HAL
open science

On Measuring Bias in Online Information

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, Gerhard Weikum

To cite this version:

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, et al.. On Measuring Bias in Online Information. SIGMOD record, 2018, pp.1-6. hal-01638069v2

HAL Id: hal-01638069

https://inria.hal.science/hal-01638069v2

Submitted on 20 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On Measuring Bias in Online Information

Evaggelia Pitoura, Panayiotis Tsaparas
Department of Computer Science and Engineering
University of Ioannina, Greece
{pitoura,tsap}@cs.uoi.gr

Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos
Institute of Computer Science, FORTH, Greece
{fgeo,fundul,papadako}@ics.forth.gr

Serge Abiteboul
INRIA & ENS, Paris, France
serge.abiteboul@inria.fr

Gerhard Weikum
Max Planck Institute for Informatics, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Bias in online information has recently become a pressing issue, with search engines, social networks and recommendation services being accused of exhibiting some form of bias. In this vision paper, we make the case for a systematic approach towards measuring bias. To this end, we discuss formal measures for quantifying the various types of bias, we outline the system components necessary for realizing them, and we highlight the related research challenges and open problems.

1. INTRODUCTION

We live in an information age where the majority of our diverse information needs are satisfied online by search engines, social networks and media, news aggregators, e-shops, vertical portals, and other online information providers (**OIPs**). For every request we submit to these providers, a combination of sophisticated algorithms produce a ranked list of the most relevant results tailored to our profile. These results play an important role in guiding our decisions (e.g., where should I dine, what should I buy, which jobs should I apply to), in shaping our opinions (e.g., who should I vote for), and in general in our view of the world.

Undoubtedly, the various OIPs help us in managing and exploiting the abundance of available information. But, at the same time, the convenient and effective way in which the OIPs satisfy our information needs has limited our information seeking abilities, and has rendered us overly dependent on them. We rarely wonder whether the returned results represent all different viewpoints, and we seldom escape the echo chambers and filter bubbles created by personalization. We have come to accept such results as the "de facto" truth.

There are increasingly frequent reports of OIPs exhibiting some form of bias. For instance, in the recent US presidential elections, Google was accused of being biased against Donald Trump[1] and Facebook of contributing to the post-truth politics[2]. Google search has been accused of being sexist or racist when returning images for queries such as "nurse" or "hairstyling"[3], and prejudiced when answering queries about holocaust[4]. Similar accusations have been made for Flickr, Airbnb and LinkedIn. In fact, the problem of understanding and addressing bias is considered a high-priority problem for machine learning algorithms and AI for the next few years[5].

The problem has attracted some attention in the data management community as well [34]. In this paper, we make the case for a systematic approach to addressing the problem of bias in the data provided by the OIPs. Addressing bias involves many steps. Here, we focus on the very first step, that, of defining and measuring bias.

According to the Oxford English Dictionary[6], bias is *"an inclination or prejudice for or against one person or group, especially in a way considered to be unfair"*, and as *"a concentration on or interest in one particular area or subject"*. When it comes to bias in OIPs, we make the distinction between *user bias* and *content bias*. User bias appears when different users receive different content based on user attributes that should be protected, such as gender, race, ethnicity, or religion. Content bias refers to biases in the information received by any user, such as, when some aspect is disproportionately represented in a query result or in news feeds.

In the rest of this paper, we present the related work, describe formal measures for both user and content bias, and outline the basic components of a system for realizing these measures. Finally, we provide a synopsis of challenges in identifying bias in online information.

[1] https://www.theguardian.com/us-news/2016/sep/29/donald-trump-attacks-biased-lester-holt-and-accuses-google-of-conspiracy
[2] https://www.theguardian.com/us-news/2016/nov/16/facebook-bias-bubble-us-election-conservative-liberal-news-feed
[3] http://fusion.net/story/117604/looking-for-ceo-doctor-cop-in-google-image-search-delivers-crazy-sexist-results/
[4] http://www.bbc.com/news/technology-38379453
[5] https://futureoflife.org/ai-principles/
[6] https://en.oxforddictionaries.com/definition/bias.

