Assessing and Addressing Algorithmic Bias—But Before We Get There...

Aaron Springer, Jean Garcia-Gathright, Henriette Cramer

Paper on the next page, accompanying info:

Interest in the workshop:
Our work resolves around the feedback loop between UX & Machine Learning based recommendations. Beyond general community building, we'd like to connect on the specific topic discussed in our paper. There is a growing amount of current research into reducing bias in machine learning, but it is very difficult to transform this research into actionable items for practitioners. We share our experience reviewing this research and creating actionable projects within the fast-paced competing demands that exist in industry.

Format: Should this be accepted we propose a presentation slot by Henriette.

Bios:
Aaron Springer
I am a PhD candidate at University of California Santa Cruz in the HCI Lab under Steve Whittaker. My research focuses on the intersection between machine learning and user experience, specifically the relationships between sensemaking, transparency, and trust. At Spotify, my research targets data and algorithmic bias in voice applications.

Jean Garcia-Gathright
I am a machine learning engineer at Spotify, where I develop and evaluate data-driven models of user preferences and satisfaction. I earned my PhD from the University of California Los Angeles, where I studied the representation and summarization of biomedical knowledge for assisting clinical decision making. Before that, I worked on software instrumentation for the Keck Interferometer at Jet Propulsion Laboratory.

Henriette Cramer
I'm a research lead at Spotify, where I focus on research on voice interactions, the human sides of machine learning, and how people, data and machines talk to each other about music. Before joining Spotify, I was part of Yahoo Labs, researching user engagement, recommendation, search and bot-related projects. Prior, I was a researcher at the Mobile Life Centre in Stockholm, Sweden working on human-robot interaction and location-based services. My original research background is in people's responses to adaptive and autonomous systems.

Assessing and Addressing Algorithmic Bias—But Before We Get There...

Aaron Springer[1,2], Jean Garcia-Gathright[1], Henriette Cramer[1]

[1]Spotify, 48 Grove St., Somerville, MA, USA
[2]University of California Santa Cruz, 1156 High St., Santa Cruz, California, USA
alspring@ucsc.edu, {jean, henriette}@spotify.com

Abstract

Algorithmic and data bias are gaining attention as a pressing issue in popular press—and rightly so. However, beyond these calls to action, standard processes and tools for practitioners do not readily exist to assess and address unfair algorithmic and data biases. The literature is relatively scattered and the needed interdisciplinary approach means that very different communities are working on the topic. We here provide a number of challenges encountered in assessing and addressing algorithmic and data bias in practice. We describe an early approach that attempts to translate the literature into processes for (production) teams wanting to assess both intended data and algorithm characteristics and unintended, unfair biases.

Introduction

There has been around 20 years of early research into the topic of algorithmic fairness and understanding of its outcomes (Friedman & Nissenbaum, 1996). The explosion of widespread machine learning has pushed algorithmic and data bias to the front lines of both the tech press and mainstream media. In parallel, specialized research communities are forming. Promising new initiatives such as the AI Now Institute have been initiated. The FATML workshop has turned into a full conference (FAT*, 2017), a new AAAI/ACM Conference on AI, Ethics, Society has been formed (AAAI, 2017), the ACM has now presented guidelines for algorithmic fairness (Dopplick, 2017). Pragmatically speaking, this increased attention to the topic is great, but these communities' calls to action are still very hard to apply. Pragmatic methods and tools are absolutely necessary to translate nascent research into work in industry practice - also pointed out by Kate Crawford in her WSJ op-ed (Crawford, 2017).

The proliferation of different communities, and the scattered literature presents industry practitioners with challenge to keep up, even when they're highly motivated. Reported studies or methods may also not be fully applicable in practice. We here outline a number of (early) lessons learnt from conversations with Machine Learning-oriented product teams, and thinking through the pragmatic translation of literature into practice.

