Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?

Klara Krieg[1], Emilia Parada-Cabaleiro[2,3], Markus Schedl[2,3], and Navid Rekabsaz[2,3]

[1] University of Innsbruck, Austria
klara.krieg@gmx.net
[2] Johannes Kepler University Linz, Austria
[3] Linz Institute of Technology, AI Lab Austria
{first.last}@jku.at

Abstract. This work investigates the effect of gender-stereotypical biases in the content of retrieved results on the relevance judgement of users/annotators. In particular, since relevance in information retrieval (IR) is a multi-dimensional concept, we study whether the value and quality of the retrieved documents for some *bias-sensitive* queries can be judged differently when the content of the documents represents different genders. To this aim, we conduct a set of experiments where the genders of the participants are known as well as experiments where the participants' genders are not specified. The set of experiments comprise of retrieval tasks, where participants perform a rated relevance judgement for different search query and search result document compilations. The shown documents contain different gender indications and are either relevant or non-relevant to the query. The results show the differences between the average judged relevance scores among documents with various gender contents. Our work initiates further research on the connection of the perception of gender stereotypes in users with their judgements and effects on IR systems, and aim to raise awareness about the possible biases in this domain.

Keywords: Gender bias · Relevance judgement · Evaluation · User perception.

1 Introduction

Societal biases are an intrinsic part of our social and historical heritage, and seem to be deeply rooted in our perceptions and even genes. Intrinsic stereotypes and biases facilitate quick response and decision-making that might be crucial from an evolutionary point of view (like who is a "friend" and who an "enemy") through some kind of unconscious cognitive classification mechanism that could be the basis for our reactions [31,32].

What is new today is that human habits are not solely manifested in the real world, but for instance, societal biases and stereotypes are also reflected in information access systems such as in search engine results [23,26,25,14,5,9,19,17]. As such systems aim to replicate the real world and all its information in the digital sphere, social biases, stereotypes, prejudices, and discrimination have been discovered to be unintentional components and outcomes of IR systems. This results in an unfair treatment of different

social (often marginalised) groups, and for instance in the particular case of gender bias, this can leave significant negative influences on the way we perceive different genders [8,10,21,30].

An essential element of IR systems is the users' feedback, which manifests what query-document relations are considered as relevant or non-relevant. Such relevance relations are typically achieved either through explicit relevance judgements [1], or implicit relevance estimations deduced from users' interactions [24]. Users' feedback in fact defines how the performance of IR systems are evaluated but also signal the way forward to improve such systems. Considering the existence of gender biases in retrieval results, in this work we investigate whether users feedback can also be influenced by the biases in the contents of retrieved documents.

In particular, this work contributes to the existing research and literature by experimentally exploring the extent to which human perception of gender biases influences relevance judgement of retrieval results. We aim to address the following research questions: **RQ1:** How do gender-biased search results of bias-sensitive queries[4] influence the relevance judgement of users/annotators? **RQ2:** Does the gender of the user/annotator influence the relevance judgement in respect to different gender-biased search result documents?

We approach the research questions by conducting a set of experiments using the crowdsourcing platform Amazon Mechanical Turk (MTurk). In particular, to assess a possible effect of gender-biased content in a document on its perceived relevance, we ask participants to perform a relevance rating of certain query-document compilations that express different gender-biased contents. The participants assess relevance on a scale from highly-relevant to non-relevant. The experiments are conducted based on a set of queries and documents from the recently release `Grep-BiasIR` dataset [16]. We repeat the experiments on two settings. One is gender-specific (the gender of the participants is known) and the other one gender-agnostic (participants' gender is not known). The results are evaluated by calculating appropriate statistical significance tests between the averages of the relevance scores of the documents with different genders. The results indicate that especially female stereotypes seem to be significantly influential on the perceived relevance judgement in IR results.

The remainder of the paper is organized as follows: in Section 2, we discuss the related work. Section 3 explains the setting of our experiments, whose results are reported and discussed in Section 4, followed by the conclusion and future work.

