Investigating User Perception of Gender Bias in Image Search

The Role of Sexism

Jahna Otterbacher
Open University of Cyprus
jahna.otterbacher@ouc.ac.cy

Alessandro Checco
University of Sheffield
a.checco@sheffield.ac.uk

Gianluca Demartini
University of Queensland
g.demartini@uq.edu.au

Paul Clough
University of Sheffield
p.d.clough@sheffield.ac.uk

ABSTRACT

There is growing evidence that search engines produce results that are socially biased, reinforcing a view of the world that aligns with prevalent social stereotypes. One means to promote greater transparency of search algorithms - which are typically complex and proprietary - is to raise user awareness of biased result sets. However, to date, little is known concerning how users perceive bias in search results, and the degree to which their perceptions differ and/or might be predicted based on user attributes. One particular area of search that has recently gained attention, and forms the focus of this study, is image retrieval and gender bias. We conduct a controlled experiment via crowdsourcing using participants recruited from three countries to measure the extent to which workers perceive a given image results set to be subjective or objective. Demographic information about the workers, along with measures of sexism, are gathered and analysed to investigate whether (gender) biases in the image search results can be detected. Amongst other findings, the results confirm that sexist people are less likely to detect and report gender biases in image search results.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Image search**;

KEYWORDS

Gender stereotypes; Search engine bias; User perceptions

ACM Reference Format:
Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210094

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210094

1 INTRODUCTION

Algorithmic processes that are opaque to users (and researchers) are having increasing influence on users' access to information. In the case of search, these processes also influence users' view of the information landscape and highly ranked results can shape public opinion [6, 11]. This is compounded by the trust users show in the results of search engines; often seen as objective and truthful [14]. More recently, studies have been undertaken to understand the impact of users' beliefs and unconscious biases on their search interactions and experiences [11, 18]. This paper continues this line of investigation, focusing on *user bias* within image search that can manifest itself through the decisions people make about the items they select (e.g., perceived quality and relevance), the queries they pose, etc. In particular, we seek to understand how user biases in the form of sexism, an area receiving less attention, shape peoples' views of image search and can reinforce gender stereotypes.

2 RELATED WORK

2.1 Gender Bias in Image Search Results

Image search engines play a powerful role in shaping peoples' views of the social world. In a study surrounding professions, Kay et al. [10] found that Google systematically returned more/fewer images of stereotype congruent/incongruent people, compared to labour statistics. Also, when users were shown gender-biased search results and were asked to estimate the corresponding labour statistic, this skewed their view of the distribution of men/women in a given profession. The study demonstrated that when search engines reproduce gender stereotypes, this has the effect of confirming and exacerbating already prevalent gender stereotypes.

Beyond the direct effect on users' perceptions, gender-biased search results increase the retrievability of some images, at the expense of others [4, 17]. Biased results increase the chances that stereotype-congruent images will be circulated more widely in society as compared to those that challenge stereotypes. Professionals, such as marketers or journalists who rely on image search, often in time-pressured environments [9, 12], may then be more likely to include stereotypical images in their work.

Recent work has moved beyond the case of gender stereotypes in image search related to professions, to those surrounding character traits [13]. Grounded in social psychology theories of person perception, it found that women were more often associated with warm (e.g., kind emotional) character trait searches whereas men were more often depicted in searches on agentic traits (e.g., assertive,

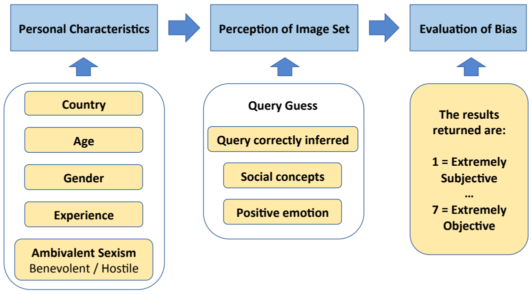

Figure 1: Conceptual model.

intelligent). In contrast to the work to date, which concerned the measurement of gender bias in search engine results, the present work aims to gauge the extent to which users actually perceive gender-based bias, and who (including their characteristics) might be more/less likely to report subjective results.

2.2 The Nature of Gender Stereotypes

The Stereotype Content Model [5] holds that we use two key dimensions in our perceptions of others – warmth (i.e., the extent to which we think someone has pro-social intentions) and competence/agency (i.e., the extent to which we believe someone is capable of achieving goals). This theory is widely accepted by researchers of person perception [1] and the two dimensions appear to be culturally universal [2]. These stereotypes describe women as being high warmth/low competence, and vice versa for men [15].

