Measuring and Mitigating Bias and Harm in Personalized Advertising

Muhammad Ali
Northeastern University
Boston, MA, USA
mali@ccs.neu.edu

ABSTRACT

Online personalized advertising is often very effective in identifying relevant audiences for each piece of content, which has led to its widespread adoption. In today's internet, however, these advertising systems are used not only to market products, but also consequential life opportunities such as employment or housing, as well as socially important political messaging. This has led to increasing concerns about the presence of algorithmic bias and possible discrimination in these important domains — with results showing problematic biases along gender, race, and political affiliation, even when the advertiser might have targeted broadly.

A growing body of work focuses on measuring and characterizing these biases, as well as finding ways to mitigate these effects and building responsible systems. However, these results often emerge from different scientific communities and are often disconnected in the literature. In this paper, I attempt at bridging the gap between isolated efforts to either measure these biases, or to mitigate them. I discuss how the need to measure bias in advertising, and the efforts to mitigate it, despite being distant in the literature, are complementary problems that need to center their methodolgy around user studies.

This paper presents a research agenda that focuses on the need for user-centric measurements of bias, by collecting real ads from users, and using surveys to understand user perceptions for these ads. My approach also calls for incorporating user sentiments into the mitigation efforts, by constraining optimization on user values that emerge from surveys. Finally, I also emphasize the need for involving users in the evaluation of responsible advertising systems; efforts to mitigate bias eventually need to be contextualized in terms of benefits to users instead of simple performance tradeoffs. My focus on the users is motivated by the fact that they are stakeholders in personalized advertising, vulnerable at the hand of algorithmic bias and harm, and therefore crucial in both efforts to measure and mitigate these effects.

ACM Reference Format:
Muhammad Ali. 2021. Measuring and Mitigating Bias and Harm in Personalized Advertising. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460231.3473895

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8458-2/21/09.
https://doi.org/10.1145/3460231.3473895

1 INTRODUCTION

Personalized advertising is ubiquitous throughout the internet today, and supports a large part of its economy, with players like Facebook and Google controlling a major share of this market [15]. These large-scale advertising systems provide their users the promise of serving *relevant* content (and ads), while giving advertisers precise targeting tools to reach said users. Advertising on these platforms also extends beyond products and services to include particularly important life opportunities, such as employment and housing.

The expansion of advertising into such sensitive domains has recently raised concerns of discrimination, as these opportunities might be shown to certain demographic groups more than others, or some groups might be disproportionately exposed to lower quality or predatory offers. One reason these differences might arise is because of the advertiser building discriminatory audiences using the platform's tools [3, 17]. To avoid such advertiser behavior, platforms build policies around sensitive ad categories — Facebook, for instance, enforces policies around housing, employment and political ads [1], as well as dating ads [2]. While there is precedent for advertisers evading these policies [8], discrimination arising from advertisers can by and large be curtailed through stronger enforcement of targeting policies.

Arguably a much more insidious way for these differences to arise can be through the ad delivery process, where the ad platform needs to subselect the most relevant users (by its measures) from an advertiser's larger targeted audience. This introduces the complex issue of algorithmic bias in advertising, where different kinds of opportunities might be skewed by demographic group because of training on biased data. It could also have broadly pernicious effects on its users — such as exposing them to misinformation, trapping them in ideological echo chambers, or pushing them towards more extreme political views, in an effort to increase user engagement.

These ad delivery algorithms (and advertising systems in general) are often powered by recent advances in recommender systems research, and more broadly by machine learning methods for personalization [11, 14, 19, 21, 22] (representation learning, multi-armed bandits etc.). Considering the important domains such methods end up being deployed in, it is imperative that personalization systems themselves account for the possibility of user harm and bias during their optimization. The potential for such harm is particularly illustrated on platforms like Facebook and Twitter — where ads are embedded into a browsing feed and the divide between organic

[1] Facebook Business Help: Choosing a Special Ad Category, https://www.facebook.com/business/help/298000447747885

[2] Facebook Business Help: Run a Compliant Dating Ad, https://www.facebook.com/business/help/143949649021372

content and advertising is practically indistinguishable. On these platforms, multiple personalization systems, such as feed ranking, content selection, and ad auctions, work in conjunction. Therefore, designing all these optimization systems without understanding their downstream uses, can possibly compound their harms, leading to technology that is ultimately harmful to its users, and society at large.

While there has been work on mitigating possible biases in recommendation systems, much of it is often designed with the premise of minimizing performance tradeoffs [16], and often resorts to tangential measures, such as popularity bias [12], or theoretical fairness metrics [10]. The larger problem of designing recommendation systems that can avoid perpetuating stereotypes, or locking users in filter bubbles, remains to be meaningfully tackled.

