the music industry; 1 to 10 records released), age (26 to 55 years), and gender (2 female, 7 male). According to research practice [13, 35], the sample size is adequate and we reached a high level of thematic saturation [23] with the same topics being repeatedly mentioned across the interviews. Anonymity of data was guaranteed.

The recordings of the interviews amount to 420 minutes with a total of 33,669 words in the transcriptions. Following the methodology of Qualitative Content Analysis [31], we developed an annotation scheme inductively from raw data and annotated the transcriptions accordingly. In the interviews, the artists addressed a wide variety of topics (e.g., lack of control, context of music, transparency). In this work at hand, we focus on the parts of the interviews related to *gender fairness*. Specifically, we asked whether and how music platforms should intervene and nudge users concerning the music that they consume. Only in follow-up questions, we asked for concrete scenarios, and pointed to genres that are not widely consumed and gender representation.

### 3.2 Interview Results

In general, the interviewed artists had a strong tendency against influencing users concerning music style. One participant argues, *"I don't see why we should tell the users which genres they should listen to."* In sheer contrast to this, there was a clear agreement among all participants—even though the sample is male-dominated—that it is important to promote content of female artists to reach gender fairness. One participant argues, *"I think there should be actions to correct some biases. The question is in which cases it should be corrected and in which not. In heavy metal music, I imagine that there aren't many female singers. Maybe we could give them more visibility, otherwise they would never be seen."* Another artist suggests maintaining gender balance in the recommendations. He states, *"[Platforms have] a huge responsibility in making recommendations."*

A newcomer artist, who argues against purposefully influencing users concerning the genres that they listen to, voices the need to do so concerning gender balance: *"[...] the population of the world is 50% women. So it would be ridiculous if the system wouldn't recommend them."* This artist suggests a progressive change towards gender balance, *"otherwise the users could perceive it as something bad and leave the platform."* Another artist proposes to enforce a 50% gender balance, because many other factors than gender (e.g., the music style) define whether a user will like a recommendation. Another artist argues for 50% of female music, and likewise 20–30% of local artists, and suggests to consider also proportions for other minorities (e.g., ethnicity, sexual orientation). Finally, an artist postulates that every recommendation influences a user in some way or other: *"It is impossible to be impartial; so it is better to do it as equally as possible."*

In summary, the interview results suggest that the artists are concerned about gender fairness. The artists voiced that the recommendations could be used as a means to change the consumers' listening behavior by promoting content of female artists and suggest gender balance in the recommendations (i.e., *positive disparate treatment*). Yet, they emphasize to increase this proportion only gradually until gender balance is reached to avoid reactance.

## 4 QUANTITATIVE APPROACH

We build on the interview results (Section 3) with a two-part quantitative analysis. In the first part, we evaluate a recommendation algorithm that is widely used in the music domain with respect to gender fairness. For this evaluation, we device two large real-word datasets of music listening events (Section 4.1). The goal of this analysis is to understand (i) how the datasets are distributed in terms of the artists' gender and (ii) how the algorithm performs for those distributions with respect to gender fairness.

As collaborative filtering is the prevalent approach for recommendations in the music domain, we choose the Alternating Least Square (*ALS*) algorithm [24] for our analysis. As the number of tracks per artist may vary per gender both in the dataset as well as in the recommendations, we evaluate—where possible—for both, recommendations on the artist level (*Last.fm 360K* dataset and *LFM-1b* dataset) as well as recommendations on the track level (*LFM-1b* dataset). In addition, we compare *ALS* with two baselines, one that generates random recommendations (*RND*) and one that recommends the same most popular items to all the users (*POP*).

We use a simulation to mimic feedback loops to study how the recommendations can affect user behavior in the longer term, following the procedure used in previous works [21, 26, 55]. 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, repeatedly for a total of 20 iterations. We demonstrate that by employing a (simple) re-ranking mechanism, we can break the feedback loop and gradually increase the exposure of female artists. We provide an in-depth analysis, using a set of metrics and compare the re-ranking mechanism to the baseline without re-ranking.