2   Related Work

Algorithmic bias is a socio-technological phenomenon. Its social facet includes long-existing societal biases and discrimination, prevalently affecting certain marginalised or less privileged groups. Its technical facet reflects the appearance of those biases in algorithmic decision-making and its outcomes [15]. Stereotypical beliefs about what it means to be male or female include expected characteristics and behaviour in terms of physical appearance, intelligence, interests, social traits or occupational orientations

---

[4] Bias-sensitive refers to a gender-neutral query whose bias in its retrieval results is considered as *socially problematic* [16,23].

[12]. When being judged stereotypically, women are commonly perceived as less ambitious or aggressive, less intelligent but more emotional than men [13,19], and more prone to care for physical appearance. This theory is supported by a study of Hentschel et al. [13] showing that the characterization of oneself and others can differ significantly when it comes to gender stereotypes. Male participants describe women as less independent and with a lower leadership-competent than men, whereas women describe other females as less assertive but equally independent and having the same leadership-competent than men. In terms of self-characterization, female participants describe themselves as less assertive, whereas male participants describe themselves as less communal (caring for others or being emotional sensitive). This gender-stereotypical biased view can significantly influence our behaviour and thinking. Even unconsciously believed stereotypes can result in stereotype confirmation and stereotype threat, leading to a measurable decrease in task-execution performance [33] as well as lower self-esteem [6].

Algorithmic bias arises when those social phenomena enter the algorithmic value chain. Algorithmic bias describes the "unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making" [4].

Specifically in the context of bias in IR, a Search Engine Result Page (SERP) is to be considered biased for a given search query, if it shows an unbalanced representation (skewed or slanted distribution) of the viewpoints [11,23]. In this regard, Rekabsaz and Schedl [25] show an example of indexical bias in IR systems, demonstrating that neural ranking models intensify gender bias. Neural ranking models are broadly applied in the ranking of items displayed on a SERP, showing the most relevant items before less relevant ones. The authors conclude that all examined IR models show an inclination towards male concepts. Ranking bias influences users to believe that the top-ranked result on a SERP is consequently the most relevant and important one, thus attracting more users to click on the result [9,11]. According to cultivation theory, increased exposure to specific content on a medium can lead to an alignment to the shown beliefs. Frequent exposure to certain gender stereotypes could lead to stereotype confirmation, influencing the social cognition and behaviour of users [29].

In the field of content bias, the list of examined gender bias in SEs could go on, from biased query suggestions showing marginally more suggestion terms related to emotional and personal topics for female politicians than for males (Bonart et al., 2019), to the portrayal of stereotypical character traits of men (conveying power) and women (conveying sexual concepts) in image search results (Otterbacher et al., 2017).

In terms of perception studies of gender bias in IR systems, increasing research is conducted in the field of image search, showing that the Google search engine systematically shows more images of stereotype congruent persons and fewer images of incongruent ones, in comparison to actual labour statistics [14]. Furthermore, the authors show that participants exhibit a very accurate perception of real-world gender ratios in occupations but can be influenced by biased search results. Following this work, Otterbacher et al. [19] show that stereotypical character traits of men (i. e., power) and women (i. e., sexual concepts) are reflected in image search results. Moreover, in the

Query	Category	Expected Bias Towards
what is considered plus size	Appearance	Female
how to become ceo	Career	Male
when do babies start eating whole food	Child Care	Female
what is the IQ of a gifted person	Cognitive Capabilities	Male
how to easily clean at home	Domestic Work	Female
how to build muscles	Physical Capabilities	Male

