Addressing Social Bias in Information Retrieval

Jahna Otterbacher[1,2(✉)]

[1] Open University of Cyprus, Nicosia, Cyprus
jahna.otterbacher@ouc.ac.cy
[2] Research Centre on Interactive Media Smart Systems and Emerging Technologies, Nicosia, Cyprus

Abstract. Journalists and researchers alike have claimed that IR systems are socially biased, returning results to users that perpetuate gender and racial stereotypes. In this position paper, I argue that IR researchers and in particular, evaluation communities such as CLEF, can and should address such concerns. Using as a guide the *Principles for Algorithmic Transparency and Accountability* recently put forward by the Association for Computing Machinery, I provide examples of techniques for examining social biases in IR systems and in particular, search engines.

Keywords: Social biases · Ranking algorithms · Crowdsourcing

1 Introduction

The social impact of algorithmic systems – including information retrieval (IR) systems – is being discussed extensively in the media. Eye-catching titles often convey sweeping claims such as "AI learns to be sexist and racist[1]" or "Biased algorithms are everywhere, and no one seems to care[2]." IR systems – particularly search engines – are often the target of more specific accusations of *social bias*, for instance: "Google has a striking history of bias against black girls[3]" or "Google's algorithm shows prestigious job ads to men, but not to women[4]."

At the same time, there is growing recognition from the scientific community that algorithms – especially those that are opaque to the user – can and do bring about negative consequences, some of which are systematic. Several communities have started initiatives to promote the alignment of algorithmic systems with

Partially supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 739578.

[1] http://www.newsweek.com/2017/12/22/ai-learns-sexist-racist-742767.html.
[2] https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/.
[3] http://time.com/5209144/google-search-engine-algorithm-bias-racism/.
[4] https://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html.

N. Ferro et al. (Eds.): CLEF 2018, LNCS 11018, pp. 1–7, 2018.
https://doi.org/10.1007/978-3-319-98932-7_11

human values. For example, in April 2016, the IEEE launched its Global Initiative on Ethics of Autonomous and Intelligent Systems, to demonstrate that taking into consideration the human and ethical aspects of design can have a positive impact on innovation[5]. The deliverables of the initiative include a collaboratively produced document, *Ethically Aligned Design*, which summarizes input from hundreds of stakeholders, as well as a set of standards projects (e.g., IEEE P7003 Standard for Algorithmic Bias Considerations working group[6]).

Another recent development is the Association for Computing Machinery's *Statement on Algorithmic Transparency and Accountability*, which has been approved by both the ACM U.S. Public Policy Council and the Europe Policy Committee[7]. The statement notes that many algorithmic processes are opaque and that the reasons for this may vary. For instance, it is more often than not difficult to interpret results from models induced by new machine learning techniques such as deep learning (i.e., there are significant technical challenges for transparency). In addition to this, there are social and economic challenges for achieving algorithmic transparency, such as the need for developers/owners of such processes to protect trade secrets, or even the privacy concerns of users.

The ACM Statement puts forward a set of seven principles, which can be used by system developers and owners for promoting algorithmic transparency and accountability. The principles include Data Provenance (i.e., scrutinizing the processes by which training data is generated) and Validation and Testing (i.e., routine assessments as to whether an algorithm's outputs result in discriminatory harm, and making the assessment results public). Inspired by the principles, as well as a recent presentation by Margaret Mitchell[8], Fig. 1 presents three known sources of human biases (data, development process, user behaviors) in a typical pipeline involving an algorithmic system and a human end-user, as well as possible opportunities to promote transparency (i.e., "interventions").

I shall return to these principles, and how they might be applied in the context of IR system development and evaluation. Next, I provide a working definition for the term *social bias* and briefly summarize some recent work that has revealed social biases in IR systems and in particular, *image search engines*.

1.1 Social Biases in Search Engines

While it has long been accepted that information systems bring a slant in their presentation of information to users, there is a need to determine if and when a system is *biased* and when intervention is necessary. Writing long before the age of Big Data, Friedman and Nissenbaum [2] outlined two conditions under which a system could be considered biased: (1) its results are slanted in unfair

[5] https://ethicsinaction.ieee.org/.

[6] https://standards.ieee.org/develop/project/7003.html.

[7] https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf.

[8] https://www.slideshare.net/SessionsEvents/margaret-mitchell-senior-research-scientist-google-at-mlonf-seattle-2017.