## 4.1 Datasets

We use two public datasets obtained from Last.fm. *LFM-1b* [42] is a large dataset of more than one billion listening events containing playcounts with timestamp by 120K users covering 32M tracks by 3M artists. The second dataset is *LFM-360k* [10], which contains 17M interactions between users and artists (359K users and 260K artists). We extend the datasets with gender information of the artists collected from MusicBrainz.org (*MB*). For complexity reduction, we focus on 'solo' artists—thus, where the artist is an individual person—and consider those artists for which *MB* reports the gender (in *MB*: female or male). While we are aware that this binary gender classification is inapt to reflect the multitude of gender identities [47], to the best of our knowledge, there is no dataset that goes beyond this binary gender classification. For *LFM-1b*, we collected the gender of 64, 745 artists, whereof 15, 055 are classified female and 49, 690 are male. For *LFM-360k*, we collected gender information for 46, 469 artists, whereof 10, 535 are female and 35, 922 are male [20]. Note, the gender imbalance in the datasets reflect the current reality in the music business [19, 54].

We consider only users and tracks or artists, respectively, with more than 30 interactions to have sufficient data for training and evaluation. Thus, we remain with the following data: For *LFM-1b*, we have 112, 291 users and 465, 064 tracks by 33, 325 artists, and for *LFM-360k*, we have 220, 444 users and 12, 900 artists. For both datasets, we split in train and test set by randomly selecting for each user 80% of the items for training and 20% for test.

## 4.2 Metrics

We apply several metrics to understand the system' behavior from different perspectives. We use metrics that assess the probability of female artists being recommended. 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) [12]. To this end, we average for each user the *position of the first occurrence of content by a female* (with the highest rank on position 0) in the recommendation ranking and the *percentage of content by females* in the recommendations. We use *Hellinger distance* (as in [2]) to measure the similarity of the gender distribution in the recommendations compared to the users' original listening behavior. With *Coverage*, we measure the number of different artists (or tracks) globally recommended (differentiated by gender).

**Table 1: Results for artist recommendation (both datasets).**

| | Algo | Avg position | | % females | Hellinger | Precision | | nDCG |
|---|---|---|---|---|---|---|---|---|
| | | 1st female | 1st male | rec. | distance | P@1 | P@10 | @10 |
| LFM-1b | ALS | 6.7717 | 0.6142 | 25.44 | 0.0988 | 0.4505 | 0.2997 | 0.3409 |
| | POP | 0.1325 | 1.7299 | 32.44 | 0.1577 | 0.1033 | 0.0919 | 0.1118 |
| | RND | 3.3015 | 0.3046 | 23.30 | 0.1346 | 0.0010 | 0.0010 | 0.0019 |
| LFM-360k | ALS | 8.3165 | 0.7136 | 26.27 | 0.2102 | 0.1781 | 0.0863 | 0.2804 |
| | POP | 0.9191 | 0.2713 | 29.31 | 0.2670 | 0.0247 | 0.0205 | 0.0978 |
| | RND | 3.3973 | 0.2951 | 22.77 | 0.2597 | 0.0003 | 0.0003 | 0.0025 |

We use precision and *nDCG* [40] to measure the accuracy of the algorithms. We report precision for all recommendations and also separately by gender. Given a track ($t$) and a user ($u$), $hit@K(t, u)$ returns 1 only if $t$ is recommended in the top-$K$ to the user $u$ and is in the test set for that user. We follow these steps: 1) Generate ranked recommendations for user $u$, referred to as $A$; 2) divide items in $A$ by the artists' gender into $F$ for female and $M$ for male; 3) for each user, the precision is computed as: $P@k = \frac{1}{|K|} \sum_{t \in T} hit@K(t, u)$, where the group of items $J$ corresponds to: $A$ when we compute $P@K_{all}$, $F$ when we compute $P@K_{female}$ and $M$ when we compute $P@K_{male}$. Thus, $P@K_{female}$ and $P@K_{male}$ add up to $P@K_{all}$.

# 5 GENDER IN MUSIC RECOMMENDATION

In this section, we report the performance of the algorithms with respect to gender fairness. We analyze the recommendations on the artist level using both datasets—*LFM-1b* and *LFM-360k*—(Section 5.1), and on the track level using the *LFM-1b* dataset (Section 5.2). Finally, we present the results of the simulation of artist recommendations using the *LFM-1b* dataset (Section 5.3). Note that for all analyses we run the experiments three times and see stable results in all cases.[1]

## 5.1 Gender Fairness on the Artist Level

Table 1 summarizes the results of the analysis on the artist level, considering the top-10 artists recommended by the algorithms.