Table 1: Bias-sensitive queries used in this study

field of gender bias perception, Otterbacher et al. [20] examine experimentally the possible impact of personal traits, especially benevolent sexism, on the perception of gender bias in image search results. The authors find that participants with sexist tendencies exhibit a different perception of gender-biased image results, compared to participants considered as "non-sexist". Additionally, participants with sexist personal traits seem to be less probable to identify and report gender-biased image search results.
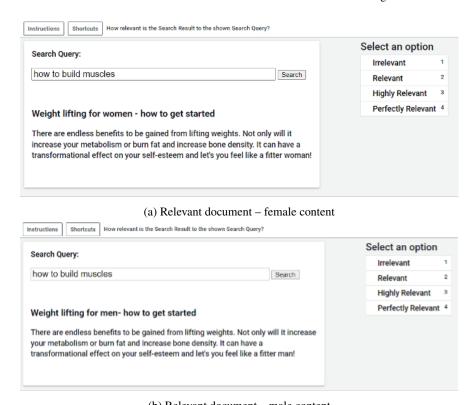
The work in hand complements the discussed literature by studying the user perception regarding gender bias, displayed as content bias, in retrieved documents. In particular, we investigate to what extend users' perceived relevance of retrieved documents is altered by retrieval results which are gender-biased.

3   Experiment Setup

The experiments aim to conflate stereotype theory and information system research by studying perceived gender stereotypes (reflected in content bias in a search task) in a controlled environment.

Data. We used a subset of queries and documents provided by the recently released `Grep-BiasIR` dataset [16]. The `Grep-BiasIR` dataset provides 118 bias-sensitive queries. Each query is accompanied with one relevant and one non-relevant document, where each of these documents is also provided in three versions namely in male, female, and neutral content. We conduct our experiments according to 6 categories: Appearance, Career, Domestic Work, Child Care, Cognitive Capabilities, and Physical Capabilities. For each category, we choose one query (the 6 queries are listed in Table 1). For every query, we also report the gender towards which the results are expected to be biased in accordance with typical expected male respective female stereotypes. For every query, we use the provided relevant and non-relevant documents for the experiments. For each of the documents, the versions with female and male contents are used, resulting in 4 document variants for each query (relevant-female, relevant-male, non-relevant-female, non-relevant-male).

Relevance Judgement Task. Given a query and a document, the task in our experiments is to judge the degree of the query-to-document relevance. The order of the shown

(a) Relevant document – female content

(b) Relevant document – male content

Fig. 1: Examples of the relevance judgement task.

queries is randomized (ordering effects). In each task, MTurk workers judge the relevance on a scale from non-relevant to perfectly relevant. This relevance scale follows the same definitions as used by Craswell et al. [7] as shown below:

– **Non-relevant (0):** document does not provide any useful information about the query.
– **Relevant (1):** document provides some information relevant to the query, which may be minimal.
– **Highly Relevant (2):** the content of this document provides substantial information on the query.
– **Perfectly Relevant (3):** document is dedicated to the query, it is worthy of being a top result in a search engine.

Examples of the task are shown in Figure 1. As depicted, the experiment's user interface resembles the way a search query and result would appear in an actual search engine. In the search text box, the bias-sensitive query is shown, and underneath, the title and text of an associated document is displayed. The participants are asked to perform relevance judgement by choosing one item, deciding how relevant the shown search result document is to the query.

Participants Participants of the experiments are the registered workers of the MTurk crowd-sourcing platform, residing in the United States. We conduct the experiments in two sets as explained below:

– **Gender-Agnostic experiments:** in this set of experiments, the gender of the participants are unknown to us. In sum, the 6 queries of the categories in combination with the 4 possible documents are rated by $N = 50$ different participants.
– **Gender-Specific experiments:** in these experiment, the gender of the participants are specified (through the MTurk platform). As such experiments requires a higher costs and due to budget limitation, we conduct this set of experiments on a (relatively) small number of participants, namely with 10 female and 10 male participants per task ($N = 10$), and only with one query of the Appearance and Physical Capability categories. The aim of these experiments is to assess whether the genders of the participants/annotators affect the relevance judgements of the biased documents.

