Gender Fairness in Information Retrieval Systems

Amin Bigdeli
abigdeli@ryerson.ca
Ryerson University, Canada

Negar Arabzadeh
narabzad@uwaterloo.ca
University of Waterloo, Canada

Shirin SeyedSalehi
shirin.seyedsalehi@ryerson.ca
Ryerson University, Canada

Morteza Zihayat
mzihayat@ryerson.ca
Ryerson University, Canada

Ebrahim Bagheri
bagheri@ryerson.ca
Ryerson University, Canada

ABSTRACT

Recent studies have shown that it is possible for stereotypical gender biases to find their way into representational and algorithmic aspects of retrieval methods; hence, exhibit themselves in retrieval outcomes. In this tutorial, we inform the audience of various studies that have systematically reported the presence of stereotypical gender biases in Information Retrieval (IR) systems. We further classify existing work on gender biases in IR systems as being related to (1) relevance judgement datasets, (2) structure of retrieval methods, and (3) representations learnt for queries and documents. We present how each of these components can be impacted by or cause intensified biases during retrieval. Based on these identified issues, we then present a collection of approaches from the literature that have discussed how such biases can be measured, controlled, or mitigated. Additionally, we introduce publicly available datasets that are often used for investigating gender biases in IR systems as well as evaluation methodology adopted for determining the utility of gender bias mitigation strategies.

CCS CONCEPTS

• **Information systems → Presentation of retrieval results**.

KEYWORDS

Bias, Fairness, Information Retrieval Systems

ACM Reference Format:
Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. Gender Fairness in Information Retrieval Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3477495.3532680

1 MOTIVATION AND OVERVIEW

There have been both qualitative and quantitative studies that have effectively shown that societal biases have become prevalent in various Natural Language Processing (NLP) and Information Retrieval (IR) techniques, models, and datasets [2, 3, 9, 10, 12, 14, 16, 17, 23, 28, 34, 36]. Given these tools are often deployed at scale, such biases have the potential to directly impact the lives of many people. More specifically within the context of information retrieval, biased retrieval methods can exacerbate biases by exposing users to a set of biased documents in response to user queries. Such biases can have a potentially harmful impact on the users' judgments when exposed to unfair and biased search results. This is concerning especially given the fact that not only do a large number of search engine users heavily rely on retrieval systems on a daily basis but also due to the fact that search results often constitute a major component of important practical systems such as recommendation systems, question answering systems, intelligent assistants, to name a few. Researchers such as Draws et al. [13] have recently shown that when search results are biased, the users who are exposed to the biases results will tend to favor the biased viewpoint. This aligns very well with several forms of cognitive bias identified by Azzopardi [1] including *Availability bias*, which points to user biases towards content that are more easily accessible, and *Anchoring Bias* that reports that users are more likely to focus on the first piece of information that they receive.

Thus, it is important to systematically control the degree of biases that are exhibited by such retrieval systems to avoid their detrimental effects on the users' beliefs and decisions. In order to systematically address such biases, various researchers have proposed methods that can help control and/or mitigate biases, such as gender biases, in information retrieval systems. In this tutorial, we will provide a classification of existing work in the literature [2, 5–9, 11, 15, 18–20, 22, 30, 32, 34, 37–39] and introduce the state-of-the-art methods that are available for managing gender biases within IR systems. The structure of this tutorial can be summarized as follows:

(1) The tutorial will present concrete evidence, using real-world examples of cases where gender biases are introduced and intensified in natural language processing and information retrieval systems;
(2) The tutorial will draw inspiration from and provide adequate contextual information from experience reports and methodological work in natural language processing that have already explored gender biases [33, 34];
(3) A systematic classification of possible sources for gender biases will be presented and details of how biases can be transferred from these sources will be provided. These sources include relevance judgement collections, ranker characteristics, objective functions, and neural embeddings, to name a few.
(4) The tutorial will review existing methods that have attempted to control or mitigate gender biases and will also provide an in-depth treatment of the retrieval effectiveness-bias tradeoff. This

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3532680

tradeoff is concerned with the right balance between maximizing retrieval effectiveness and minimizing gender bias.
(5) A clear description of evaluation methodology, datasets, and metrics that have been used in the literature for investigating gender bias will be provided.