Except for *POP* using *LFM-1b* (0.1325 vs. 1.7299), the average position of the first female artists is lower than the one of the first male artist. Compared to the baselines, *ALS* delivers the largest gender gap, which is even larger using *LFM-360k* compared to *LFM-1b* (*LFM-360k*: 8.3165 vs. 0.7136; *LFM-1b*: 6.7717 vs. 0.6142, for average first position of female vs. male artists, respectively).

Using *ALS* and *LFM-1b*, 25.44% of the recommendations are female artists, which is close to what is reflected in the users' previous listening behavior (25.26%); thus, indicating statistical parity. For both datasets, the Hellinger distance suggests that the recommendations computed via *ALS* are the closest to the gender distribution as reflected in the users' previous listening behavior.

The last three columns of Table 1 show the performance of the analyzed algorithms. For each of the datasets, the parameters of *ALS* were optimized to provide a higher precision. Consequently, we used a 300-dimensional space in *LFM-1b*, and a 200-dimensional space in *LFM-360k*. In both cases, the results clearly suggest that although *POP* gives better results concerning gender fairness, the performance with respect to precision and nDCG are below *ALS*.

An additional analysis for coverage (considering the top-10 recommendations for each user using *LFM-1b*) shows a far lower coverage using *POP* compared to *ALS* (336 vs. 15, 194 unique artists appearing in the top-10). Likely, the low coverage using *POP* is not in the interest of the overall artist population.

[1] https://github.com/andrebola/gender-recs

Table 2: Results of track recommendation (*LFM-1b*).

| Algo | Avg position | | % females | Hellinger | Precision | | nDCG |
|---|---|---|---|---|---|---|---|
| | 1st female | 1st male | rec. | distance | P@1 | P@10 | @100 |
| ALS | 24.9162 | 4.6993 | 28.99 | 0.1374 | 0.4730 | 0.3237 | 0.2392 |
| POP | 0.8726 | 0.8239 | 66.66 | 0.3404 | 0.0509 | 0.0310 | 0.0239 |
| RND | 3.6422 | 0.2819 | 21.72 | 0.1507 | 0.0002 | 0.0002 | 0.0002 |

Table 3: Performance of track recommendation (*LFM-1b*).

| Algo | P@1 | | P@10 | | nDCG@100 | |
|---|---|---|---|---|---|---|
| | female | male | female | male | female | male |
| ALS | 0.1701 | 0.3176 | 0.2193 | 0.1142 | 0.1323 | 0.1802 |
| POP | 0.0200 | 0.0329 | 0.0261 | 0.0073 | 0.0317 | 0.0092 |
| RND | 0.0001 | 0.0002 | 0.0001 | 0.0002 | 0.0001 | 0.0002 |

## 5.2 Gender Fairness on the Track Level

Table 2 summarizes the results on the track level considering the top-100 recommendations of the algorithms, using the *LFM-1b* dataset. *ALS* shows a large gender gap in the average first position (4.6993 vs. 24.9162 for male vs. female artists, respectively); by far larger than on the artist the level. Using *RND* provides a similar picture as on the artist level, and using *POP* results in similar positions for male and female artists when analyzed on the track level.

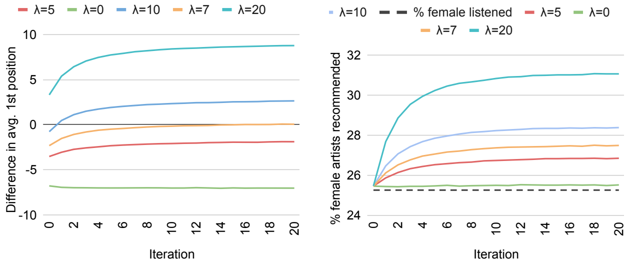
Overall, *ALS* recommends slightly more content by female artists compared to the users' listening behavior in the given dataset (28.99% vs. 25.33%). *POP* delivers a far higher percentage of content by female artists (66.66%) compared to other approaches. The Hellinger distance indicates that *ALS* delivers recommendations that are the closest to the gender distribution as reflected in the users' previous listening behavior. Also *RND* comes close to the original distribution, while *POP* does not.