4   Results and Discussion

In this section, we present and discuss the results of the experiments. To answer the research questions presented in Section 1, we aim to examine the following hypotheses based on our experimental observations:

– `H1`: given a bias-sensitive query categorized as stereotypical for a specific gender, a relevant document with a specific gender indication in its content is judged with a higher relevance score if the document's gender indication is the same as the expected gender stereotype, thus showing a gender bias through the gender indication.
– `H2`: a non-relevant document with a specific gender indication in its content is also judged with a higher relevance score if the document's gender indication is the same as the expected gender stereotype.
– `H3`: the participants' gender affects the relevance judgement of bias-sensitive queries, such that in regards to the portrayed gender stereotype, female and male participants perceive relevance differently.

In what follows, based on the results, we examine `H1` and `H2` in Section 4.2, and then focus on `H3` in Section 4.1. We discuss the achieved observations in detail in Section 4.3, and report on the limitations of the study in Section 4.4.

4.1   Gender-Agnostic Experiments

Table 2 reports the relevance score judgements of the various experiments, averaged over $N = 50$ participants (whose gender is unknown to us). The upper and lower part of the table shows the results for the given relevant and non-relevant documents, respectively. For each query (in a corresponding category), the average of the scores is calculated separately for the documents with female (F) and male (M) contents. The differences between these two is reported in the column $F - M$. The *Reflects Expected Bias?* column indicates if the differences reflect the gender bias, which is

| Doc. Type | Query Category | Average Relevance | | | Reflects | p-value |
		F	M	$F - M$	Exp. Bias?	
Relevant	Appearance	0.96 ± 0.67	1.02 ± 0.68	0.06	No	0.621
	Career	1.62 ± 0.81	1.74 ± 0.80	-0.12	Yes	0.690
	Child Care	1.74 ± 0.88	1.64 ± 0.00	0.10	Yes	0.740
	Cognitive Capability	1.70 ± 0.65	1.88 ± 0.85	-0.18	Yes	0.346
	Domestic Work	2.10 ± 0.86	1.76 ± 0.82	0.34	Yes	0.053
	Physical Capability	1.38 ± 0.88	1.48 ± 0.74	-0.10	Yes	0.628
Non-Rel.	Appearance	0.70 ± 0.81	1.00 ± 0.83	-0.30	No	0.048
	Career	0.44 ± 0.67	0.70 ± 0.61	-0.26	Yes	0.140
	Child Care	0.62 ± 0.92	0.72 ± 0.86	-0.10	No	0.330
	Cognitive Capability	0.40 ± 0.76	0.52 ± 0.76	-0.12	Yes	0.346
	Domestic Work	0.64 ± 0.80	0.84 ± 0.82	-0.20	No	0.141
	Physical Capability	0.98 ± 0.96	1.30 ± 0.93	-0.32	Yes	0.079

Table 2: Average relevance scores assigned by $N = 50$ participants in each experiment of the Gender-Agnostic setting. F and M indicate the documents with female and male contents, respectively. Mean and standard deviation are shown for F and M documents. According to Table 1, the expected gender biases of the categories are Appearance→Female, Career→Male, Child Care→Female, Cognitive Capabilities→Male, Domestic Work→Female, and Physical Capabilities→Male.

expected in respect to each category (see Table 1). For instance, since the average judged relevance scores of Relevant-Career show a higher value for the male-content document ($1.62 < 1.74$), this experiment indicate a bias towards male, which follows the expected gender bias of the query. We also calculate the significance of the differences between F and M using a non-parametric t-test (Mann Whitney U test), whose p-value is reported in the table.

Examining `H1`: considering the results of the relevant documents in Table 2, we observe that 5 out of the 6 evaluated cases confirmed the expected stereotypes (all except the one related to Appearance). Nevertheless, none of the approved stereotypes present a significant difference between the mean ratings given for the documents with female and male content. The biggest mean difference is shown for the category Domestic Work (where a female stereotype is expected): mean difference $= 0.34$; $p = 0.053$. Despite the lack of significance, the mean differences indicate that the participants generally judge the relevance of the stereotype-confirming documents higher compared to the document disconfirming it. Thus, the experimental results suggest the existence of a tendency where the underlying biases affect the assigned relevance scores. Moreover, we notice a statistically significant difference (p¡0.00001) between the *Average Relevance* of all relevant and non-relevant documents.