Background

A wide variety of bias literature and a wide variety of definitions of bias exist. Bias, as a term in Machine Learning contexts, is used in somewhat divergent ways. Bias can be defined as unfair *discrimination*, or it can be framed as a system having certain *characteristics*, some intended and some unintended. Any dataset, and any Machine Learning-based application is 'biased' in the latter interpretation. This means we need to distinguish between unfair/unintended and intended biases. We base our work for practitioners on the pragmatic principle that any dataset is 'biased' in some way, that no dataset completely represents the world, and that human decisions in Machine learning systems inherently have tradeoffs that can result in (un)intended biases. The goal for product teams is to consider which characteristics of data, algorithms, and outcomes are aligned with the goals that they want to achieve - and side-effects.

For the purposes of this discussion, we take specific example definitions and frameworks. We use an adjusted definition from Friedman & Nissenbaum on 'Computational bias' as placeholder for (unfair) algorithmic bias: 'Discrimination that is systemic and unfair in favoring certain individuals or groups over others in a computer system' (Friedman & Nissenbaum, 1996). Where we use 'systemic' in our definition, Friedman and Nissenbaum used 'systematic'; we made this change to emphasize that algorithmic bias often arises through unintentional oversights rather than requiring specific biased intents as systematic implies. As definition for data bias, we use Olteanu et al.'s 'a systemic distortion in the data that compromises its representativeness' (Olteanu, Castillo, Diaz, & Kiciman, 2016) as starting point. Note however that this raises an immediate dilemma: if data is completely representative of reality, it will also reflect the very real societal biases and existing disadvantages, and could potentially simply echo or amplify these. This means that 'biasing' the data against these biases may be important (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). We here supplement that definition with representativeness 'necessary for the application at hand' and perhaps 'representative' of the world teams would like to represent.

Frameworks and Types of Biases

Friedman and Nissenbaum present a taxonomy of biases in computational systems with top level categories of Preexisting Bias, Technical Bias, and Emergent Bias (Friedman & Nissenbaum, 1996). While Friedman and Nissenbaum's work was often prescient, it is difficult to use this taxonomy to address algorithmic and data bias issues in practice. Their categorization does not point to underlying causes, making it somewhat challenging to use the framework in a solutions oriented manner.

More recent taxonomies of algorithmic and data bias allow us to classify problems in a way that points out how to intervene and correct biases. The Baeza-Yates taxonomy consists of 6 types of bias: activity bias, data bias, sampling bias, algorithm bias, interface bias, and self-selection bias (Baeza-Yates, 2016). These biases form a directed cycle graph; each step feeds biased data into the next stage where additional and new biases are introduced. The cyclical nature of bias makes it difficult to discern where to intervene; models like Baeza-Yates' help break down the cycle and find likely targets for initial intervention.

Though biases exist and are propagated through all types of data, one of the most common types of data that practitioners currently use is social data. Social data encompasses content generated by users, relationships between those users, and application logs of user behaviors (Olteanu et al., 2016). The framework presented by Olteanu et al. comprehensively 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.

Translation Into Bias Identification Processes

A major challenge is translating the growing, but scattered literature into a step-by-step process that works in practice. Unfortunately, in many cases the methods to assess, and certainly how to address a problem are not yet available.

The first step to correcting algorithmic biases is *identification* of *potential biases,* for which we have three possible entry points:

- Biases in input data
- Computational biases that may result from algorithm + team decision
- Outcome biases, for example for specific user groups (gender, age) or for specific domains (e.g. having really good recommendations in one genre over another).

Per definition, the first two categories can be done even before a project has been started, whereas the third category requires domain knowledge and at least a predictive model. Particularly challenging is that to be able to measure outcomes, we need to not only assess which facets would be important to explore, but also which evaluation metrics are actually valid - which is many cases can be very large projects themselves.

After the identification of potential issues, a prioritization has to be made of which of these issues are most pressing, and how to assess them. While eliciting bias targets from the bottom up is a positive initial route, it is still essential to prioritize which biases to tackle first. The problem is often not that identifying potential algorithmic biases is a difficult task; it is that looking for candidate algorithmic biases will surface a large number that it becomes difficult to determine which biases to tackle first and which are currently intractable and better suited as long term goals. Some bias targets are clearly long term, e.g. finding that a highly used metric loses much of its predictive power for subpopulations, while others may require simpler changes like modifying a data sampling paradigm for training models. It is essential that these bias targets are prioritized by evaluating impact on users. Note that prioritizing simply on size of the affected population alone would lead to biases in itself, and that, on the other hand, slightly degraded user experience for a subpopulation may not be fixable or require effort best spent elsewhere. Weighing these bias targets against each other involves a complex decision involving level of harm, ubiquity of bias, and business driven priorities.