While the last three columns of Table 2 present the performance metrics for all artists, we show these metrics differentiated by the artists' gender in Table 3. The results suggest for all precision metrics as well as for the ranking quality (nDCG) that lower performance is achieved for recommended female artists than for male artists when using the *ALS* algorithm. Using *POP* flips those results.

An additional analysis shows that the recommendations generated with *POP* cover the limited number of 130 tracks by female artists, compared to 18, 825 tracks with *ALS* and 100, 722 with *RND*.

## 5.3 Simulating Feedback Loops

We propose an ad-hoc approach to improve the exposure of female artists by penalizing male artists by moving them $\lambda$ positions in the ranking. We study the impact of different values for $\lambda$ on the exposure of female artists in the long term. Thereby, $\lambda = 0$ represents the baseline *ALS* without re-ranking. To this end, we use recommendation on the artist level and simulate the interaction of users with the top-10 recommendations for each iteration. We visualize two different aspects of exposure: First, Figure 1a shows the difference between the average first position of female and male artists for each iteration. Increasing $\lambda$ gives a more balanced exposure to female artists compared to the baseline without re-ranking ($\lambda = 0$). Depending on how fast the change is desired, different values of $\lambda$ may be preferred. Second, Figure 1b shows the evolution of the average percentage of female artists across the iterations for the different values of $\lambda$, and compares those to the consumers' listening behavior according to the *LFM-1b* dataset. Compared to the users' current listening behavior, using $\lambda = 7$, the percentage increases by almost 2 percentage points, whereas with $\lambda = 20$, it increases by more than 6 percentage points. Considering both views on gender fairness (Figures 1a and 1b) provides a good basis to decide on a $\lambda$ value. Using $\lambda = 7$ achieves a good balance in the long-term, which is aligned with the idea expressed by the artists (see Section 3.2) of progressively inducing a change in the behavior to a balanced exposure of female and male artists. An even higher exposure of female artists could be achieved with $\lambda > 7$.



(a) Avg. difference between the first position of female and male artist.

(b) Percentage of female artists recommended and originally listened.

Figure 1: Simulation of the exposure of female and male artists in the recommendations using different values of $\lambda$.

To investigate the potential performance loss when increasing the exposure of female artists, we compare the prediction accuracy achieved with the baseline *ALS* without re-ranking (i.e., $\lambda = 0$) and those achieved with different values for $\lambda$ for each iteration. Our analysis suggests that, in comparison to the baseline ($\lambda = 0$), the nDCG@10 is, on average, reduced by 2.2% for $\lambda = 5$, 4.9% for $\lambda = 7$, 6.7% for $\lambda = 10$, and 15.0% for $\lambda = 20$.

In addition, we analyzed the intervention of the re-ranking by looking at the average number of items that are re-ranked for each user in each iteration. Results suggest that the number of re-ranked items decreases with increasing iterations. In short, *ALS* starts recommending more females over time compared to the initial recommendations, and the effect of the feedback loop decreases once the users start changing their behavior.

## 6 CONCLUSION

In this work, we build on interviews with music artists, which suggest that artists would like to see balanced recommendations in terms of the artists' gender. Motivated by this finding, we investigated the effects of a collaborative filtering approach (here: *ALS*) on gender fairness. Results suggest that there is a considerable difference with respect to the average first position of female and male artists in the recommendation ranking. In short, the exposure of content by female and male artists is not balanced.

We follow the interviewed artists' expressed request to gradually give more exposure to female artists and propose a simple re-ranking approach. By simulating the feedback loop, we show that gender can be better balanced in a longer term when gradually increasing the exposure of female artists in the recommendations. This balance is achieved without severely affecting performance.

Future research should investigate alternative algorithms. A crucial path of research will be to study how consumers perceive the changes introduced by the re-ranking strategy in a real-world setting and how it impacts their listening behavior in the long-term.

## ACKNOWLEDGMENTS

This research was partially supported by Kakao Corp.

## REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation. In *Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. ACM, New York, NY, USA, 726–731. https://doi.org/10.1145/3383313.3418487

[3] Luis Aguiar, Joel Waldfogel, and Sarah Waldfogel. 2018. *Playlisting Favorites: Is Spotify Gender-Biased?* JRC Technical Reports JRC113503. European Commission, Seville, Spain. JRC Digital Economy Working Paper 2018-07.