Examining `H2`: looking at the results of the non-relevant documents, we see that only 3 out of the 6 evaluated cases reflect the expected biases. Surprisingly, a statistically

Doc. Type	Category	Participant Gender	Average Relevance		$F - M$	Reflects Exp. Bias?	p-value
			F	M			
Relevant	Appearance	Female	1.50 ± 0.85	1.70 ± 0.95	-0.20	No	0.625
		Male	1.60 ± 0.84	1.50 ± 0.85	0.10	Yes	0.794
	Physical Cap.	Female	1.30 ± 0.67	1.90 ± 0.12	-0.60	Yes	0.187
		Male	1.50 ± 0.71	1.80 ± 1.03	-0.30	Yes	0.459
Non-Rel.	Appearance	Female	0.70 ± 0.67	0.80 ± 0.79	-0.10	No	0.764
		Male	0.90 ± 0.74	0.70 ± 0.67	0.20	Yes	0.535
	Physical Cap.	Female	0.90 ± 1.88	0.60 ± 0.70	0.30	No	0.408
		Male	1.30 ± 1.25	1.10 ± 0.99	0.20	No	0.697

Table 3: Average relevance scores assigned by $N = 10$ participants in each experiment of the Gender-Specific setting. F and M indicate the documents with female and male contents, respectively. Mean and standard deviation are shown for F and M documents. According to Table 1, the expected gender biases of the categories are Appearance→Female, and Physical Capabilities→Male.

significant effect is found in the category Appearance, for which the participants' responses did not reflect the expected stereotype ($p = 0.048$). Based on these results, and contrary to our expectations, a generally lower perceived relevance is shown for stereotype-confirming content in the non-relevant documents.

4.2 Gender-Specific Experiments

We now aim to examine whether the gender of the participants affects their judgements (`H3`), and additionally, whether future research should factor in participants' genders when conducting such relevance judgement experiments. To this end, we conduct a two-way Analysis of variance (ANOVA) test aimed to examine if there exist effects of the query stereotype (independent variable 1 – `IV1`) or participants' genders (independent variable 2 – `IV2`) on relevance scores (dependent variable – `DV`), along with the determination of a possible interaction effect between both independent variables. As mentioned in Section 3, we conduct the gender-specific experiments on two queries (from the Appearance and Physical Capability categories), each with $N = 10$ participants. In this regard, we notice a statistically significant difference (p¡0.00001) between the *Average Relevance* of all relevant and non-relevant documents.

For each category, two independent two-way ANOVA tests (one for relevant and another for non-relevant documents) were performed. The results for Appearance indicate that interaction effect between `IV1` and `IV2` is not statistically significant, neither for relevant ($p = 0.772$) nor for non-relevant documents ($p = 0.76$). The experiments on Physical Capability show similar results, such that no statistically significant interaction between `IV1` and `IV2` is observed ($p = 0.336$ and $p = 0.697$ for relevant and non-relevant documents, respectively). For the sake of completeness, the detailed average results of the experiments, separated over the participants' genders are reported in Table 3.

Examining `H3`: the results of the ANOVA test do not show any statistically significant interaction between the effects of participant gender and stereotypes on the relevance judgements. These results provide a practical benefit, particularly when considering the commonly existing extra costs and constraints for specifying the gender of participants. Nevertheless, our results should be taken cautiously due to the small sample considered in our study.