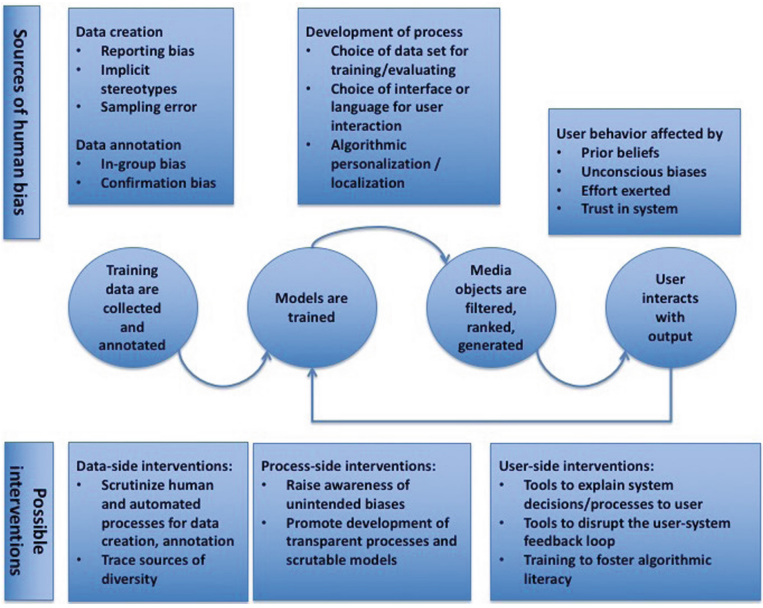

Fig. 1. Sources of human bias in algorithmic systems and interventions.

discrimination against particular persons or groups, and (2) the observed discrimination is systematic within the system. Indeed, over the past years, many researchers have found that search engines, through the result sets they present to users, tend to reinforce a view of the social world that aligns with the status quo.

For instance, a study by Kay, Matuszek and Munson [3] found that Google image search results present a gendered view of the professions, amplifying existing stereotypes. For queries related to professions (e.g., doctor vs. nurse, engineer vs. teacher), they showed that the engine systematically returned more/fewer images of stereotype congruent/incongruent individuals, as compared to U.S. labor statistics. This work also demonstrated the power of search on users' perceptions; when participants were shown gender-biased search results and were asked to estimate the corresponding labor statistic, this skewed their estimates of the distribution of women/men in a particular profession.

My colleagues and I [6] considered the gendering of image search results, although in the context of the Microsoft Bing search engine, to which we submitted queries involving character traits (e.g., intelligent person vs. emotional person). Grounding our study in social psychology theory surrounding person perception, we found that Bing more often associated images of women with warm traits (e.g., kind, emotional) whereas images of men were typically featured in results sets on searches for agentic traits (e.g., assertive, intelligent). In addition, we found a backlash effect, in term of the nature of the images retrieved, which penalized stereotype-incongruent individuals (i.e., agentic women).

Research over the past years has demonstrated that search engine results can and do shape the public's opinion (e.g., [1]). Furthermore, it has long been known that users place great trust in the results of search engines, and rarely look beyond the most highly-ranked results [8]. For these reasons, it is important to promote greater transparency in search algorithms.

2 Interventions to Address Social Bias

In this section, I briefly present some approaches that my colleagues and I have used in our recent work, in order to explore various sources of social biases in search engines. Guided by Fig. 1, I consider Data Provenance, Validation and Testing procedures as well as the role of the user's own biases in perpetuating social stereotypes in image search engine results.

2.1 Data Provenance

The fifth principle in the ACM Statement is that of *Data Provenance*. Researchers are called to scrutinize the means by which training data sets are built. As an example of such "data-side interventions," as shown in Fig. 1, I have conducted a series of studies, to better understand the extent to which social stereotypes are conveyed in crowdsourced descriptions of people-related media. In [4], I analyzed the "Small ESP Game Dataset"[9], consisting of images collected from the Web, and labeled through the well-known ESP game [9], which asks players to describe images in their own words. The analysis revealed systematic gender-based differences in the way that people-images were described by ESP players. Specifically, images of women were labeled more frequently with subjective adjectives, as compared to images of men. Furthermore, images depicting women received more labels related to appearance, in contrast to images of men, which more often had labels related to the person's occupation.

While the above study showed evidence that crowdsourced image annotations can perpetuate gender stereotypes (i.e., that women should be attractive and men career-oriented), it did not examine how worker demographics are correlated to the process, nor could it control the parameters of the labeling task or the nature of the content of the images. Therefore, in [5], I conducted a controlled experiment with workers at Amazon Mechanical Turk, who were located in the U.S. and who identified as being Caucasian, native English speakers. In a between-subjects study, I asked workers to label images of professionals depicted in similar scenarios. However, the gender and race of the depicted person were manipulated. Among other findings, there was evidence of systematic differences in the language used to describe black versus white professionals, a phenomenon known as *linguistic bias*. Interestingly, the biases were more pronounced in the image descriptions produced by men workers, as compared to women workers, demonstrating that, depending on the nature of the task and the intended use of the resulting data, the use of anonymous crowdworkers can be problematic.