[4] Christine Bauer. 2019. Allowing for equal opportunities for artists in music recommendation: a position paper. In *Proceedings of the 1st Workshop on Designing Human-Centric Music Information Research Systems* (Delft, The Netherlands) *(wsHCMIR '19)*. 16–18.

[5] Christine Bauer, Marta Kholodylo, and Christine Strauss. 2017. Music Recommender Systems: Challenges and Opportunities for Non-Superstar Artists. In *30th Bled eConference* (Bled, Slovenia), Andreja Pucihar, Mirjana Kljajić Borštnar, Christian Kittl, Pascal Ravesteijn, Roger Clarke, and Roger Bons (Eds.). University of Maribor Press, Maribor, Slovenia, 21–32. https://doi.org/10.18690/978-961-286-043-1

[6] Laura Beckwith, Margaret Burnett, Valentina Grigoreanu, and Susan Wiedenbeck. 2006. Gender HCI: What About the Software? *Computer* 39, 11 (2006), 97–101. https://doi.org/10.1109/MC.2006.382

[7] Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. 2005. Effectiveness of end-user debugging software features: Are there gender issues?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. ACM, New York, NY, USA, 869–878. https://doi.org/10.1145/1054972.1055094

[8] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. 2011. Precision-Oriented Evaluation of Recommender Systems: An Algorithmic Comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, IL, USA) *(RecSys '11)*. ACM, New York, NY, USA, 333–336. https://doi.org/10.1145/2043932.2043996

[9] Margaret M Burnett, Laura Beckwith, Susan Wiedenbeck, Scott D Fleming, Jill Cao, Thomas H Park, Valentina Grigoreanu, and Kyle Rector. 2011. Gender pluralism in problem-solving software. *Interacting with Computers* 23, 5 (2011), 450–460. https://doi.org/10.1016/j.intcom.2011.06.004

[10] Oscar Celma. 2010. Music recommendation. In *Music Recommendation and Discovery*. Springer, Berlin Heidelberg, Germany, Chapter 3, 43–85.

[11] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, BC, Canada) *(RecSys '18)*. ACM, New York, NY, USA, 224–232. https://doi.org/10.1145/3240323.3240370

[12] 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

[13] John W Creswell and Cheryl N Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches.* Sage Publications, Thousand Oaks, CA, USA.

[14] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. ACM, New York, NY, USA, 525–534. https://doi.org/10.1145/3351095.3372878

[15] Paul de Revere. 2015. A Bechdel Test for Music. *Pitchfork* (March 2015). https://pitchfork.com/thepitch/699-a-bechdel-test-for-music/

[16] Michael D Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Recommendation and Retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. ACM, New York, NY, USA, 576–577. https://doi.org/10.1145/3298689.3346964

[17] 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 *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, New York, NY, USA, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[18] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, BC, Canada) *(RecSys '18)*. ACM, New York, NY, USA, 242–250. https://doi.org/10.1145/3240323.3240373

[19] 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) *(ISMIR 2020)*. ISMIR, 248–254.

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

[21] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jsson 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.

[22] Andres Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring Longitudinal Effects of Session-Based Recommendations. In *Proceedings of the Fourteenth ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. ACM, New York, NY, USA, 474–479. https://doi.org/10.1145/3383313.3412213

[23] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82. https://doi.org/10.1177/1525822X05279903 arXiv:https://doi.org/10.1177/1525822X05279903

[24] 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

[25] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. ACM, New York, NY, USA, 49–58. https://doi.org/10.1145/3287560.3287600

[26] 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

[27] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

[28] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 88 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274357

[29] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x

[30] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-Centered Approaches to Fair and Responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375158

[31] Philipp Mayring. 2004. Qualitative content analysis. In *A companion to qualitative research*, Uwe Flick, Ernst von Kardoff, and Ines Steinke (Eds.). Sage Publications, London, United Kingdom, Chapter 5.12, 159–176.

[32] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Making Recommendations Better: An Analytic Model for Human-Recommender Interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI EA '06)*. ACM, New York, NY, USA, 1103–1108. https://doi.org/10.1145/1125451.1125660

[33] Danaë Metaxa-Kakavouli, Kelly Wang, James A. Landay, and Jeff Hancock. 2018. Gender-Inclusive Design: Sense of Belonging and Bias in Web Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3174188

[34] Rob Mitchum and Diego Garcia-Olano. 2018. Tracking the Gender Balance of This Year's Music Festival Lineups. *Pitchfork* (May 2018). https://pitchfork.com/features/festival-report/tracking-the-gender-balance-of-this-years-music-festival-lineups/

[35] Janice M. Morse. 1994. Designing funded qualitative research. In *Handbook of qualitative research*, Yvonna S. Lincoln and Norman K. Denzin (Eds.). Sage Publications, Thousand Oaks, CA, USA, 220–235.