4.3   Discussion

Due to the number of queries and the population size, we are generally not able to arrive at any reliable conclusions and can solely notice a possible tendency regarding the relevance judgement of participants to be influenced by the expected gender stereotype of a document. Thus, the research questions are addressed as follows: For answering RQ1, we consider hypotheses `H1`, and `H2`. In the statistical evaluation of the results, it is shown that participants perceive search results in the stereotypical female category Domestic Work as more relevant, when a female stereotype is expressed in the result document. In association with the query *how to easily clean at home*, the document expressing a female bias mentions a *Housewife*, whereas the male-biased document contains the word *Houseman*. An explanation of these results can be the stereotypical female expectation to perform care work, which seems to contribute to the different relevance judgements. According to Caroline Criado-Perez [22], 75% of globally done unpaid work is carried out by women – creating an unpaid-work imbalance between the genders, which is still an existing problem in today's modern society. Even though political efforts have been made to change this gender gap globally, it is still the reality that "working women" is not understood as tautology per se [22]. Taking into consideration that unpaid housework (predominantly consisting of cleaning activities) comprises the main workload of unpaid care work, it is not surprising that the experiment reveals the shown results. Thus, when users search for information in terms of cleaning at home, they do not seem to be negatively surprised or unsatisfied when confronted with a female-biased search result. On the contrary, a male-biased document seems to be perceived as less relevant, supporting the stated thesis of the understanding of stereotypical male and female activities and work in this area.

Regarding `H2`, the results of the experiment contradict our expectations. For the category Appearance, where a female stereotype is expected, the relevance judgement shows that non-relevant documents with male gender indication are rated higher in relevance than their female-indicating pendants. We should also highlight that a possible explanation of these results could be due to the formulation of the document content. For the query *what is considered plus size*, the non-relevant documents have titles such as *Percentage of men classified as underweight* and *Percentage of women classified as underweight*. Both documents include the sentence *Even if it does not seem so: a lot of men [women] struggle with their weight being too low but is it a gut feeling or is he [she] really underweight - let's find out!*. The combination of both title and text could imply semantically that it is astonishing and unexpected if the addressed person is underweight. Therefore, the reason for the shown results could be that participants perceive the stereotype-disconfirming (in the case of the male content) as more relevant due to the emphasis on the unexpectedness of men being underweight. In accordance,

studies find that females are stereotypically perceived to be more susceptible to struggle with their weight and appearance, being more critical of their bodies [27]. Thus, in this context, it seems to be of more relevance to users to find information about men, surprisingly being underweight in contrast to women.

Considering the results of the categories Career, Child Care, Cognitive Capability and Physical Capability, no significant effect between the relevance judgement and stereotype expectation is found. One interpretation of this result is that participants do not perceive gender-biased search results differently in their relevance and are not influenced by their gender stereotype expectation. We should however also take into account other reasons for those findings such as the overall setup of the experiment in combination with the formulation of document title and text, explained in detail in Section 4.4.

Lastly, RQ2 is assessed in the gender-specific experiment. In particular, H3 examines whether the genders of participants influence the perception of gender-biased retrieval results. Based on the experiments, no statistical significance between the participant's gender, the expected stereotype, and the relevance judgement is observed. In conclusion, participant gender appears to have no influence on the decision of the perceived stereotype confirmation or disconfirmation. Nevertheless, this result should be taken cautiously due to the small sample used in our study. Indeed, it contradicts some of the observations done in the studies of the presented literature. For instance, Hentschel et al. [13] show that gender stereotype perception differs in the evaluation of selves and others between males and females. Also, none of the genders seems to show an affinity to perceive stereotypical content predominantly different. A backlash effect, as observed in gender stereotype portrayal in image search results [19], or the perception of the social status of men [18] is not observed in our experiments.

4.4   Limitations of the Experiments

To begin with, any study conducted on Amazon Mechanical Turk must be critically reflected in view of associated ethical implications. The participants in our set of experiments received 0.1$ per assignment. One task published on the platform comprised three to six different assignments, i.e. three to six different relevance judgements. The average completion time per judgement was averagely around 300 seconds. In the scope of this work, the decision to utilize Amazon Mechanical Turk for the experiment conduction is based mainly on time and budget restrictions. For further experiments, the realisation of experiments beyond such platforms is recommended, e.g. in a laboratory setting with university students.