2. RELATED WORK

In the field of machine learning, there is an increasing concern about the potential risks of data-driven approaches in decision making algorithms [2, 3, 17, 24, 31, 34], raising a call for equal opportunities by design [21]. Biases can be introduced at different stages of the design, implementation, training and deployment of machine learning algorithms. There are reports for discriminatory ads based on either race [33, 35], or gender [11], and recommendation algorithms showing different prices to different users [19]. AdFisher [9] runs browser-based experiments to explore how user behaviors and profiles affect ads and if they can lead to seemingly discriminatory ads. Consequently, there are efforts for defining principles of accountable algorithms[7], for auditing algorithms by detecting discrimination [32] and for debiasing approaches [1, 44]. There is a special interest for racial fairness and fair classifiers [20, 41, 42, 7], ensuring that groups receive ads based on population proportions [11] and reducing the discrimination degree of algorithms against individuals of a protected group [15]. Other efforts try to ensure temporal transparency for policy changing events in decision making systems [14]. Recently, tools that remove discriminating information[8], help in understanding opposing opinions[9], flag fake news[10], increase transparency of personalization algorithms[11], or show political biases of Facebook friends and news feed[12] have started to appear.

Another branch of research focuses on how bias can affect users. According to field studies, users of search engines trust more the top-ranked search results [29] and biased search algorithms could shift the voting preferences of undecided voters by as mush as 20% [12]. Since most users try to access information that they agree with [22], the personalization and filtering algorithms used by search engines lead to echo chambers and filter bubbles that reinforce bias [4, 18]. This is also evident in social media where platforms strengthen users existing biases [25], minimizing the exposure to different opinions [37]. Rating bubbles emerge especially when positive social influence accumulates, while crowd correction neutralizes negative influence [27].

Previous studies have looked at individual aspects of bias, such as geographical (i.e. whether sites from certain countries are covered more) [36], or temporal (recommending recent and breaking news) [6]. Other approaches try to examine how bias can be measured [26] and if search engines can partially mitigate the rich-get-richer nature of the Web and give new sites an increased chance of being discovered [16]. The presence of bias in media sources has been studied based on human annotations [5] and by exploiting affiliations [38] and the impartiality of messages [40], while [23] tries to quantify bias in Twitter data. There is clearly a need for a systematic approach to identifying bias in online information, and in this paper, we outline some required steps and related challenges to this end.

[7]http://www.fatml.org/resources/principles-for-accountable-algorithms
[8]http://www.debiasyourself.org/
[9]https://www.escapeyourbubble.com/
[10]http://www.theverge.com/2016/12/15/13960062/facebook-fact-check-partnerships-fake-news
[11]https://facebook.tracking.exposed/
[12]http://politecho.org/

3. TYPES OF BIAS

We consider bias in terms of *topics*. In particular, we would like to test whether an OIP is biased with respect to a given topic. A topic may be a very general one, such as, politics, or a very specific one down to the granularity of a single search query. For example, we may want to test whether an OIP provides biased results for events such as"Brexit" and "US Elections", people such as "Donald Trump", general issues such as "abortion" and "gun control", transactional queries such as "air tickets", "best burger", or even topics such as "famous people". An OIP may be biased with respect to one topic and unbiased with respect to another one.

We distinguish between two types of bias, namely *user* and *content* bias. User bias refers to bias against the users receiving the information, while content bias looks at bias in the information delivered to users.

For user bias, we assume that some of the attributes that characterize the user of an OIP are *protected* (e.g. race, gender, etc.). User bias exists when the values of these attributes influence the results presented to users. For example consider the case of a query about jobs, where women receive results of lowered paid jobs than men. User bias can also appear due to hidden dependencies between protected and unprotected attributes, even when such protected attributes are not used directly in computing the results (e.g., see [13]). For instance, the home location of users may imply their race.

Content bias refers to bias in the results provided by the OIP and may appear even when we have just a single user. For example, an instance of this kind of bias occurs when an OIP promotes its own services over the competitive ones, or, when the results for queries about a political figure take an unjustifiable favorable, or unfavorable position towards this politician (independently of the user receiving the results).

In most cases, the OIP content is presented in the form of a ranked list of results. Results are often complex objects, such as news feeds, web pages, or, even physical objects, in the case of recommendations. We assume that results can be described by features, or attributes, either explicitly provided, or intentionally extracted. In analogy to protected attributes for users, we consider *differentiating attributes* for topics. For instance, for a controversial topic such as "abortion" or "gun control", the differentiating attribute could be the stance (pro, or against). For a topic such as "famous people", we may want to test whether the results are biased towards men over women, or, favor people from specific countries, or, over-represent, say, artists over scientists. Finally, for a topic such as "US Elections", a differentiating attribute may be the political party (with values, "Democrats" or "Republicans").