Our tutorial will build on four recent invited talks that we have delivered at Microsoft Research, Center for Intelligent Information Retrieval at UMass Amherst, the keynote talk at the BIAS workshop at ECIR 2022, and Radboud University. The central focus of these talks have been on methods for controlling and mitigating gender biases, which can broadly be classified as follows:

(1) *Relevance Judgement Collections*: Relevance judgment documents are often considered as gold standard benchmark datasets used for training and evaluating ranking models. As such, they can significantly impact the characteristics of IR systems. Researchers have introduced methodological processes for studying possible traces of gender bias in relevance judgment collections [8, 26], and show that stereotypical gender biases can be observable in these collections, which are capable of making their way into the algorithmic aspects of ranking models that are trained and evaluated based on them. We will also introduce those approaches that have been introduced in the literature for de-biasing relevance judgment collections. We will report on the findings from these studies that when neural ranking models are trained based on de-biased relevance judgments, the level of gender biases may be reduced while retrieval effectiveness is maintained.
(2) *Neural Representations*: Neural embeddings have been widely adopted in IR systems for different tasks such as document retrieval [4]. Since neural representations have often been pretrained on large corpora, they may have picked up existing gender biases. Many research works [5, 9, 11, 34, 38, 39] have investigated gender biases within these neural representations, and have proposed methods for mitigating the levels of bias using different approaches such as data augmentation and embeddings de-biasing techniques [9, 16, 24, 29]. We will cover how such techniques can be adopted in practice to manage gender biases.
(3) *Query Representation*: The query submitted by the user can itself be highly influential on the retrieved list of documents. For instance, Kulshrestha et al. [21] examine the impact of such biases in the context of political search queries. Therefore, we will report on studies that explore the gendered nature of search queries [35], as well as those that present query reformulation mechanisms that attempt to revise an initial query in a way that will lead to a less biased list of documents while maintaining (or possibly increasing) retrieval effectiveness [7].
(4) *Retrieval Methods*: The emergence of neural architectures has resulted in a dramatic shift from traditional methods to neural ranking models that leverage neural embeddings for finding relevant documents [27]. Recent studies show that neural-based retrieval methods are capable of intensifying the level of gender biases within the retrieved list of documents [15, 31]. Therefore, it is important to manage the level of gender bias at the ranker level. Researchers have already looked into how neural rankers can be made less biased (or in other terms more fair) through approaches such as introducing bias-aware loss functions or bias-aware negative sampling strategies. In the tutorial, we will cover various existing work in this space. For instance, we will introduce methods such as AdvBert [30] that leverages adversarial components within the BERT reranker loss function for decreasing the level of bias in neural rankers. We will also introduce the bias-aware neural ranker [32], which incorporates a notion of gender bias and hence control how bias is expressed in the retrieved documents. We will also cover bias-aware negative sampling strategy that consider the degrees of gender bias when sampling documents to be used for training neural rankers [6].

We highlight that this tutorial will build on but significantly expand the scope of our talks by providing comprehensive information about evaluation metrics, available datasets, and bias measurement techniques. We will discuss the limitations of existing work from both technical and conceptual perspectives. For instance, we will at least highlight the following two limitations: (1) existing work in the literature have focused on the notion of sex as a binary construct and assumed that search queries and results can be analyzed from their association with the male or female gender. This is a major limitation that needs to be addressed in future work; and (2) Most existing work assume that gender bias can be measured based on the frequency of gendered terms. This overlooks the complexity associated with the stance and position of documents with regards to different gender identities in favor of simplifying computation.

2 OBJECTIVES

The objectives of this tutorial can be enumerated as follows:

(1) Show the presence of gender biases in information retrieval systems and large scale corpora relevance judgments;
(2) Introduce bias measurement metrics used for calculating the level of gender biases within the retrieved list of documents;
(3) Present datasets used for investigating gender biases in information retrieval results;
(4) Introduce de-biasing methods for reducing the level of bias in relevance judgment datasets;
(5) Present existing methods for reducing the bias through revising the input query;
(6) Describe existing methods for mitigating the level of bias within neural ranking models.
(7) Highlight important theoretical and conceptual limitations of existing work when dealing with the concept of gender.

The aforementioned topics will give participants a thorough understanding of existing datasets and bias measurement metrics used for investigating gender biases within information retrieval results. Besides, they become familiar with methods used for reducing gender biases within IR systems. As a result, they can take advantage of these techniques to release models that are aware of gender biases and expose users to a less biased list of documents without being worried about the retrieval effectiveness of their model.

To facilitate future development in this research area, we introduce all the required materials including datasets used for investigating bias, performance and bias measurement metrics, and methods used for reducing bias.

In addition, these topics can be beneficial for researchers who are conducting research in a similar area in terms of applying introduced de-biasing methods for other *types of societal biases* and can serve as a useful starting point.

3 FORMAT AND SCHEDULE

This tutorial presents a comprehensive survey about the exploration and mitigation of gender biases in information retrieval systems. In particular, this tutorial covers the following sections:

Session A [20 Minutes]: Background and Introduction to the Topic of Gender Biases in IR. The tutorial begins with a session about the basics of the information retrieval systems as well as the datasets that will be used through the tutorial. Followed by that, we show the footprints of gender biases in the IR systems as well as NLP and introduce the bias measurement methods that will be used for measuring the level of biases within a list of documents.