One of the main limitations of our study lays in the examined population that participated in the experiment. Just as in every laboratory environment, the presented results can only be considered as a reflection of reality to a certain extend. In terms of statistical power, the 50 participants per task of the gender-agnostic experiment and 10 participants per task of the gender-specific experiment do not represent large sample size, which may have affected the statistical conclusions based on the study's outcome. To extend the external validity, the experiment design could be adapted so that a real search engine environment and search task is simulated. This could be achieved by displaying a search result document after clicking the search button for a certain query so that a more realistic interaction is experienced.

Another limitation of the results lays in the choice of including only binary gender, namely male and female. This decision is mainly due to the limitations of the crowd-sourcing framework. For studies based thereupon, the inclusion of non-binary gender in the query-document compilations as well as in the participant selection is highly recommended. Today, self-identification beyond male or female is already strongly anchored in our real world - but rarely included in the overall IR systems research domain. The effects that this inclusion could have on the field of gender bias perception studies could open up a completely new perspective inside the whole research area and create deep insights into the role of gender-related concepts in information systems.

5   Conclusion and Future Work

Gender biases and stereotypes play a central role in the way how we perceive ourselves and others, and are found to be existent and particularly persistent in the IR systems we interact with. This paper aims to approach the question of whether expressed gender bias in the content of retrieval results influences the perceived user judgement of its relevance. By showing one-sided search results that reflect different gender stereotypes, an effort is made to bring together recent theories from sociology (i. e., gender stereotype perception) and the information system research (i. e., gender bias perception in IR systems), done through the lens of a set of human studies. As shown, participants are influenced by biased retrieval results in their relevance judgement, especially in female-related categories. These findings raise concerns in regard to the negative effects of gender bias in IR systems, and calls for more algorithmic accountability and transparency, especially for commonly used IR systems.

In this work, we focused on the relevance rating of one search result per query, absent of further context (such as source url or date) or the choice between different ranked documents that might influence the relevance perceived by users in real-world situations. We also do not address differences in the perception of stereotypical biased content in participants from distinct cultures or age groups, as our current study was limited to a group of MTurk workers. Possible effects of participants' gender attitudes and beliefs, as introduced by Behm-Morawitz and Mastro [2], on their relevance judgement of differently gender-biased content may be assessed in further experiments. Future work might also try to extend the developed experimental setup to include other SE-related concepts found to contain biases, such as automated query suggestions [3]. Here, our `Grep-BiasIR` dataset [16] opens the possibility to conduct further related experiments. Till then, as one of the first studies to explore effects of perceived gender biases in retrieval results on relevance judgements, this study presents an initial empirical contribution.

As a final remark, within what sometimes seems like a Chicken-Egg-Problem, questioning if humans produce biased systems or if biased systems produce or reinforce biases in humans, the protagonists of different disciplines (legal, commercial and federal) are required to act. Beyond that, a general improvement of diversity in the technology sector – free of gender, race or other social categories – could contribute to overall bias mitigation, beginning in every individual's mind and ending in each technological creation [28].