This paper lays out a research agenda for advancing the state-of-the-art in both measuring and mitigating the effects of bias and harm in recommender systems, particularly in the context of personalized advertising. The remaining paper is structured as follows: Section 2 discusses the measurement and accountability results that have been accomplished in the current literature, the domains in which advertising has been measured, and the techniques that have been used. Section 3 lists the efforts to control bias in advertising, and in personalization at large. Both Section 2 and Section 3 are structured to first review the current literature, and then identify open problems that remain to be solved. Section 4 lays out a plan on how the highlighted remaining questions around measurement and mitigation can be tackled in a realistic manner. Section 5 provides a concluding discussion, with an emphasis on why understanding bias and harm in recommendation and personalization systems is an inevitably important task, and one that is extremely timely given the discourse around fairness and accountability in artificial intelligence.

2 MEASURING BIAS AND HARM

2.1 Related Work

One of the earliest studies identifying the potential for bias in advertising came from Sweeney [18], empirically showing that searching for African-American names was more likely to return ads for background-checking websites, compared to searching for European-sounding names. Datta et al. [7] expanded the result, showing through randomized controlled experiments, that setting a participant's gender to female led to seeing fewer ads for executive career coaching in Google's advertising system. Furthermore, Datta et al. also train a classifier to predict which demographic group the ads belong to, and establish through permutation testing that the differences learnt by the classifier were significant — suggesting a causal connection between gender and the career coaching ads [6, 7]. While the opacity and complexity of personalized advertising on Google makes it challenging to pinpoint the exact reason of these differences, prior work has attempted to explain them. Factors such as advertiser targeting or manual curation might be an explanation, but optimizations done for better personalization (often resulting from user behavior) could also clearly be a contributing factor [6].

The magnitude of bias that personalization itself can cause has been highlighted in the literature recently; studies that investigate the bias in ad targeting attributes (used by advertisers) provide a window into these differences. For Facebook's advertising, prior work has shown the extent of demographic skew that can exist simply during the targeting of the ad [17]. This leads to potential for discrimination on the advertiser's end, where even if racial targeting is prohibited, they can choose features that have high correlation with race to achieve discriminatory ad delivery [17]. The potential to combine these high correlation targeting attributes to compound their effects has also been documented for major platforms like Facebook, Google, and LinkedIn [20]. This leads to situations where malicious advertisers have access to tools that enable discriminatory ads, and sincere advertisers might be unaware of how biased their audiences are to begin with. While bias that can arise during ad targeting is a major concern, it is arguably much more insidious when these differences can appear during the ad delivery phase, where the ad platform's personalization algorithms are primarily responsible for them.

Towards understanding ad delivery biases, Ali and Sapieżyński et al. have demonstrated how these effects can appear even in the absence of the advertiser's intent [1]. They design a series of controlled trials to show, even under identical targeting, the optimizations done during the ad delivery phase (based on ad content) can create skews along gender and race of the targeted audience. These effects are shown to persist even for legally protected ad categories such as employment and housing, coming dangerously close to discrimination under U.S. law. Imana et al. [9] have further strengthened these results, demonstrating the existence of gender bias in job ads, even when controlling for qualifications of the advertised job. These effects exist beyond employment and housing as well, including for political advertising [2], which can reinforce users' filter bubbles and harm the democratic process.

The ubiquity of these biases across platform and domains suggests that it might be a fundamental byproduct of over-optimizing for user engagement. While enforcing advertising policies for different domains is a viable short-term solution, there is a clear need for designing advertising systems that are able to account for the harm they might produce, and avoid it.

2.2 Open Questions

As discussed in Section 2.1, the current literature is rich in unique approaches for measuring personalization systems. Owing to the interdisciplinary nature of the problem, techniques from internet measurement, computational social science, causality, and many others, have contributed towards an understanding of the harms that can afflict users.

What remains missing, however, is an understanding of how users perceive these harms. Domains such as housing, credit and employment are legally protected; and political ads have a normatively sensitive and important place in our society, but beyond these clearly defined areas, there is room to better understand what users themselves find harmful or helpful. There is currently a dearth of work that involves user studies and attempts to understand user perceptions of ads. This is an essential question because advertising, at its core, is a user-facing technology, and in order to meaningfully fix it, users must be part of the process. Section 4 goes into further details on how, concretely, users can be integrated into the measurement process to better understand their perceptions.

3 MITIGATING BIAS AND HARM

Owing to the popularity of multi-armed bandits in advertising systems [11, 16, 19], a majority of the reviewed literature revolves around bandit based optimization.

