Translation, Tracks & Data: An Algorithmic Bias Effort in Practice

Henriette Cramer
Spotify
San Francisco, CA, USA
henriette@spotify.com

Jean Garcia-Gathright
Spotify
Boston, MA, USA
jean@spotify.com

Sravana Reddy
Spotify
Boston, MA, USA
sravana@spotify.com

Aaron Springer
University of California, Santa Cruz
Santa Cruz, CA, USA
alspring@ucsc.edu

Romain Takeo Bouyer
Spotify
New York, NY, USA
romaintakeo@spotify.com

ABSTRACT

Potential negative outcomes of machine learning and algorithmic bias have gained deserved attention. However, there are still relatively few standard processes to assess and address algorithmic biases in industry practice. Practical tools that integrate into engineers' workflows are needed. As a case study, we present two tooling efforts to create tools for teams in practice to address algorithmic bias. Both intend to increase understanding of data, models, and outcome measurement decisions. We describe the development of 1) a prototype checklist based on existing literature frameworks;

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3299057

and 2) dashboarding for quantitatively assessing outcomes at scale. We share both technical and organizational lessons learned on checklist perceptions, data challenges and interpretation pitfalls.

CCS CONCEPTS

• **Information systems → Personalization**;

KEYWORDS

Algorithmic bias; algorithmic accountability; bias and data checklist; industry practice

ACM Reference Format:

Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: An Algorithmic Bias Effort in Practice. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland Uk*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3290607.3299057

INTRODUCTION

After at least twenty years of academic research into computational bias [6], calls to address undesirable algorithmic outcomes have now gained mainstream traction. While much attention rightfully focuses on highly consequential algorithmic decisions, more subtle effects of algorithmic bias are ubiquitous and can negatively affect users' experience. Unfair algorithms may lead to a loss of trust from users [11]. Biases in voice recognition for example, lead to difficulties retrieving specific types of content, thus leading to a poor performance for specific subgroups of users [9]. Rather than waiting for potential negative effects to arise, it would be preferable to consider the potential effects of algorithmic decisions as a regular part of the development process. This requires new tools and processes that organizations can implement in practice.

As a concept in machine learning, bias has multiple connotations. For example, bias can refer to properties of data not fully representing the population being sampled [7]. It is also a property of machine learning models that underfit the training data. Finally, bias can refer to machine learning models unfairly favoring or disfavoring particular groups or individuals [6]. As all datasets and models are approximations of reality, we base our work on the principle that every dataset is 'biased' to some degree, and that all (non-)decisions made in a machine learning effort inherently have tradeoffs. We note that the phenomenon of negative consequences of algorithms do not require intent; rather, it often arises through unintentional oversights. The goal for practitioners then is to consider which characteristics of data, models, and outcomes are aligned with the goals that they want to achieve, and to guard against undesirable outcomes.

As a case study, we present the learnings from the early stages of establishing an algorithmic accountability effort within our own organization. The activities reported here are part of a wider algorithmic bias effort that combines research (e.g., case studies, metrics, literature review) with the development of organization-wide tools and processes. While we do not claim our experiences are universally applicable, nor that we have all the answers, we aim to provide initial results about tools that can help organizations address algorithmic bias.

APPROACH

Our approach to addressing algorithmic bias in practice roughly consists of three types of activities:

- **External engagement** with wider research and stakeholders.
- **Research**. Translating external research into Spotify-applicable insights (performing case studies on our products and data, informed by literature discussed in a reading group).
- **Methods and tools** within a shared organizational framework.

In this case study, we focus on lessons learned while developing two types of shared tools: 1) a 'bias checklist' for translating frameworks from literature into actionable insights, and 2) organization-wide dashboarding and data efforts needed to facilitate quantitative auditing.

CHECKLIST

Method

We summarized existing bias frameworks from literature into an easy-to-digest checklist prototype for practitioners within our organization. These summaries would enable product teams to educate themselves about different categories of biases and ask themselves questions about the characteristics of their data, models and outcomes in a structured way. We addressed assessment and possible interventions at three entry points:

(1) Training data characteristics.
(2) Model characteristics and team expertise. While these could be separated, this combination reflects an 'algotorial' approach, combining algorithmic and editorial decision making.
(3) Outcomes and their measurement, incl. potential benefits or harms for specific groups.