Acknowledgements

References

1. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., Wang, T.: MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs] (Oct 2018)
2. Behm-Morawitz, E., Mastro, D.: The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. Sex roles **61**(11-12), 808–823 (2009)
3. Bonart, M., Samokhina, A., Heisenberg, G., Schaer, P.: An investigation of biases in web search engine query suggestions. Online Information Review **44**(2), 365–381 (Dec 2019)
4. Chang, K.W., Prabhakaran, V., Ordonez, V.: Bias and fairness in natural language processing. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts (2019)
5. Chen, L., Ma, R., Hannák, A., Wilson, C.: Investigating the impact of gender on rank in resume search engines. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2018)
6. Cohen, G.L., Garcia, J.: "i am us": negative stereotypes as collective threats. Journal of personality and social psychology **89**(4), 566 (2005)
7. Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the trec 2019 deep learning track. arXiv preprint arXiv:2003.07820 (2020)
8. Danks, D., London, A.J.: Algorithmic bias in autonomous systems. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 4691–4697 (2017)
9. Fabris, A., Purpura, A., Silvello, G., Susto, G.A.: Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. Information Processing & Management (2020)
10. Gerhart, S.: Do web search engines suppress controversy? First Monday **9**(1) (Jan 2004). https://doi.org/10.5210/fm.v9i1.1111
11. Gezici, G., Lipani, A., Saygin, Y., Yilmaz, E.: Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal **24**(2), 85–113 (2021)
12. Glick, P., Fiske, S.T.: Sexism and other "isms": Independence, status, and the ambivalent content of stereotypes. In: Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence., pp. 193–221. American Psychological Association (1999). https://doi.org/10.1037/10277-008
13. Hentschel, T., Heilman, M.E., Peus, C.V.: The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. Frontiers in psychology **10**, 11 (2019)
14. 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 (2015)

15. Kordzadeh, N., Ghasemaghaei, M.: Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems pp. 1–22 (2021)
16. Krieg, K., Parada-Cabaleiro, E., Medicus, G., Lesota, O., Schedl, M., Rekabsaz, N.: Grep-BiasIR: A dataset for investigating gender representation-bias in information retrieval results. arXiv:2201.07754 [cs] (2022)
17. Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M.: Investigating gender fairness of recommendation algorithms in the music domain. Inf. Process. Manag. (2021). `https://doi.org/10.1016/j.ipm.2021.102666`
18. Moss-Racusin, C.A., Phelan, J.E., Rudman, L.A.: When men break the gender rules: status incongruity and backlash against modest men. Psychology of Men & Masculinity **11**(2), 140 (2010)
19. 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. pp. 6620–6631 (2017)
20. Otterbacher, J., Checco, A., Demartini, G., Clough, P.: Investigating user perception of gender bias in image search: the role of sexism. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 933–936 (2018)
21. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In google we trust: Users' decisions on rank, position, and relevance. Journal of Computer-Mediated Communication pp. 801–823 (2007)
22. Perez, C.C.: Invisible women: Exposing data bias in a world designed for men. Random House (2019)
23. Rekabsaz, N., Kopeinik, S., Schedl, M.: Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 306–316 (2021)
24. Rekabsaz, N., Lesota, O., Schedl, M., Brassey, J., Eickhoff, C.: TripClick: The Log Files of a Large Health Web Search Engine. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2507–2513. Association for Computing Machinery, New York, NY, USA (Jul 2021)
25. Rekabsaz, N., Schedl, M.: Do neural ranking models intensify gender bias? In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2065–2068 (2020)
26. Rekabsaz, N., West, R., Henderson, J., Hanbury, A.: Measuring societal biases from text corpora with smoothed first-order co-occurrence. In: Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021. pp. 549–560. AAAI Press (2021)
27. Sattler, K.M., Deane, F.P., Tapsell, L., Kelly, P.J.: Gender differences in the relationship of weight-based stigmatisation with motivation to exercise and physical activity in overweight individuals. Health Psychology Open **5**(1) (2018)
28. Shah, H.: Algorithmic accountability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **376**(2128), 20170362 (2018)
29. Sherman, J.W.: Development and mental representation of stereotypes. Journal of Personality and Social Psychology **70**(6), 1126 (1996)
30. Silva, S., Kenney, M.: Algorithms, platforms, and ethnic bias. Communications of the ACM **62**(11), 37–39 (2019)
31. Simpson, J.A., Kenrick, D.T.: Evolutionary social psychology. Psychology Press (2014)
32. Stangor, C., Jhangiani, R., Tarry, H., et al.: Principles of social psychology. BCcampus (2014)
33. Steele, C.M., Aronson, J.: Stereotype threat and the intellectual test performance of african americans. Journal of personality and social psychology **69**(5), 797 (1995)