[36] G. Oestreicher-Singer and A. Sundararajan. 2012. Recommendation Networks and the Long Tail of Electronic Commerce. *MIS Quarterly* 36, 1 (2012), 65–83.

[37] Ricardo S Oliveira, Caio Nóbrega, Leandro Balby Marinho, and Nazareno Andrade. 2017. A Multiobjective Music Recommendation Approach for Aspect-Based Diversification. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* (Suzhou, China) *(ISMIR 2017)*. ISMIR, 414–420.

[38] Yoon-Joo Park and Alexander Tuzhilin. 2008. The Long Tail of Recommender Systems and How to Leverage It. In *Proceedings of the 2008 ACM Conference on*

*Recommender Systems* (Lausanne, Switzerland) *(RecSys '08)*. ACM, New York, NY, USA, 11–18. https://doi.org/10.1145/1454008.1454012

[39] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. arXiv:cs.CY/1712.09124 http://arxiv.org/abs/1712.09124

[40] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, Boston, MA, USA, 1–35.

[41] Dimitris Sacharidis, Carine Pierrette Mukamakuza, and Hannes Werthner. 2020. Fairness and Diversity in Social-Based Recommender Systems. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. ACM, New York, NY, USA, 83–88. https://doi.org/10.1145/3386392.3397603

[42] Markus Schedl. 2016. The LFM-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (New York, NY, USA) *(ICMR '16)*. ACM, New York, NY, USA, 103–110. https://doi.org/10.1145/2911996.2912004

[43] Vaughn Schmutz and Alison Faupel. 2010. Gender and Cultural Consecration in Popular Music. *Social Forces* 89, 2 (2010), 685–707. http://www.jstor.org/stable/40984552

[44] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring Artist Gender Bias in Music Recommendation. In *Proceedings of the Workshops on Recommendation in Complex Scenarios and the Impact of Recommender Systems co-located with 14th ACM Conference on Recommender Systems (RecSys 2020)* (Online, 2020-09-25) *(ComplexRec-ImpactRS 2020)*, Vol. 2697. CEUR-WS.org, Article 1, 9 pages. http://ceur-ws.org/Vol-2697/paper1_impactrs.pdf

[45] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. 2020. Exploring User Opinions of Fairness in Recommender Systems. arXiv:cs.IR/2003.06461 https://arxiv.org/abs/2003.06461

[46] Stacy L. Smith, Marc Choueiti, and Katherine Pieper. 2018. *Inclusion in the Recording Studio?: Gender and Race/Ethnicity of Artists, Songwriters & Producers across 600 Popular Songs from 2012–2017.* Report. Annenberg Inclusion Initiative. http://assets.uscannenberg.org/docs/inclusion-in-the-recording-studio.pdf.

[47] 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

[48] Harald Steck. 2011. Item Popularity and Recommendation Accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, IL, USA) *(RecSys '11)*. ACM, New York, NY, USA, 125–132. https://doi.org/10.1145/2043932.2043957

[49] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, New York, NY, USA, 645–651. https://doi.org/10.1145/3308560.3317303

[50] Mihaela Vorvoreanu, Lingyi Zhang, Yun-Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From Gender Biases to Gender-Inclusive Design: An Empirical Investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300283

[51] Yixue Wang and Emőke-Ágnes Horvát. 2019. Gender Differences in the Global Music Industry: Evidence from MusicBrainz and The Echo Nest. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (7 2019), 517–526. https://ojs.aaai.org/index.php/ICWSM/article/view/3249

[52] 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.

[53] Ian Youngs. 2019. Pop music's growing gender gap revealed in the collaboration age. *BBC* (February 2019). https://www.bbc.com/news/entertainment-arts-47232677

[54] Ian Youngs. 2019. Pop music's growing gender gap revealed in the collaboration age. *BBC* (February 2019). https://www.bbc.com/news/entertainment-arts-47232677

[55] 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