For each of these categories, teams need to know what to assess and address. Existing frameworks provide useful starting points, but need to be turned into practical tools. In our case, we evaluated a selection of bias frameworks (including Friedman & Nissenbaum [6], Olteanu et al [7], Baeza-Yates [1], and Crawford [3]), and translated them into a summary of bias types. In addition, we added team composition and expertise as additional factors. Our aim was to allow practitioner teams to use the checklist as part of the product design process, to anticipate bias-related issues, and to prioritize their prevention or mitigation.

Table 1: Data biases included in checklist prototype.

Type	Description
Population	Differences between characteristics of the dataset population and the target population
Behavioral	Differences in user behavior across platforms (e.g., mobile, voice) or contexts (e.g., work, party, family)
Temporal	Differences in populations or behaviors over time
Redundancy	Data items that appear in multiple copies, or artificially often (e.g., bots)
Content	Lexical, syntactic, semantic, or structural differences in how content is produced
Linking	Differences in the attributes of networks or user connections
Interface	Biases that result from interface design or shown rankings
Sampling	Biases resulting from data sampling choices

Lessons learned

Characteristics of different frameworks. We found that while Friedman and Nissenbaum's 1996 work on bias was prescient, their categories are harder to translate into underlying causes, which in turn makes it harder to use them to actively address issues. More recent frameworks put forth by Baeza-Yates, Olteanu et al., and Crawford point to *direct* and *indirect* harm, allowing us to classify problems in a way that could guide practitioners towards assessment and interventions.

For our purposes, Baeza-Yates' bias categories were a particularly usable example taxonomy as it forms a directed cycle graph; each step feeds biased data into the next stage where additional biases are introduced. The cyclical nature of bias makes it difficult to discern where to intervene, but this model can help break down the cycle and potential intervention targets. It also aligns with the informal feedback received from teams on the difficulties of assessing where to intervene when biases intersect. The framework presented by Olteanu et al. examines biases introduced at different levels of social data gathering and usage, including: user biases, societal biases, data processing biases, analysis biases, and biased interpretation of results. In order to create the initial checklist, we consolidated the categories from Baeza-Yates, Olteanu et al. and Crawford into the three bias entry points (see Tables 1, 2, and 3), providing us with a shared language. Each row of the checklist describes a bias category and asks 3 questions: whether this bias affects the project's outcome, what its priority would be, and what could be done. The outcome section of the checklist aimed to solicit consideration of intended benefits, neutral effects, and unintended harms.

Table 2: Model and team biases included in checklist.

Type	Description
Parameters	Side effects from model and parameter choices
Self-selection	Bias originating from those who choose (or choose not to) interact with the product
Team	Knowledge/experience gaps within the team (i.e., would you be able to recognize "obvious" problems?)

Challenges. As part of our efforts to educate product managers and engineers on algorithmic bias, we presented an overview of the three entry points for bias in internal machine learning course sessions and described the checklist. We discovered several challenges. First, applying the checklist in practice was a significant task. Representatives of product teams reported that the checklist required a great deal of effort to fill out. While perhaps useful as a didactic tool, the checklist appeared too overwhelming as a practical 'self-serve' tool.

Practically, producing a checklist does not surface specific stakeholders, characteristics or impacts to focus on; thus the list can become overwhelming on where to start. This information has to come from other qualitative insights, a deep familiarity with characteristics of content, and knowledge of the involved stakeholders. For example, in a streaming context, we must consider both needs of content creators, as well as consumers. A domain-specific choice has to be made on which characteristics are relevant and feasible to examine. In addition, the cyclical nature of bias' intersection with organizational structures also meant that a change in one pipeline may affect multiple services, with potentially unforeseen consequences for products in the wild. Getting organized using a shared framework becomes even more important to help deal with cases that span multiple teams. We therefore included a specific 'which team owns this' consideration in later checklist iterations. Simplification and support are however both still necessary.

QUANTITATIVE AUDITING

Data & dashboarding

In order to make an impact in organizations driven by metrics and data, a theoretical overview of potential biases needs to be accompanied early-on with concrete numbers. This especially applies in Agile contexts that strive for early and iterative delivery, which have become industry standard. This need for early numbers leads to a dilemma: in the beginnings of an algorithmic accountability effort, data and metrics are usually not readily available. It may not even be clear what the 'right' measures are and what their impact may be. At the same time, showing early numbers will paint a more concrete picture.