3.1 Related Work

Joseph et al. [10] formally introduce the notion of fairnes in bandit optimization, with a particular focus on situations where the algorithm might be choosing between different demographic groups for a decision. Their fairness algorithm focuses on merit, enforcing at each time step that a worse candidate is not preferred over a better one. While a direct translation to advertising in particular isn't provided, Joseph et al.'s results are broadly motivated by Sweeney [18]. Celis et al. [4] provide a general-purpose framework to broadly curtail *polarization* in personalization, which is directly applicable to advertising, alongside other domains that use content selection. They achieve this by specifically defining *groups* of arms in a bandit that might correspond to different content themes at risk of polarization, and constraining the likelihood of picking content at the group level. While generally applicable and useful, their approach requires clearly defined content themes (groups), as well as mapping of all content to these themes, for each kind of harm to avoid. As an alternative to directly controlling the content selection algorithm, modifying the bidding strategies has also been proposed as a solution to avoid gender discrimination [13].

Similarly, approaches arising from industrial research choose to focus on some notion of fairness or user health, while balancing performance tradeoffs. Mehrotra et al. [12] present a group fairness criterion and constrain their personalization system to ensure equity of attention by popularity, in the context of music recommendations. Singh et al. [16] focus on recommendation *trajectories* in particular, defining a notion of user health while watching videos. Their definition of harm isn't based on individual pieces of content, but rather a trajectory of videos that a user might explore. Their proposed method constrains on harm to the worst-off users, and is able to control this worst-case harm without a significant drop in overall performance.

3.2 Open Questions

While the literature discussed in Section 3.1 provides novel, valuable approaches, often times these results aren't contextualized for users themselves. Controlling a defined measure of harm for the worst-off users [16], or a theoretical definition of fairness [10] are valuable contributions, but are not trivial to translate to personalized advertising. There is also a need to understand whether such a translation from prior work would align with the users' definition of bias.

Furthermore, current literature often does not avoid the pitfall of using tradeoffs as an evaluation metric. Since mitigating bias and harm is a fundamentally user-centric endeavor, evaluating these efforts with users in the loop is crucial. Framing these evaluations as tradeoffs [12, 16] shifts the objective from protecting users, and can make these changes seem like technical drawbacks instead of safeguards for long-term user value. Section 4 goes into further details about my proposed methodology on how to integrate user values into algorithm design.

4 PROPOSED WORK

This section proposes approaches to address the limitations in literature identified in Section 2 and Section 3. I suggest focusing on the user for both measurements and mitigations of bias and harm. This focus on the user translates to multiple steps in the personalization pipeline, as discussed below.

Understanding User Values. As briefly mentioned in Section 2.2, there is currently a need for understanding how users themselves perceive harm in different advertising domains. This can help inform new domains of advertising (e.g. food, scams etc.) that users are concerned about, and where ad delivery algorithms might be biased. Unlike employment, credit, or housing, these domains might not have clear legal protections, but building a diverse set of domains can lead to a broader understanding of user harms. Furthermore, in addition to understanding what users find uncomfortable, it is also important to understand what they find helpful — since these measurements can help inform optimization objectives that need to be emphasized for a user-centric advertising platform. The eventual goal of involving users in this process should be to (a) elicit an understanding of what they consider sensitive, and (b) understand the objectives that are valuable to users.

Concretely, I propose collecting real, in situ advertisements from a diverse panel of participants (e.g. through a browser extension), and evaluating participants' perceptions of the collected ads through surveys. In line with the objective of identifying new sensitive advertising domains, I also propose asking users to identify ads that they think are likely predatory and could cause outsized harm to them.

Encoding User Values into Constraints. Similar to the current literature where fairness/accountability goals are achieved through constraints on optimization [4, 16], I propose designing constraints corresponding to the set of user values obtained through surveys. As an example, Celis et al.'s [4] framework provides a natural fit for such a goal, where each sensitive advertising domain can constitute a content *group*, and users' sentiments towards the domain can be used to judge the strength of the constraints. Other frameworks such as Mehrotra et al.'s [12] also allow constraints to satisfy secondary stakeholders in a personalization system, which in this case are the users. I argue that encoding the perceptions of real users into personalization can help ensure that we are not solving towards a contrived fairness goal, but something that is eventually valued by users themselves.

User-centric Evaluations. Finally, as mentioned in Section 3.2, changes to personalization need to be evaluated in terms of gains to users, and not as performance tradeoffs. I propose evaluating the constrained personalization system with the help of both conventional measurements such as accuracy, relevance etc., as well as user surveys of the eventual output. Prior work has shown valuable insights that can be obtained through mixed methods, pairing qualitative user interviews with quantitative log analysis in the case of information retrieval [5]. I suggest similarly expanding the evaluation for advertising systems by reporting both accuracy as well as user perceptions from surveys.