In a sense, addressing user bias can be regarded as a

counterweight to machine-learning and personalization algorithms that try to differentiate the needs of various user groups, so that these algorithms do not discriminate over specific protected attributes. On the other hand, addressing content bias has some similarity to result diversification [10]. However, diversity is related to coverage, since we want all various aspects of a topic, even the rarest ones, to appear in the result. For content bias, we want the differentiating attributes to be represented proportionally to a specific "ground truth".

A commonly encountered case is the case of a combined user and content bias appearing when a specific facet is over-represented in the results presented to a specific user population, e.g., democrats get to see more pro-Clinton articles than republicans. This type of bias is also related to *echo chambers*, i.e., the situation in which information, ideas, or beliefs are amplified, exaggerated or reinforced inside groups of equally-minded people. Since similar people may be interested in specific aspects of a topic, as a result the content they create, consume, or prefer is biased towards these aspects. Then, the information presented to them may reflect this bias and by doing so possibly amplify the bias, creating a bias-reinforcement cycle. In such cases, there is often some relation between the protected attributes of the users and the differentiating attributes of the topic.

4. BIAS MEASURES

In this section, we present measures for user and content bias. Our goal is not to be overly formal, but instead we provide such measures as a means to make the related research challenges more concrete.

We assume that the information provided by an OIP is in the form of a ranked list R. In the core of each bias measure lies a definition of similarity between lists of results. For now, let us assume that given two ranked lists of results R_1 and R_2, there is a distance function $D_R(R_1, R_2)$ that measures the distance between these two rankings. D_R can be defined by employing existing distance metrics between ranked lists, or using a geometric embedding of the ranked lists that takes into account both the similarity between results in the list and the importance of their position. We will revisit this issue when we talk about content bias.

To simplify the discussion, in the following, we assume that the topic for which we want to measure bias is a single query q. We can generalize the definitions to a set of queries by adopting some aggregation measure of the metrics for a single query.

User Bias. Let U be the OIP user population. For simplicity, assume a binary protected attribute that divides users into a protected class P and an unprotected class $\bar{P}$. For example, if the protected attribute is gender, P may denote the set of women and $\bar{P}$ the set of men. Intuitively, we do not want the information provided to users to be influenced by their protected attributes.

The problem of user bias is somehow related to fairness in classification, where individuals are classified in a positive or negative class. Example applications include among others hiring, school admission, crime risk factor estimation, medicine (e.g., suitability for receiving a medical treatment) and advertisement selection.

There are two general approaches to defining fairness, namely *group* and *individual fairness* [11]. Group fairness imposes requirements on the protected and unprotected class as a whole. A common example of group fairness is *statistical parity*, where the proportion of members in the protected class that receive positive classification is identical to the proportion in the general population. Individual fairness requires that similar people are treated similarly. A problem with group fairness is that it does not take into account the individual merits of each group member and may lead in selecting the less qualified members of a group. On the other hand, individual fairness assumes a similarity metric of the individuals for the classification task at hand. Such metrics are very hard to define.

A technical difference between fairness and user bias is that most work in fairness focuses on classification tasks, while, in our case, results are ranked. Very recent work addresses fair ranking (where the output is a ranked list of individuals) by adopting a group based approach that asks for a proportional presence of individuals of the protected class in all prefixes of the ranked list [39, 43]. A conceptual difference between the two problems is that in the case of fairness, users are the ones who are being classified (or ranked), whereas in user bias, the users are the ones who receive ranked information.

An individual-based approach to user bias assumes that it is possible to define an appropriate distance measure D_u between the users in U. The distance should capture when two users are considered similar for the topic under consideration. For instance, if the topic is jobs, individuals with the same qualifications should be considered similar independently of their gender. The following definition is based on the premise that similar users should receive similar result lists.

Definition 1 (Individual User Bias). *An online information provider is individual user unbiased if for any pair of users u_1 and u_2, it holds:*

$$D_R(R_{u_1}, R_{u_2}) \leq D_u(u_1, u_2)$$

where R_{u_1} and R_{u_2} are the result lists received by u_1 and u_2 respectively.

There are many ways of capturing group-based user bias. We will discuss one. Let $\mathcal{R}_{\mathcal{P}}$ be the union of the result lists seen by the members of the protected class and $\mathcal{R}_{\bar{\mathcal{P}}}$ be the union of the result lists seen by the members of the non-protected class. We could aggregate the results in each of them to create two representative ranked lists, R_P and $R_{\bar{P}}$, for $\mathcal{R}_{\mathcal{P}}$ and $\mathcal{R}_{\bar{\mathcal{P}}}$, respectively. We can define user bias using these representative ranked lists.