Session B [50 Minutes]: Exploration and Mitigation of Gender Biases in IR Relevance Judgment Datasets. In this session, we first investigate the existence of stereotypical gender biases within the relevance judgement datasets through a methodological approach. After presenting the findings, we proceed further and introduce systematic de-biasing methods for balancing the relevance judgment datasets and reducing the level of gender biases within these documents. We also show that when neural rankers are trained based on de-biased relevance judgments, the level of biases within their ranked list of documents decreases significantly while utility remains unchanged.

Session C [70 Minutes]: Reducing Gender Biases within IR Retrieval Methods. In this session, we review existing methods for reducing gender biases through different classes of retrieval methods, namely, term-frequency-based methods as well as neural ranking models. We first start by introducing bias-aware query reformulation methods that revise the initial query in a way that would lead to a less biased ranked list of documents. Subsequently, we introduce state-of-the-art methods for mitigating the level of gender biases in neural ranking models. We will also demonstrate that leveraging these methods allows for maintaining utility and at the same time mitigating the level of gender biases considerably. Finally, we demonstrate how each of the proposed methods can be applied for other societal attributes other than gender.

Session D [20 Minutes]: Limitations and Future Work. This session will discuss major theoretical and conceptual limitations of existing work and will present avenues for future work.

4 RELEVANCE TO INFORMATION RETRIEVAL COMMUNITY

Fairness and ethical issues surrounding the practice of IR has become a major topic of concern among IR researchers [7–9, 30, 31, 38, 39]. One of the main reasons is that the existence of gender biases in the IR systems can influence an individual's judgments, leading to unfair treatments and outcomes. In an ideal world, the expectation from IR systems is to be fair towards different gender identities and avoid reflecting unfair prejudices that may exist within society. We hope that our work contributes to the growing body of knowledge in this area, and helps the IR community to become familiar with the datasets, metrics, and methods that can be used for reducing the level of such biases in retrieval methods. It is worth mentioning that there have been many attempts by industrial entities to address biases from a practical sense. For instance, we can point to the investigation of fairness in the practice of neural-based models by Microsoft, the responsible machine learning initiative at Twitter, which tackles gender and racial biases, or the PAIR group at Google Brain that explores responsible AI in different Google systems.

We note that while there have been similar tutorials related to investigating fairness issues in retrieval systems in other venues, this tutorial distinguishes itself by focusing on proposing systematic and well-validated methods for reducing gender biases in retrieval results. The following tutorials can be considered complementary and synergistic to the theme of our proposed tutorial:

(1) *Addressing Bias and Fairness in Search Systems* by Ruoyuan Gao and Chirag Shah at SIGIR 2021. This tutorial focuses on introducing the issue of bias in data, algorithms, and search process. It also presents some concepts about fairness, diversity, and bias and the ways for creating a fairer retrieval system.
(2) *Fairness of Machine Learning in Recommender Systems* by Yunqi Li, Yingqiang Ge, Yongfeng Zhang at CIKM 2021. This tutorial introduces fairness definitions as well as evaluation metrics and describes the fairness topics in recommender systems.
(3) *Fair Graph Mining* by Jian Kang, Hanghang Tong at CIKM 2021. The purpose of this tutorial is to introduce state-of-the-art techniques for increasing fairness on graph mining and describe challenges as well as future directions.

Our goal in this tutorial is to provide comprehensive knowledge about the methods and techniques that can be used for reducing gender biases in information retrieval systems, while past tutorials are not related to retrieval tasks.

5 INTENDED AUDIENCE

The target audience for this tutorial will be those who have interest in IR methods especially neural ranking models and well-known datasets. The tutorial will provide an overview of some of the IR concepts and components for those who are new to the field of IR. As such, sufficient details will be provided as appropriate so that the content will be accessible and understandable to those who only have a basic understanding of IR principles. This tutorial will only assume that the audiences is familiar with different topics included in an undergraduate IR course such as those covered in [25].

6 PRESENTERS

Amin Bigdeli is a Data Scientist at Warranty Life and a Research Associate at Ryerson University. His research work focuses on issues of fairness in information retrieval systems. Amin has published multiple research papers in this area in top information retrieval venues such as SIGIR, CIKM, EDBT, and ECIR.

Negar Arabzadeh is a Research Scientist at Spotify Research working on PodCast Retrieval Evalaution. She is also completing her Ph.D. studies at the University of Waterloo. Her research is aligned with Ad hoc Retrieval and Conversational search in IR and NLP domains. Negar has published and presented relevant papers in prestigious conferences and journals such as SIGIR, CIKM, ECIR, and IP&M and recently interned at Microsoft Research.