I hypothesize that comparing user evaluations pre- and post-treatment can provide insights into the realistic impact of responsible recommendations. An effective advertising system should be able to minimize any discomfort to users that might arise from seeing harmful ads — therefore, comparing measures such as the fraction of ads that cause discomfort, or the fraction of ads found useful, would be just as important in this evaluation as the overall relevance.

My proposed approach involves the user at every step of the design process and informs the values from the users themselves. This is an important and conscious distinction from the current literature, and shifts the focus back to users, who are arguably the biggest stakeholders in the face of algorithmic bias and harm. A limitation of such participatory design is that it might make the process more involved and challenging, requiring translation between users and machine learning algorithms, and then back again. However, given the human nature of the task, the investment to incorporate user feedback into the process is justifiable.

5 CONCLUSION

In this paper, I present a research agenda around measuring and mitigating algorithmic bias and harm in online advertising. I highlight approaches in prior work that measure bias in advertising along race, gender, and political affiliation, for consequential advertising domains such as employment, housing and politics. I also provide an overview of work in the recommender systems and personalization literature that works on mitigating these biases. I identify the lack of involvement of users as a limitation in the current literature, and propose approaches to bridge this gap. My proposed approach involves designing a user study to collect ads seen by users, and surveying recruited users to understand their perceptions. I hypothesize that a more nuanced understanding of sensitive advertising domains (beyond employment, housing, and politics etc.) lies in understanding user perceptions of ads. My proposal also suggests incorporating users' understanding of harms as constraints while optimizing personalization systems, to ensure that the artifacts of bias can be curtailed. Finally, I also propose that efforts to mitigate bias in advertising be measured through user studies in conjunction with accuracy and relevance, and not merely as performance tradeoffs.

I argue that centering questions of bias on users is a necessary step for long term solutions. Given that users might be at the receiving end of the harms, the mitigations need to factor in their sensitivities into the design process. I emphasize that moving towards mitigations with user-defined objectives could be a more meaningful approach than defining these objectives ourselves.

ACKNOWLEDGMENTS

I am grateful to my advisor, Alan Mislove, for help in structuring and framing the conversation for a diverse set of readers. This work was funded by NSF grant CNS-1955227.

REFERENCES

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery can Lead to Biased Outcomes. In *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*.

[2] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2021. Ad Delivery Algorithms: the Hidden Arbiters of Political Messaging. In *Proceedings of ACM Conference on Web Search and Data Mining (WSDM)*.

[3] Julia Angwin and Terry Parris Jr. 2016. Facebook Lets Advertisers Exclude Users by Race. ProPublica, https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/.

[4] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 160–169.

[5] Praveen Chandar, Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, and Jennifer Thom. 2019. Developing Evaluation Metrics for Instant Search Using Mixed Methods Methods. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 925–928.

[6] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K. Mulligan, and Michael Carl Tschantz. 2018. Discrimination in Online Advertising: A Multidisciplinary Inquiry. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, New York, NY, USA, 20–34.

[7] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proc. of PETS*.

[8] Laura Edelson, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. An Analysis of United States Online Political Advertising Transparency. *arXiv preprint arXiv:1902.04385* (2019).

[9] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. (2021).

[10] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in Learning: Classic and Contextual Bandits. (2016).

[11] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 31–39.

[12] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*. 2243–2251.

[13] Milad Nasr and Michael Carl Tschantz. 2020. Bidding strategies with gender nondiscrimination constraints for online ad auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 337–347.

[14] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019). https://arxiv.org/abs/1906.00091

[15] Kenneth Olmstead and K Olmstead. 2014. As Digital Ad Sales Grow, News Outlets Get a Smaller Share. *Pew Research Center* (2014).

[16] Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, H Chi, Jilin Chen, and Alex Beutel. 2020. Building Healthy Recommendation Sequences for Everyone: A Safe Reinforcement Learning Approach. (2020).

[17] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 5–19.

[18] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[19] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. 2013. Automatic Ad Format Selection via Contextual Bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1587–1594.

[20] Giridhari Venkatadri and Alan Mislove. 2020. On the Potential for Discrimination via Composition. In *Proceedings of the ACM Internet Measurement Conference*. 333–344.

[21] Xinxi Wang, Yi Wang, David Hsu, and Ye Wang. 2014. Exploration in Interactive Personalized Music Recommendation: a Reinforcement Learning Approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 1 (2014), 1–22.

[22] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A Deep Reinforcement Learning Framework for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 167–176.