Definition 2 (Group User Bias). *An online information provider is group user unbiased if it holds:*

$$|D_R(R_P, R_{\bar{P}})| \leq \epsilon$$

for some small $\epsilon \geq 0$.

Aggregating result lists is just one possibility. For instance, another definition is to require the probability that a member of P receives any of the lists in $\mathcal{R}_{\mathcal{P}}$ to be the same with the probability that a member of $\bar{P}$ receives it (and vice versa). All group-based definitions ignore the profiles of individual users; i.e., they do not capture the fact that a result list should be relevant to the specific individual in the group who receives it.

Content Bias. Let us first assume that there is just one user. Let A be a differentiating attribute, and let $\{a_1, ..., a_m\}$ be the values of A. For example, in the case of a query about elections, $a_1, ..., a_m$ may correspond to the different parties that participate in the elections. We also assume that each result is annotated with the values of attribute A, meaning that the result is about these values.

A distinctive characteristic of content bias is that it can be defined only relatively to some "ground truth", or "norm". But what is the "ground truth"? One option is to consider the actual data used by the OIP for computing the content delivered to users as the ground truth. For example, this is the approach taken in [23] that compares the political bias of Twitter search with the bias in all tweets that contain the search terms. However, user-generated content may include biases inflicted by the design and affordances of the OIP platform, or by behavioral norms emerging on each platform. Bias can also be introduced by the OIP during the acquisition of data (e.g. during crawling and indexing for a search engine). See [28] for a complete analysis of the different biases and pitfalls associated with social data. There are also cases, where the actual data used by the OIP may not be available, as with search engines.

Ideally, we would like to have an indisputably unbiased ranked list of results. Such lists could be constructed using an aggregation of OIPs and other external sources such as knowledge bases, or domain experts. Crowdsourcing could also be utilized in creating such lists. In some cases, an estimation of the distribution of values of the differentiating attributes in the general population may be available. For example, for the election query, we could use external sources, such as polls, to estimate the actual party popularity and user intention to vote. One could also think of creating bias benchmarks consisting of reference sample topics and result lists similar to TPC benchmarks for evaluating database system performance, and TREC tracks for evaluating relevance in information retrieval.

Given the ground truth as an "ideal unbiased ranking" R_T, we could define content bias looking at its distance from the ground truth.

Definition 3 (content bias). *An online information provider is content unbiased if it holds:*

$$|D_R(R_u, R_T)| \leq \epsilon$$

for some small $\epsilon \geq 0$.

One way of defining D_R is by looking at the distribution of the values of the differentiating attribute in an ideal ranking. Assume that we have the "ground truth" in the form of probabilities $Pr_T(a_i)$ for all the attribute values which captures the relative popularity of each value (e.g., the support of a party as measured by polls). Let $Pr(u, a_i)$ be the probability that user u receives a result annotated with value a_i (e.g., one possible definition is this to be defined as the fraction of the top-k results that are about a_i). The following equation could serve as a definition of D_R.

$$D_R(R_u, R_T) = \max_i |Pr(u, a_i) - Pr_T(a_i)| \quad (1)$$

Combined User-Content Bias. We can refine user bias, using content-aware distance definitions, such as the one in Equation (1). For example, in Definition 1, we could use:

$$D_R(R_{u_1}, R_{u_2}) = \max_i |Pr(u_1, a_i) - Pr(u_2, a_i)| \quad (2)$$

Equation (2) looks at the relative bias of the content seen by two users. Although both users may receive biased content with respect to ground truth, there is no user bias as long as content is equally biased.

To look for echo chambers, we need to test the content bias in the result lists seen by different users. For instance, adopting the representative list approach to user bias, we may look at the distance of R_P and $R_{\bar{P}}$ from the ground truth to test, for example, whether specific attribute values are over-represented in the results shown to a population group.

5. A SYSTEM FOR MEASURING BIAS

We now look at some of the challenges involved in realizing a system for measuring the bias of an OIP. The OIP may be a search engine, a recommendation service, the search or news feed service of a social network. In Figure 1, we present the main components needed by a system, called BiasMeter, for measuring bias. We treat the OIP as a black-box and assume that BiasMeter can access it only through the interface provided by the OIP, e.g., through search queries. For simplicity, we assume that the set of protected and differentiating attributes are given as input to BiasMeter.