In our case, we started with a number of case studies, including an analysis of which content was less accessible in voice contexts [9], which indicated that genres with more creative language usage (hip-hop, country), or creators with non-English names and tracks were much less accessible than other types of content. Product-specific case studies also included performance snapshots of flagship recommendation features. However, while such case studies were informative, it became clear that it would be more effective to establish an ongoing effort to provide across-organization tooling to facilitate better understanding of characteristics of content across products. That way teams would not have to repeat basic data engineering work, analyses would be comparable, and we would move beyond rapidly outdated snapshots. This meant we had to: 1) identify the important consumption and creator characteristics that would be feasible to include in a data pipeline and dashboard, 2) identify a baseline of how to assess outcomes, and 3) implement iterative dashboarding and interpret the consumption patterns displayed. We describe a number of insights on potential analysis pitfalls.

Table 3: Outcome biases included in checklist prototype.

Type	Description
Content/creators	Intended or expected content gaps Unintended content gaps or negative consequences to test for
User	Intended or expected performance or satisfaction gaps (i.e., for which users is this going to work well, and for whom will it not?)

Lessons learned

Data challenges. When looking at all content consumed in larger organizations, a multitude of teams can be involved, all with their specific instrumentation efforts and content categorizations. Organizations, particularly large ones, may not necessarily have that data all in one place, nor use the same schemas or datasets. It is well advised to plan for a considerable amount of time to identify which data is available and appears reliable. In order to move beyond a 'data impasse', an executive decision may be necessary to pick a dataset that is good enough, while noting its imperfections. We found that existing datasets may be incomplete or limited in different ways.

1. Categories. We took advantage of prior knowledge of a creator-focused team on which characteristics were feasible as a first step (including popularity, gender, and locale), and already well-known to enable communication. Similar to Ekstrand et al. [4], we found that existing data is imperfect; in the case of gender, classifications were usually limited to male/female and catch-alls for other identities or missing data. Similarly, ensemble-based creators lead to questions on how to represent multiple band members. In our case, categories were female, male, unknown/other, and mixed for mixed gender bands. While this is neither ideal nor inclusive [10], it provided a starting point.

2. Data and types of content. In our case, 40+ million tracks are available on the platform from over 3.5 million creators, and new tracks come in every day[1]. Detailed data in a desired format is not always available from the start. For example, we found that 12.6% of a sample of top popular streams did not have gender data, with data rapidly declining for less popular artists. This has different impacts for different types of content, and makes it imperative to explicitly include missing data in analysis efforts. This coverage suffices to perform meaningful aggregate analyses, but impacts analyses in different ways. In general, our data on gender reflects that the music industry is not gender-balanced as a whole (reports estimate 22% of top pop artists being female, with producer estimates as low as 2% [1]). For example, for one sample we looked at that included all types of streams or playlists, female and mixed group creators represented 19% of programmed/recommended streams; while low, this was actually *higher* than non-programmed (actively user-chosen) streams, 17%. However, in one specific playlist focused on completely new content, we found that *all* gender categories were 'underrepresented' in programmed content compared to non-programmed consumption, simply because the 'missing' category was overrepresented (up to 75% of streams). If we had looked at only the representation of one gender, and not explicitly included the missing/other category in our analysis, this would lead to wrong conclusions.

3. Self-representation. Categorizations require in-depth understanding [10], and it may be undesirable to let others 'judge', especially when in conflict with desired self-presentation. This applies to gender, but also self-identification with cultural movements, or genres as in music or art. These are not trivial questions. From creator-oriented teams, we heard that not everyone wants to self-identify, and that questions may arise on what the consequence of information - or its absence - will be. In contrast, we had access to *on-platform behavior*, such as data based on the platform's popularity of a creator in relation to other creators. We decided to focus on outcomes (in Crawford's terms 'direct' benefits), defined as (programmed) streams. We, however, did not look at representation, see for example Epps and Dixon [5] on potential stereotyping through lyrical themes in hip-hop.

Challenges in interpretation. Metric interpretation is fraught with pitfalls, especially without full understanding of the factors that play into existing consumption patterns. In our case, genres as well as user preferences are crucial to understand. For example, holiday music, children's music and

[1] https://www.nytimes.com/2018/01/25/arts/music/music-industry-gender-study-women-artists-producers.html

spoken word content appear ‘underrepresented’ in our dashboard when we look at the percentage of programmed content compared to non-programmed consumption. However, having these types of content appear in recommendations would be a suboptimal experience for many. ‘Balanced’ consumption and recommendation for every user across all content creators is likely not possible.

