## Thesis –title page, Table of Contents, Acknowledgements

Gender Bias in Spotify Recommendation Algorithms

## Abstract—summarize work and results

## Background—provides all relevant information for the audience to understand prior work

## and experimental or design tools (broken up into subsections)

1. Gender disparity in the music industry
2. Gender
3. Gender and algorithmic bias
4. Assessing algorithmic bias
   1. Types of algorithmic bias
   2. Statistical criteria for fairness (in recommendation algorithms)
5. Spotify specifically (experiments)
   1. Spotify discover weekly
   2. Spotify black box algorithm

Women in Music

It is no secret that the music industry, like many others, is dominated by men. The extent is difficult to measure and analyze due to the creative and individual nature of music. Since 2012, the USC Annenberg Inclusion Initiative has attempted to capture these inequities by analyzing inclusion on the Billboard Hot 100 Year-End Chart. In 2022, there were a total of 160 artists credited on the 100 songs featured on the Hot 100 Year-End Chart—69.4% men, 30% women, and 0.6% gender non-binary (Smith, 2023, 4). This is a 7.3 percentage point improvement since 2012 for women. They also analyzed songwriters and producers. Of the 451 songwriters on the Chart in 2022, 14% were women, and only 3 credits belonged to gender non-binary songwriters (Smith, 2023, 12). There was a total of 231 producers credited on the Chart in 2022, 3.4% of whom were women, and only one of whom was gender non-binary (Smith, 2023, 16). It is notably that all of these non-binary metrics—artists, songwriters, and producers—belong to the same nonbinary artist, Sam Smith (Smith, 2023). The USC Annenberg report captures the inequities in the music industry; however it falls short in explaining their causes and only addresses the most popular mainstream music. It largely ignores the way music consumption has become an individual experience through the rise of streaming.

While the USC Annenberg report addresses popular music, interviews from 17 “female digital musicians” describe the experience of being a self-taught musician performing online without a record label. The study focused on musicians using YouTube, Soundcloud, and Vine, however Spotify is another online platform where independent artists can promote their work. The musicians cite expectations of their physical appearance as a barrier to success. They see the differences in portrayal of male and female musicians in the music industry and rewarding of female sexualized body parts as reinforcing oppressive societal expectations about appearance for female musicians, even those who are not famous (Choi, 2016). This study captures the individual experience of being a female musician promoting work online including barriers to success related to the gender disparity in the music industry and gender-based oppression as a whole.

Gender

All the previously mentioned studies convolute sex and gender, excluding other gender identities and expressions. In her work “Queering feminist technology studies”, Landström suggests that heteronormativity influences feministic research on gender and technology. She calls for a theoretical shift in framework to account for technology’s coproduction of gender to avoid empirical research that relapses into old patterns of masculine and feminine binary (Landström, 2007).

Many studies on the classification of gender or technology and gender fall into this archaic thinking. For example, a study from Verizon, Yahoo, and Worcester Polytechnic Institute attempted to create a machine learning model to predict a user’s gender based on their username. They only allowed for two outcomes: male or female. They found a character-based logistic regression approach was most accurate for gender classification, though there are problems with all attempted gender classification models (Hu, 2021). Their work demonstrates that name is not enough information to make accurate gender predictions as well as the shortcomings of binary gender classification.

Gender classification is often attempted when a user’s gender is not supplied. Even as sites have expanded past the gender binary in account creation, the gender binary is reinforced elsewhere behind the scenes. An immerse ethnographic study of the 10 most popular English speaking social media sites revealed that while many sites are expanding gender options outside of the binary, demographic information is still heavily used in advertising. They stress the social implication of gender in social media and technology as they become ingrained in our society, especially for identity curation and performance (Rena, 2016).