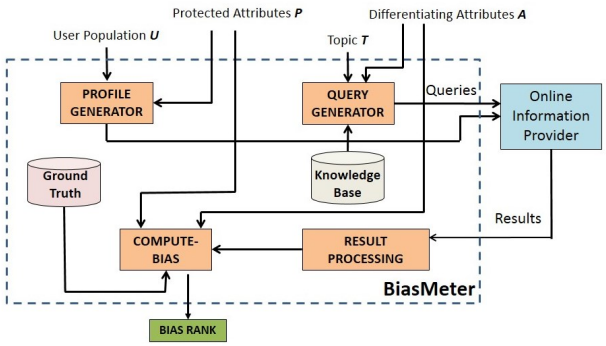

Figure 1: System components.

Given the topic T and the differentiating attributes A, the goal of the *query generator* is to produce an appropriate set of queries to be submitted to the OIP under

consideration. For instance, if the OIP is a search engine, to test about the topic "US elections", the generator may produce a variety of search queries, including queries referring to specific political parties. To produce queries that best represent the topic and the attributes, the query-generator may need to use background knowledge, such as, a related knowledge base.

The *profile generator* takes as input the user population U and the set of protected attributes P and produces as output a set of user profiles appropriate for testing whether the OIP discriminates over users in U based on the protected attributes in P. For example, if we want to test gender bias in job search queries, we need samples of men and women, that have very similar characteristics with respect to other attributes such as grades, skills, background, ethnicity, etc, to avoid differences that may appear due to attribute correlations.

There are many issues of both a theoretical and a practical nature in generating profiles. For example, we must ensure that the profiles are an appropriate sample of U that represents all values of the protected attributes. Furthermore, we should ensure that the characteristics of the users in the sample are similar with respect to all other attributes, so as to avoid the effect of confounding factors. This raises issues similar to those met when selecting people for opinion polls, surveys, etc. From a more practical view, we need to assemble users with the specific profiles and ask them to issue the queries (for example using a crowd-sourcing platform, such as Mechanical Turk), or generate artificial accounts of such users. An important step to automated profile generation is offered by AdFisher, a tool for testing discrimination in Google Ads [9]. AdFisher builds user profiles by just using the Ad profile setting and by simulating visits at specific webpages.

The *result processing* component takes as input the results from the OIP and applies machine learning and data mining algorithms such as topic modeling and opinion mining to determine the values of the differentiating attributes. For example, if the topic is "gun control", we need to determine whether a specific result takes a positive, neutral or negative stand.

Finally, the *compute-bias* component calculates the bias of the OIP, using our bias metrics and the *ground-truth*. Note that the cause of bias is not specified in our result; we just detect bias with respect to specific user and content attributes.

6. RESEARCH CHALLENGES

Obtaining the ground truth. Defining the ground truth is the most formidable task in identifying bias. One approach could be a human-in-the-loop approach where users take the role of data processors characterizing the bias of online information, similarly to users evaluating the relevance of search results. One can even envision novel crowdsourcing platforms specifically targeting bias evaluation. However, such tasks are hindered by strong cognitive biases, such as confirmation bias, that may lead users in discrediting as biased any information that does not fit their own believes. Furthermore, bias, as opposed to relevance, may involve political, ideological, or, even, ethical connotations. Besides crowdsourcing, one can envision a form of data-driven validation that integrates information from large data repositories, knowledge bases, and multiple OIPs. Besides this long-term quest for ground truth, a more realistic approach is to rely on comparative evaluations. For instance one could compare the bias between the results of two OIPs or between the results of an OIP and content found in traditional media.

Defining bias measures. Bias is multifaceted. We abstracted the many forms of bias, through the notions of protected attributes for users and differentiating attributes for content. However, there are often correlations among the attributes making it hard to single out the effects of each of them in the results. Further, our measures are high level, and a lot of work is needed to come up with rigorous mathematical formulations.

Engineering and technical challenges. To measure bias with respect to a protected attribute P (e.g. gender), we need to generate large samples of user accounts for the different values of P (e.g., women and men), making sure that the distribution of the characteristics for the other attributes is near identical. Careful statistical analysis is also required to ensure statistical significance of our results. In addition, the query generation and result processing components involve a variety of data mining and machine learning algorithms for identifying keywords to describe an information need, or understanding the topic and stance of a specific result. To this end, we need modules for knowledge representation, record linkage, entity detection and entity resolution, sentiment detection, topic modeling, and more.

Auditing. Bias detection can be simplified, if access is given to the internals of the OIP (e.g., for sampling users with specific demographics, or getting non personalized results). Clearly, this is impossible for an entity outside the OIP and it requires the cooperation of law and policy makers. Such access would also help in differentiating between bias in the source data and bias in the results. There is a growing literature advocating the systematic auditing of algorithms [8, 30, 32].