2.3 Ambivalent Sexism and the ASI

According to a classic definition [3, p. 9] prejudice is an "antipathy based upon a faulty and inflexible generalization." Prejudice towards a particular social group involves holding a stereotypical view of its members, reducing them to restricted social roles and/or characteristics. Therefore, sexism can be understood as gender-based prejudice. However, Glick and Fiske [7] pointed out that sexism differs from other forms of prejudice (e.g., racism) in that it is not unidimensional, consisting only of antipathy toward the target group. Rather, sexism is characterized by an ambivalence towards women: people may be hostile toward women, or they may hold benevolent attitudes, stereotyping them into limited, traditional roles. To measure the dimensions of sexism, they developed the Ambivalent Sexism Inventory (ASI), consisting of 22 items.

The ASI has been shown to reliably tap the two components of sexism: Hostile Sexism (HS) and Benevolent Sexism (BS). While HS and BS are positively correlated, HS tends to be associated with holding negative stereotypes of women, while BS tends to predict having a positive - although traditional - view of women. In addition, in a large-scale study, Glick and colleagues [8] demonstrated that HS and BS are universal across cultures. However, sexism is related to culture, and national averages on these measures are correlated to the levels of gender inequality in the society.

3 RESEARCH QUESTIONS

We use the ASI as a tool for understanding how users of image search engines may perceive gender bias. We present study participants with images returned by Google in response to character trait searches. Some result in image sets that are heavily gender-biased; while one is a neutral query ("hot air balloon"). As illustrated in Figure 1, extrapolating from theory, we expect that a user's level of sexism will be correlated to the manner in which she/he perceives the search results, which in turn influences the evaluation of the results or the degree to which they are seen as objective or biased. Specifically, we address the following questions:

RQ1: Are sexist/non-sexist people less/more likely to evaluate a heavily gender-imbalanced result set as being subjective?
RQ2: Is there evidence that sexist/non-sexist people perceive a given image result set differently?

4 METHODOLOGY

We conduct an experiment on the crowdsourcing platform Crowdflower, where we ask participants a set of questions to assess their perceived bias of image searches, together with other indicators related to web proficiency and cognitive behavior. We also administer the ASI, deriving scores for the BS and HS dimensions, as detailed in [7]. We repeat this experiment for UK, USA, and India.

To assess perceived bias we use a novel technique we call *reverse image search*: images are retrieved through a search engine, and then we ask participants to describe them. We then reveal the actual query used to retrieve the images and ask the users to compare this query with the description they provided. In this way, we can assess perceived bias without priming users on the topic of the experiment: when users are describing a set of images, they are not aware that those images have been obtained through a web search, nor that they will be asked about search engine objectivity. This allows us to jointly estimate user bias and user perception of bias for an image search results set. The task is structured as follows:

Part 1 (guess the query): The main part of the task consists of showing the users a grid of 9 images and asking them what keywords best represent/describe the images. This question is repeated for multiple grids, each obtained from a different query, as shown in Table 1. The phrasing of the question is carefully selected to avoid disclosing the fact that the image grid has been generated with an image search engine.
Part 2 (search engine opinions): The users are then asked to answer questions regarding the objectivity of search engines, along with a number of proficiency self-assessment questions.
Part 3 (perceived bias): Only at this stage in the experiment is the user told that the grid of images was obtained from a search engine. For each image grid, the query is disclosed and the user is required to compare it with the description they provided in Part 1 and assess the objectivity of each image result.
Part 4 (ASI). The user completes the ASI questionnaire to assess their level of sexism.

The current paper presents analyses concerning the data collected in Parts 1 and 4 of the experiment.

4.1 Dataset

Queries were chosen based on the findings of [13]. In addition to those listed in Table 1, the full dataset includes neutral character traits, as well as queries that result in non-biased image sets[1].

[1] The full anonymized dataset, containing 2.811 query-description comparisons for 281 different users equally split across the three regions and 10 unique queries, as well

Query	smart person	aggressive person	warm person	anxious person	hot air balloon
Trait	+	-	+	-	=
Bias	M	M	F	F	na

Table 1: Query used with corresponding trait (+ for positive, - for negative) and bias (M/F for bias towards males/females).

For our initial analysis, we focus on users' perceptions of bias in the neutral query "hot air balloon", as well as in four gender-biased queries detailed in Table 1. The theory of ambivalent sexism makes predictions on the correlation between BS and HS, and holding positive/negative views of women; thus we focus on positive/negative traits. Finally, we eliminate one observation, in which the participant did not disclose his or her gender.

5 RESULTS AND ANALYSIS

We confirm the validity of the ASI scores, conducting an exploratory factor analysis with varimax rotation to examine the structure of participants' responses. The analysis revealed a two-factor solution, corresponding to the two dimensions of Ambivalent Sexism - HS and BS. The solution explained 40% of the variance and had structural coefficients (i.e., loadings) of at least 0.44 for both factors. An acceptable degree of internal consistency was found; the factor corresponding to HS had a Cronbach alpha of 0.76, while the second factor corresponding to BS had an alpha of 0.74. ANOVA revealed regional (BS: F = 117.5***, HS: F = 102.4***) and gender (BS: F = 216.7***, HS: F = 135.3***) differences on both dimensions of the ASI[2]. Generally, men scored higher than women on both BS and HS. Participants from India scored higher on BS and HS as compared to those from the USA, who in turn, scored higher on both dimensions as compared to participants from the UK.