Gender and Algorithmic Bias

This ignorance of gender identities as well as existing gender bias breeds bias in our algorithms which perpetuates the heteronormativity, ignorance, and oppression which contribute to it. Algorithmic bias extends far beyond the context and application of the algorithm itself and is already apparent in daily life. Perez’s “Invisible Women” details how the gender data gap, the systemic and historic lack of data collected on or about women, perpetuates the oppression that created it. Perez warns this biased input data leaves the impression that men are the default and majority. Ultimately, these machines are not only reflecting but also amplifying our biases (Perez, 2021).

Recommendation algorithms are particularly suited to amplify these effects. YouTube’s recommendation algorithm has already been shown to favor feminized content, creating a vicious cycle that rewards oppressive norms. YouTube systematically promoted feminized material by rewarding and promoting feminized content in beauty vlog recommendations. When failing to account for potential bias, developers create algorithms with unintended side effects which amplify existing societal problems such as discrimination, gender bias, and gendered societal expectations (Bishop, 2018). If recommendation algorithms are amplifying gender oppression on YouTube, similar algorithms may be amplifying it on Spotify.

These unintended effects of biased algorithms have unintended effects themselves. Researchers at Stanford found women receiving stereotypically “feminine” career recommendations had lower estimates of leadership ability than those who received traditionally “masculine” career recommendations and believed the recommendations were based on internal characteristics. An understanding of these systems, algorithms with implicit bias, moderates its effect on sense of self but cannot remove them entirely (French, 2018). Given the sexualization of female musicians and gender disparity in the industry, algorithmic bias could be perpetuating this disparity on a psychological level in those exposed to its effects.

Spotify

Spotify algorithms have been analyzed in a variety of ways in the past, though little is known about the gender discrepancy in their output. It is suggested that Spotify uses content-based and collaborative filtering to generate track representations. Content-based filtering includes the content of each track, such as artist name, songwriting credits, and genre, but also metrics Spotify calculates based on these attributes such as danceability as well as a semantic meaning estimated with natural language processing models. Collaborative filtering describes the track’s connection to other tracks and what other users with similar tastes are listening to. Each user on Spotify also has their own “taste profile” or Spotify’s understanding of what the user likes based on their listening history (Pastukhov, 2022).

While Spotify reveals a lot to users and artists about how its algorithms work, it is currently unknown whether artist gender is meaningfully included in the content-based filtering, and it is also unknown whether there is a data gap in music from female-identifying musicians based on the near infinite amount of content on Spotify. As such, it is difficult to know if there is gender bias in Spotify’s recommendation algorithms and where it is coming from.

These recommendation algorithms are particularly important in long-term user retention as long-term diversity of listening is associated with long-term retention. Researchers found that listeners given personalized recommendations based on their listening history had increased streams but decreased diversity compared to those given recommendations based on podcasts popular among users in their demographic group. They also stress the engagement-diversity trade-off: personalized recommendations drive immediate consumption, but long-term diversity is necessary for long-term user retention (Holtz, 2020). While this study focused on podcasts, the engagement-diversity trade-off also applies to music streamed and demonstrates the dangers both to the user and Spotify of personalized recommendation algorithms decreasing diversity.

A study of Spotify’s visual user interface as well as recommended songs attempted to understand how Spotify recommendation algorithms present gender when presenting music. They analyze the listening experiences of 80 young adults in Moscow and Stockholm using three common Spotify recommendations—Related Artists, Discover Weekly, and Browse. Related Artist recommendations are artists recommended based on a current artist the user is viewing. These recommendations were found to often have artists of the same gender and demographic group as the original. Browse algorithms are general playlists available for large communities of users and were the most diverse. The Discover playlist is a playlist of 30 songs generated weekly for each user individually and was less demographically cohesive, though still recommended music similar in genre, gender, and race. (Werner, 2020).

The study confirms that the Discover playlist is mostly unique to each user but did not address what other factors might determine the recommendations made there. Discover is made up of personalized recommendations which have been shown to be integral to the engagement-diversity trade-off and user satisfaction. As such, the diversity in these user playlists is particularly important. This particular playlist, Spotify’s “Discover Weekly”, though, has been shown to generate very little attachment between artists and fans for those who were not previously musically engaged and only relatively low attachment for those who were highly engaged (Leisewitz, 2022). Its significance and yet its lack of impact for the artists included confuse the overall role of the Discover playlist and question the accuracy of and satisfaction with its algorithmic recommendations.