7. CONCLUSIONS

In this paper, we argue about the importance of a systematic approach for measuring the bias of the information we get from online information providers. As more people rely on online sources to get informed and make decisions, such an approach is of central value. Building a system for measuring bias raises many research challenges, some of which we have highlighted in this paper. Measuring bias is just the first step; many more steps are needed to counteract bias including identifying bias sources and developing debias approaches.

8. REFERENCES

[1] G. Adomavicius, J. Bockstedt, C. Shawn, and J. Zhang. *De-biasing user preference ratings in recommender systems*, volume 1253, pages 2–9. CEUR-WS, 2014.

[2] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *SSRN eLibrary*, 2014.

[3] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *CoRR*, abs/1607.06520, 2016.

[4] E. Bozdag. Bias in algorithmic filtering and personalization. *Ethics and Inf. Technol.*, 15(3):209–227, Sept. 2013.

[5] C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

[6] A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi. Can trending news stories create coverage bias? on the impact of high content churn in online news media. In *Computation and Journalism Symposium*, 2015.

[7] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.

[8] K. Crawford. Can an algorithm be agonistic? ten scenes from life in calculated publics. *Science, Technology, & Human Values*, 41(1):77–92, 2016.

[9] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[10] M. Drosou and E. Pitoura. Search result diversification. *SIGMOD Record*, 39(1):41–47, 2010.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012.

[12] R. Epstein and R. E. Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *PNAS*, 112(20), 2015.

[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015.

[14] M. Ferreira, M. B. Zafar, and K. P. Gummadi. The case for temporal transparency: Detecting policy change events in black-box decision making systems. *arXiv preprint arXiv:1610.10064*, 2016.

[15] B. Fish, J. Kun, and Á. D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SDM*, pages 144–152, 2016.

[16] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences*, 103(34):12684–12689, 2006.

[17] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *KDD*, pages 2125–2126. ACM, 2016.

[18] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring personalization of web search. In *WWW*, pages 527–538. ACM, 2013.

[19] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Internet Measurement Conference*, pages 305–318, 2014.

[20] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[21] W. House. Big data: A report on algorithmic systems, opportunity, and civil rights. *Washington, DC: Executive Office of the President, White House*, 2016.

[22] D. Koutra, P. N. Bennett, and E. Horvitz. Events and controversies: Influences of a shocking news event on information seeking. In *WWW*, pages 614–624, 2015.

[23] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, I. Shibpur, I. K. P. Gummadi, and K. Karahalios. Quantifying search bias: Investigating sources of bias for political searches in social media. In *CSCW*, 2017.

[24] B. Lepri, J. Staiano, D. Sangokoya, E. Letouzé, and N. Oliver. The tyranny of data? the bright and dark sides of data-driven decision-making for social good. *arXiv preprint arXiv:1612.00323*, 2016.

[25] Z. Liu and I. Weber. Is twitter a public sphere for online conflicts? a cross-ideological and cross-hierarchical look. In *SocInfo*, pages 336–347, 2014.

[26] A. Mowshowitz and A. Kawaguchi. Measuring search engine bias. *Information Processing & Management*, 41(5):1193–1205, 2005.

[27] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[28] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. In *SSNR Preprint*, 2017.

[29] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In google we trust: Users decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

[30] F. Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[31] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582–638, 2014.

[32] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 2014.

[33] J. L. Skeem and C. T. Lowenkamp. Risk, race, and recidivism: predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

[34] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data, responsibly: Fairness, neutrality and transparency in data analysis. In *EDBT*, 2016.

[35] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[36] L. Vaughan and M. Thelwall. Search engine coverage bias: evidence and possible causes. *Information processing & management*, 40(4):693–707, 2004.

[37] I. Weber, V. R. K. Garimella, and A. Batayneh. Secular vs. islamist polarization in egypt on twitter. In *ASONAM*, pages 290–297, 2013.

[38] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang. Quantifying political leaning from tweets and retweets. *ICWSM*, 13:640–649, 2013.

[39] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *SSDM*, pages 22:1–22:6, 2017.

[40] M. B. Zafar, K. P. Gummadi, and C. Danescu-Niculescu-Mizil. Message impartiality in social media discussions. In *ICWSM*, pages 466–475, 2016.

[41] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2015.

[42] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017.

[43] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. A. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *CIKM*, 2017.

[44] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457, 2017.