Shirin Seyedsalehi is a Ph.D. student at Ryerson University. Her research so far is focused on fairness in Information Retrieval and Neural Rankers. She has published papers in well known conferences such as SIGIR, CIKM and EDBT. She will join Microsoft Research as for a research intern in Summer 2022.

Morteza Zihayat is an Associate Professor and co-founder of the centre for Digital Enterprise Analytics & Leadership (DEAL) at

Ryerson University. His research concerns user modeling, applied machine learning and bias, debiasing, and fairness in NLP and IR. He has published in various well-respected journals and conferences in Information Retrieval, Machine Learning, and Information Systems such as IEEE TKDE, Information Processing and Management, ACM SIGKDD, SIGIR, ECIR, PKDD, SIAM SDM.

Ebrahim Bagheri is a Professor and the Director for the Laboratory for Systems, Software and Semantics (LS^3) at Ryerson University. He holds a Canada Research Chair (Tier II) in Social Information Retrieval as well as an NSERC Industrial Research Chair in Social Media Analytics. He currently leads the NSERC Program on Responsible AI (http://responsible-ai.ca). He is an Associate Editor for ACM Transactions on Intelligent Systems and Technology (TIST) and Wiley's Computational Intelligence.

7 TYPE OF SUPPORT MATERIALS

As for the supporting materials, we will publicly share a Github repository several weeks prior to the conference so the participants of the tutorial can familiarize themselves with the content. The repository will include a comprehensive slide deck, links to code, models, datasets, and run files. Links to existing papers in the field of biases in information retrieval systems will also be available.

REFERENCES

[1] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 27–37.

[2] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Fourteenth ACM Conference on Recommender Systems*. 2–2.

[4] Ebrahim Bagheri, Faezeh Ensan, and Feras Al-Obeidat. 2018. Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management* 54, 4 (2018), 657–673.

[5] Christine Basta, Marta R Costa-Jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.08783* (2019).

[6] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. A Light-weight Strategy for Restraining Gender Biases in Neural Rankers. In *European Conference on Information Retrieval (ECIR2022)*. Springer.

[7] Amin Bigdeli, Negar Arabzadeh, Shirin Seyersalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. On the Orthogonality of Bias and Utility in Ad hoc Retrieval. In *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[8] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets. In *European Conference on Information Retrieval*. Springer, 216–224.

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[10] Shikha Bordia and Samuel R Bowman. 2019. Identifying and reducing gender bias in word-level language models. *arXiv preprint arXiv:1904.03035* (2019).

[11] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*. PMLR, 803–811.

[12] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[13] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. *This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics.* Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3404835.3462851

[14] Michael D Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness in Information Access Systems. *arXiv preprint arXiv:2105.05779* (2021).

[15] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[16] Joel Escudé Font and Marta R Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).

[17] Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2020. Bias in Conversational Search: The Double-Edged Sword of the Personalized Knowledge Graph. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 133–136.

[18] Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the Characteristics of Ranking-based Gender Bias Measures. In *WebSci'22* (2022-03-30) *(The 14th International ACM Conference on Web Science in 2022 (WebSci'22), 26 – 29, June, 2022, Universitat Pompeu Fabra, Barcelona, Spain)*.

[19] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. *arXiv preprint arXiv:2201.07754* (2022).

[20] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements? *arXiv preprint arXiv:2203.01731* (2022).

[21] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 417–432.

[22] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486* (2019).

[23] Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint arXiv:2009.13028* (2020).

[24] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, 189–202.

[25] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

[26] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

[27] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[28] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. 2021. FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, Vol. 53. ACM New York, NY, USA, 20–43.

[29] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *arXiv preprint arXiv:1908.02810* (2019).

[30] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation for BERT Rankers. *arXiv preprint arXiv:2104.13640* (2021).

[31] Navid Rekabsaz and Markus Schedl. 2020. Do Neural Ranking Models Intensify Gender Bias?. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2065–2068.

[32] Shirin SeyedSalehi, Amin Bigdeli, Negar Arabzadeh, Bhaskar Mitra, Morteza Zihayat, and Ebrahim Bagheri. 2022. Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases. In *Extending Database Technology (EDBT2022)*. Springer.

[33] Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint arXiv:2112.14168* (2021).

[34] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[35] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search. *arXiv preprint arXiv:2109.05433* (2021).

[36] Zekun Yang and Juan Feng. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9434–9441.

[37] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. *arXiv preprint arXiv:2005.00699* (2020).

[38] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

[39] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).