Researchers at Linkoping University attempted to understand user’s gender’s impact on Spotify’s “black box” algorithm. They performed 288 bot experiments within four genres—rock, gospel, rnb/hip-hop, and dance-electronic—with each bot registering as a male or female user and listening to only the top 10 songs out of the top 100 songs of their given genre. The majority of bot users were given the same recommended artists regardless of their account’s gender, though gender-skewed recommendations were found in rock as well as rnb/hip-hop and dance/electronic. Gender classification of artists in this study was performed through text mining of artists’ information for pronouns, names, and photographs (Eriksson, 2017). This experiment concludes that account gender is not a heavily weighted factor in Spotify’s recommendation algorithm but does not further analyze these “gender-skewed” recommendations by category.

Italian researchers attempted to use streaming data from Spotify to assess gender discrimination among Italian singers. The study used the number of listeners on Spotify as a measure of success for each artist. They ultimately concluded there is no gender discrimination among Italian singers as there was no significant difference in number of listeners by gender, despite the average number of listeners for male singers being consistently higher than that of female singers (Nappo, 2019). Their analysis reveals a difficulty in measuring gender bias, despite the existing gender disparity in the music industry, especially using Spotify metrics. By choosing to compare an equal number of male and female artists, the study fails to account for the forces that make a user more likely to listen to a male artist from industry bias to user interface. Furthermore, number of listeners is an unstable metric for comparison as not all listeners are created equal (or contributing to an artist’s monetary success in the same way). One listener could have listened to an artist’s song once in a 28-day period, while another could have streamed the artist’s entire discography, and they each count as one listener. Due to this discrepancy, it is entirely possible for male and female artists to have similar numbers of listeners, but vastly different numbers of streams and therefore success.

Accessing Algorithmic Bias

Algorithmic bias, for reasons seen above, is notoriously difficult to analyze or prove. It also demands an answer to the question: what counts as algorithmic bias in an already biased society? Friedman and Nissenbaum define bias in computer systems as falling into three categories: preexisting bias, technical bias, and emergent bias. Preexisting bias is rooted in “social institutions, practices, and attitudes”, much like the gendered data gap and other ideas about gender-based oppression, while technical bias arises from “technical constraints and considerations”, and emergent bias “arises in the context of use”. Given this definition, even preexisting bias can qualify as bias in a computer system if it “systematically and unfairly discriminate[s] against certain individuals or groups of individuals in favor of others”().

This notion of preexisting bias as algorithmic bias raises more moral and ethical questions about the obligations of companies created machine learning algorithms and how best to combat the existing societal bias which will inevitably be reflected in them. [ADD SOMETHING ABOUT MORALITY OF FAIRNESS AND DISCRIMINATION AND ALSO ETHIC DILEMMA OF BIASED AI HERE]

Many statistical criteria have been developed with the goal of assessing how biased a given system may be. There are eleven widely accepted statistical criterion for algorithmic fairness for both continuous risk scores and binary predictions, however recommendation algorithms are not classification problems; therefore many fairness criteria cannot apply. For example, the first statistical fairness criterion for binary predictions requires equal false positive rates, meaning that “the (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group” (Eva). In the context of Spotify’s Discover Weekly, this would mean the expected percentage of artists who should not be recommended to a user (because they would be interested) who are wrongfully recommended anyway is the same among users of different gender identities. Due to the hidden inner-workings of Spotify’s algorithm as well as the nearly infinite sample space of available songs on Spotify, this true amount of individual songs that were classified as negative is impossible to know and likely unrelated to the algorithm’s function altogether, not to mention the subjective nature of music where an “incorrect” recommendation will depend on a myriad of individual factors.