After assessment, very specific domain knowledge will be necessary to fix the bias at hand. Very promising projects exist focused on debiasing particular techniques, see (Bolukbasi et al., 2016) for debiasing word embeddings, but there is no guarantee that those methods will exist for your specific problem. In large settings, multiple issues may interact - and very pragmatic challenges can be encountered as well.

Domain Challenges

Every application will have different bias issues to assess and address. For example, voice interfaces are rapidly gaining popularity, but, unfortunately, voice interfaces may amplify bias due to their unique affordances. Voice interfaces may struggle with regional accents (Best, Shaw, & Clancy, 2013). Language dialects also may result in worse accuracy and voice recognition (Tatman, 2017). Even if dialects and accents were perfectly recognized by voice interfaces, these interfaces would still struggle to counteract biases using common solutions from other modalities. Recommender systems often suffer from popularity bias, meaning that popular content is recommended far more frequently than the long tail of less popular items (Abdollahpouri, Burke, & Mobasher, 2017). Solutions to enhancing discoverability of the long tail of content include increasing serendipity and novelty among recommendations (Vargas & Castells, 2011). Unfortunately, users are often trying to accomplish a task quickly by voice and listing 10 search results that include some popular, some novel, and some serendipitous results may degrade the user experience because of the time it takes to verbally list them. Therefore, this task of countering popularity bias may be much harder in voice where only one result is often returned. The voice realm may be challenging to properly correct biases in but that does not make the task impossible.

A major struggle with many types of bias research is understanding whether the metric differences measured are due to algorithmic/data bias or simply due to natural demographic variation (Mehrotra et al., 2017). Bias audits often require splitting the population sample in a way that we can measure metric differences across these samples but this action confounds itself because each sample may behave differently to begin with. Given this confounding challenge, a particularly effective way to measure and correct bias may be finding problems where a ground truth answer is available. Springer et al. examine the types of content that current voice interfaces underserve due to content characteristics (under submission). For example, current voice interfaces often transcribe dialect speech into Standard American English; this can result in a user asking for a music track titled "You Da Baddest" and the voice interface transcribing and searching for "You're the baddest" which may not result in finding the intended track. These entity resolution difficulties fortunately mean that ground truth is available, people can either access the track through the voice interface or not. With the availability of ground truth, we can tease apart the algorithmic bias from demographic differences and quickly identify ways to correct bias.

Every modality and every domain will require its own assessment methods and solutions. The challenge is to develop processes that are relatively light-weight to a variety of teams to communicate and implement, while still progressing towards a more equitable product.

Pragmatic Challenges

In this section, we present a few examples of pragmatic challenges that may be encountered when attempting to mitigate data and algorithmic bias in an industry setting. First, value must be established to motivate the prioritization of reducing specific unfair biases in production systems. Next, the work be developed in a way that harmonizes with the engineering practice of rapid delivery. Finally, longer-term changes in engineering culture are necessary to address bias as early as possible.

Prioritizing Correcting Bias

Engineering teams abide by a carefully planned roadmap of deliverables, with much energy devoted to maintaining their current systems and pushing new features to product. Setting aside time to measure and correct bias has to compete with other pressing priorities. It becomes hard to prioritize such projects where it's unclear how to assess their impact. Methods are not yet available, and case studies from literature demonstrate the extensive effort and expertise necessary (Mehrotra et al., 2017). Furthermore, in a situation where features built from imperfect data have already been surfaced in the product, making significant changes in the feature may be perceived as too risky. Framing such work in terms of business goals, such as improving performance across markets and improvement of quality, is a compelling argument for pursuing this work (compared with, for example, unspecified appeals that bias should be important).