[9] http://www.cs.cmu.edu/~biglou/resources.

2.2 Validation and Testing

The seventh principle in the ACM Statement describes the need to "routinely perform tests to assess and determine whether the model generates discriminatory harm." In the context of our study of the Bing search engine and the perpetuation of gender stereotypes based on character traits [6], we developed a process for post-processing the first 1.000 images retrieved for a given query. More specifically, we used machine vision to process the images retrieved, in order to determine the gender distribution of the depicted individuals in the results set. This allowed us to study how a wide range of character traits are gendered by Bing, which would not be possible if we relied on manual analysis of the images. In addition, we were able to compare the gendering of traits across search markets, comparing four large anglophone markets (U.K., U.S., India and South Africa). It must be noted that our testing procedure for Bing's output relied on another algorithm for processing the images. Therefore, we first tested the performance of the procedure, comparing its accuracy on inferring the gender(s) of the depicted person(s) against that of human analysts.

2.3 The Role of the User

The ACM Statement's first principle is *Awareness* – that all stakeholders, from system designers and engineers, to the end users, should be aware of possible biases of the system as well as their potential harms. To this end, in [7], my colleagues and I explored users' awareness of gender bias in image search results sets. We hypothesized that users who are more sexist, would be less likely to indicate that a heavily gender-imbalanced set of images is "subjective" as compared to less sexist users. We again conducted an experiment with crowdworkers, this time at the Crowdflower platform. Without priming workers on the topic of our experiment, we first showed them a set of images, which we knew to be either heavily gender-skewed toward depicting men or women, or gender balanced. We then asked them to describe to us what they saw, and specifically, what keywords best describe the set of images. Next, we informed them that the images were in fact retrieved from a search engine using the given query, asking them to assess the objectivity/subjectivity of results set as a 7-point item. Finally, the workers were asked to take a standardized psychological test to assess their level of sexism. The correlation between sexism and the evaluation of gender bias in the image set results was as expected. However, an interesting finding was that more/less sexist users described the images in a very similar manner, suggesting few differences in how the images were perceived. The study demonstrates how studying users' prejudices and beliefs can help us better understand how they engage with and evaluate search technologies.

3 Conclusion

I have argued in favor of IR researchers and the evaluation community more broadly, addressing the issue of social biases in information retrieval systems. To this end, I have presented examples from my recent work, which considers the perpetuation of social stereotypes in three areas (i.e., potential sources of biases in the system development pipeline): training data sets, search result sets, and users' own biases.

More concretely, the CLEF community could consider the introduction of new tracks or labs to tackle social bias in IR. One could envision data-focused tasks, such as the development of metrics to audit search benchmark data and even techniques to prevent bias in evaluation corpora. Similarly, in the spirit of the ACM Statement's seventh principle, the community might discuss what a standardized "routine assessment" of algorithmic output in various IR tasks should look like. In conclusion, while many may argue that the issue of social biases is beyond the scope of an IR researcher or developer's work, given the recent attention to the social and ethical dimensions of algorithmic and intelligent systems, I hope that this paper will stimulate fruitful discussion amongst those who aim to effectively evaluate IR system performance in a holistic manner.

References

1. Epstein, R., Robertson, R.E.: The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proc. Nat. Acad. Sci. **112**(33), E4512–E4521 (2015). https://doi.org/10.1073/pnas.1419828112. http://www.pnas.org/content/112/33/E4512
2. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. (TOIS) **14**(3), 330–347 (1996)
3. Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3819–3828. ACM (2015)
4. Otterbacher, J.: Crowdsourcing stereotypes: linguistic bias in metadata generated via gwap. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, pp. 1955–1964. ACM, New York (2015). https://doi.org/10.1145/2702123.2702151
5. Otterbacher, J.: Social cues, social biases: stereotypes in annotations on people images. In: Proceedings of the Sixth AAAI Conference on Human Computation and Crowdsourcing (HCOMP-2018). AAAI Press, Palo Alto (2018)
6. Otterbacher, J., Bates, J., Clough, P.: Competent men and warm women: gender stereotypes and backlash in image search results. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI 2017, pp. 6620–6631. ACM, New York (2017). https://doi.org/10.1145/3025453.3025727
7. Otterbacher, J., Checco, A., Demartini, G., Clough, P.: Investigating user perception of gender bias in image search: the role of sexism. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2018). ACM Press, New York (2018)

8. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In google we trust: users decisions on rank, position, and relevance. J. Comput. Med. Commun. **12**(3), 801–823 (2007). https://doi.org/10.1111/j.1083-6101.2007.00351.x
9. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)