Some of the logic reasons for these rules still apply to recommendation however, such as the tenth and eleventh criterion which suggest that statistical parity—“the (expected) percentage of individuals predicted to be positive is the same for each relevant group”—and equal ratios of predicted positives to actual positives—“the (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group”—should be used to judge fairness. In the context of Spotify’s recommendation algorithm, statistical parity would suggest that the rates of artists by gender included in a Discover Weekly playlist should be equivalent, while the eleventh rule would imply that if the base rates are not equal the algorithm’s output cannot be expected to be equal. Once again, there is no consensus on the moral or ethical obligation of an algorithm to combat preexisting bias, however if we choose to define bias in computer system as bias before, during, and after its use, then statistical parity is a more robust criteria for fairness.

Recommendation Algorithms

## Experimental procedure/Methodology (many pages long about all of the tools used as well)

1. General procedure
2. Specific decisions you made
   1. Individual artists conforming to the gender binary
   2. Users that are either man or woman (since gender is almost always classified into the binary when it is not supplied (Rena))
   3. Discover Weekly
   4. The genres you chose

In the context of Spotify recommendation algorithms, there are a multitude of biases at play. To differentiate them, they are defined by the population which they impact most directly, giving bias on the musician versus bias on the user. Bias on the musician would be an algorithmic bias which systematically promotes artists of one gender group over artists of other gender identities. The sources of this bias may not be Spotify’s system, but the gendered data gap, gender disparity in the music industry, societal ideas about gender, or the gender-based oppression baked into our social systems. Since this kind of bias may be the fault of society and not Spotify, it relies on our definition of algorithmic bias: is algorithmic bias only “wrong” if it is significantly worse than that existing outside of the algorithm? Or does algorithmic bias include this societal bias which may be baked into our systems? I choose to think algorithmic bias may have three sources—the before, during, and after (). Therefore, even if the bias’s source is before the context of the algorithm, the algorithm still demonstrates bias in its perpetuation. This kind of bias would mean that there is repeated and consisted underrepresentation of artists of specific gender identities across all findings, regardless of a user’s self-prescribed gender, listening history, preferred genre, or otherwise. Our data will provide insight to this kind of bias across all genres and within genres, by comparing the raw proportion of artists recommended by gender across all users (regardless of their gender identities) and genres.

The second kind of bias at play would be bias on the user, meaning bias that affects the user on the basis of their group membership. The sources of this bias may also not be malevolent. User bias is often an attempt at personalization and improved recommendations due to the user-retention-diversity tradeoff where recommendations must be both accurate and diverse enough to keep users engaged. Spotify’s subscription based model means the stakes are high for the accuracy or approval of their recommendations. Of course, there may also be factors related to societal ideas about gender and gender-based oppression. It is important to note that anything created by humans in a society will implicitly have some amount of bias.

Evidence of this kind of bias on the user would be indicated by two users of different genders (but the same in all other controllable ways) that received recommended output that is meaningfully different, as well as two users of the same gender identities which receive similar treatment. This difference in treatment could be in the proportion of artists recommended’s gender, energy score, happiness score, or numerous other metrics.

These metrics are accessed through the Spotify Web API are as follows:

1. Fdasjka
2. Fds
3. Afdhs

The experiment has been designed to analyze Spotify’s black-box recommendation algorithm for bias of both kinds. Six users—three women, three men—will be created with no previous listening history for each identified genre. These users will listen to the same ten songs every week—the top 5 songs from women in their genre and top 5 songs from men—in order to ensure the same training data is given to each user, and therefore each user’s model. At the end of each week, each user’s Discover Weekly playlist, a playlist curated by Spotify’s recommendation algorithm unique to each user, will be analyzed for discrepancies in artist gender and song metrics. Three users of each binary gender identity will allow us to compare results both within and between gender groups, and the five genres will similarly allow comparison overall and between genres to understand if some genres have more or less prevalent bias than others.

## Results—graphs, images, data tables and prose about what to notice

## Discussion—what do the results mean and why are they important (this will have to be very careful)

## Summary

## Limitations

## Conclusions

## References—this could be like 10 pages long

## Appendix