Proposing Minimum Viable Products

Agile development is arguably the dominant approach to product development in startups. In an Agile-style environment, there is an emphasis on quick delivery of minimum viable products followed by continuous iteration. In order to translate research on bias to solutions in product, it is necessary to propose a minimal solution that can be delivered and then improved. For example, is it possible to move forward with solutions on narrow use cases or with imperfect measurements? Caution is required here, to prevent the minimum viable product from simply being accepted as the final product. Long-view thinking is also necessary, so that even as imperfect products are delivered quickly, there is still a path of iteration toward a more ideal solution. In larger companies, as datasets and APIs will be developed as services for other product teams, it becomes important to develop ways of documenting data characteristics in ways understandable beyond the direct team that developed these.

Addressing Technical Debt Via Cultural Changes

In the early stages of a company's development, the issue of scaling globally seems impossibly distant. In this scenario, teams may accumulate technical debt as a result of limited access to resources and data. For example, they may train models on themselves in the absence of user data, or quality evaluations may by necessity have to be adhoc, resulting in models that reflect the demographics or tastes of the developers. Even as the user base grows, models may be overfit to current users rather than performant a global market. When company growth reaches a point where global scaling becomes a priority, new perspectives and attitudes are necessary. Diversity in hiring becomes more important. Longer-term cultural change and education toward bias-awareness would also encourage engineers to design models and features with delivery to a global audience in mind, avoiding bias-related technical debt at the outset of the design process.

To make sure that processes and tools land in practice, they have to be lightweight, pragmatic and easy to communicate to a wide variety of teams. This should include how to prioritize, and how to assess their user and business impact.

Discussion

To assess and address algorithmic biases, teams need lightweight tools to make these processes their own; rather than calls for action from elsewhere. These tools currently do not exist. We are currently encountering the pragmatic challenges in translating the growing literature into methods that are applicable across domains and easy to communicate, while still informative enough to be of help.

Actively involving teams on the ground in this process is absolutely crucial. Shared understanding within industries and sharing of developed methods and lessons learnt, combined with a bottom-up application of frameworks by teams themselves appears most fruitful. An expert researcher coming in to a new team with a model in hand to examine systems will surely identify potential biases. However, it would be difficult to understand the finer details and potential side-effects. Changing datasets can have unforeseen effects elsewhere, how infrastructures, services, data and different parts of applications interact can be hard to understand when not deeply involved in its development process. Prescribing specific methods from afar will not work. Ensuring that team-embedded data scientists and data engineers themselves have tools and easily accessible resources to understand what to look out for, would be more fruitful.

References

AAAI. (2017, September 20). AAAI/ACM Conference on AI, Ethics, and Society – February 2-3, 2018. New Orleans, USA. Retrieved November 2, 2017, from http://www.aies-conference.com/

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling Popularity Bias in Learning to Rank Recommendation. In *Proceedings of the 11th ACM conference on Recommender systems. ACM, To appear*.

Baeza-Yates, R. (2016). Data and algorithmic bias in the web (pp. 1–1). ACM Press. https://doi.org/10.1145/2908131.2908135

Best, C. T., Shaw, J. A., & Clancy, E. (2013). Recognizing words across regional accents: the role of perceptual assimilation in lexical competition. In *INTERSPEECH* (Vol. 2013, p. 14th). Retrieved from http://www.academia.edu/download/45189070/Recognizing_words_across_regional_accent20160428-24622-16pjop6.pdf

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349–4357). Retrieved from http://papers.nips.cc/paper/6227-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings

Crawford, K. (2017, October 17). Artificial Intelligence—With Very Real Biases. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/artificial-intelligencewith-very-real-biases-1508252717

Dopplick, R. (2017, January 14). New Statement on Algorithmic Transparency and Accountability by ACM U.S. Public Policy Council. Retrieved November 2, 2017, from https://techpolicy.acm.org/?p=6156

FAT*. (2017, August 5). FAT*. Retrieved November 2, 2017, from https://fatconference.org/

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., & Yilmaz, E. (2017). Auditing Search Engines for Differential Satisfaction Across Demographics. *arXiv:1705.10689 [Cs]*. https://doi.org/10.1145/3041021.3054197

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2016). *Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries* (SSRN Scholarly Paper No. ID 2886526). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2886526

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *EACL 2017*, 53.

Vargas, S., & Castells, P. (2011). Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (pp. 109–116). New York, NY, USA: ACM. https://doi.org/10.1145/2043932.2043955