In certain countries, such as Canada, the percentage of local content appeared higher than others. However, can streams of very popular Canadian artists, such as Drake or Justin Bieber, still be seen as local? Even one extremely popular creator could skew streamed results locally, making them appear ‘equal’ in aggregate. Ignoring these factors can lead to false conclusions, or even the full reversal of associations, a phenomenon known as Simpson’s paradox. Being able to look at the intersection of content characteristics (genre, gender, locale or other characteristics of interest), and popularity is important. This may appear obvious, and has been pointed out by [2], but it can require considerable skill and foresight to prepare data pipelines when dealing with millions of records. While many of the above issues are well-known within the research community, they may not be to practitioners. This presents a dilemma between providing wide, actionable access, and misinterpretation risks. Important is also to realize that while programmed content reaches up to 31% of streams[2], a track that gets recommended and streamed once is less impactful than a track later remembered and asked for.

DISCUSSION

As Sculley et al. point out, machine learning is the ‘high interest credit card of technical debt’ [8]. This also applies to the debt in terms of unintended characteristics of datasets, as well as the hidden dependencies that may occur over time. It is much easier to build in algorithmic accountability from the start of a project, rather than in a product that has been in production for years where changes may result in unpredictable side effects. It may also be more effective to use existing data and make small, visible steps that are accessible to an organization as a whole, rather than presenting individual teams with overwhelmingly detailed requests without providing help or context.

While the need for human help, organizational context and domain expertise are extremely unlikely to be automated, our experience implementing the checklist prototype and dashboards surface potential directions for tools that assist in addressing algorithmic bias. Challenges with the checklist point towards the need to (inter)actively help teams find underserved users or content, rather than leaving them with a lengthy ‘paper’ checklist. Our lessons learned point towards future research for tools that allow teams to collectively understand potential positive and negative impacts of their data decisions and algorithmic outcomes. Practical methods/tools could address the feedback regarding checklist complexity and difficulty knowing where to start. Whether this would be prioritizing, and breaking categories into simple, step-wise questions, or actively finding underserved clusters, depends

[2] http://investors.spotify.com

on the situation at hand. In addition, assisting with organizational overhead and surfacing pipeline dependencies would be beneficial. It is important to note that in some cases collecting data needed to analyze representation may be well-intended, but inadvisable. There can be serious harms in storing sensitive personal information, especially considering legality and safety in different regions on sexual orientation, religious or political affiliation.

CONCLUSION & WHAT'S NEXT

Applied researchers in industry have an important role in ensuring that calls to action on algorithmic accountability can be more effectively answered. While complex de-biasing or multi-objective machine learning methods can be very useful, they are not the primary concern in the early stages of assessing existing algorithmic biases. We encountered many practical challenges along the way in making literature more accessible, and getting to usable overviews. This points towards a distinct need to create research-informed, *and* usable, tools for data scientists, engineers and wider product teams to understand their data and algorithmic bias footprint. To improve experiences for end users and creators, we must also empower teams.

Thank you! We here build on the work and feedback from many internal and external colleagues. We would especially like to thank Rebekka Roerden, Avriel Epps, and Christine Hosey for their help and feedback.

REFERENCES

[1] Ricardo Baeza-Yates. 2016. Data and algorithmic bias in the web. In *Proceedings of ACM Web Science*. ACM, 1–1.

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[3] Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

[4] 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 RecSys'18*. ACM, 242–250.

[5] Avriel C Epps and Travis L Dixon. 2017. A Comparative Content Analysis of Anti-and Prosocial Rap Lyrical Themes Found on Traditional and New Media Outlets. *Journal of Broadcasting & Electronic Media* 61, 2 (2017), 467–498.

[6] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM TOIS* 14, 3 (1996), 330–347.

[7] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[8] D Sculley, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. Machine learning: The high-interest credit card of technical debt. (2014).

[9] Aaron Springer and Henriette Cramer. 2018. Play PRBLMS: Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 296.

[10] Elizabeth H Stokoe. 2004. Gender and discourse, gender and categorization: Current developments in language and gender research. *Qualitative Research in Psychology* 1, 2 (2004), 107–129.

[11] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 656.