5.1 Is Sexism Correlated to Evaluating Results?

We now examine whether or not sexism has a direct correlation to how participants evaluate the objectivity of image result sets, which are known to be gender-biased. Table 2 details, for each query, a logistic regression model in which we predict the event that a participant evaluates the image set as *not being objective* (i.e., a rating of 1, 2, 3 or 4), based only on demographic characteristics, including the two ASI dimensions. Table 2 details the estimated coefficients (when statistically significant) along with a measure of fit, McFadden's R^2, which ranges from 0 to 1.

On the neutral query, "hot air balloon," only 28 of 280 participants indicated that the retrieved images were *not objective*. As expected, no participant characteristics were correlated to this outcome. In contrast, for the two image sets based on character traits with positive valance ("smart" and "warm"), we observe that, even when we control for country of residence, age, gender and self-reported experience, that the Benevolent dimension of sexism is negatively correlated to having evaluated these image sets as *not objective*. In other words, benevolent sexists are less likely than others to evaluate a set of images retrieved on the query "smart person" or "warm person," which primarily features images of men/women respectively, as being biased. This result is in line with the theory, which predicts that benevolent sexists hold positive, yet traditional views of women. Thus, they would arguably not be surprised to find images of men depicting a *smart person* (agentic trait) and women depicting a *warm person* (warm trait).

Interestingly, sexism is not correlated to participants' evaluations of the images retrieved on the queries "aggressive" or "anxious person." In particular, we might have expected to see the Hostile dimension of sexism playing a role in the case of "anxious person." This negative character trait retrieves primarily images of women, and hostile sexists tend to hold negative views of women.

5.2 Do Sexists Perceive Results Differently?

Having observed that sexism is correlated to judgments on the "smart" and "warm" image sets, we now examine these queries in detail. We test the conceptual model detailed in Figure 1, addressing RQ2. As depicted in the figure, we incorporate a mediating variable that attempts to capture how the participant perceives the images returned. To this end, we analyzed the query guess provided in Part 1 of the experiment. We processed the query guess using the Linguistic Inquiry and Wordcount tool [16]. Using a custom dictionary, we evaluated the extent to which the guess matched the true query. We also evaluated the extent to which the guess incorporated "social" words (including family relations and gender-related words) and "positive emotion" words.

For each query, we generated a structural equation model (SEM) using R's *Lavaan* package[3]. As shown in Table 3, the model incorporates the estimation of two latent constructs (user characteristics, perception of the images), which are not measured directly but result from multiple other measures, one directly measured construct (evaluation of objectivity), as well as the relationships between the constructs (i.e., the structural model). The measurement model corresponds to the smaller lower boxes in Figure 1 (e.g., Country), and the structural model to the upper boxes (e.g., Personal Characteristics). The left side of Table 3 depicts the model for the "smart person" query, which has a Comparative Fit Index of 0.98 and an RMSEA of 0.05. The SEM that describes participants' perceptions of and reactions to the "warm person" image results is also presented in Table 3, on the right side. The model has a Comparative Fit Index of 0.94 and a RMSEA of 0.07. In both cases, Table 3 reports the estimated coefficient for each variable in the SEM, along with its z-score and statistical significance. For both queries, we observe a positive correlation between the latent variable comprising the participant characteristics (gender and ASI dimensions) and the manner in which she/he perceived the set of images retrieved. In fact, further analysis (not reported) reveals that participants who correctly guessed the queries had a higher mean score on both the benevolent (warm: t =2.354*, smart: t =3.047**) and hostile (warm: t =2.984**, smart: t =4.018**) dimensions of the ASI, in comparison to those who did not correctly guess the query. We observe a negative correlation between the latent variable comprising the perception measures and the evaluation measure (i.e., perception of bias / non-objective). The observed correlations, while statistically significant, are rather weak. In future work, we plan to incorporate

as a full description of each data field, is available at github.com/AlessandroChecco/gender_bias.

[2] We use the following conventions: $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

[3] http://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Query	# Non-objective	Country	Age	Gender	log(Experience)	Benevolent	Hostile	Pseudo R^2
Hot air balloon	28	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.05
Smart person	70	n.s.	n.s.	n.s.	n.s.	-0.773**	n.s.	0.17
Aggressive person	59	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.06
Warm person	92	n.s.	n.s.	-0.690*	-0.247*	-0.510*	n.s.	0.13
Anxious person	60	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	0.04

Table 2: Logistic regression models to predict instances where image result set is evaluated as not objective.

Measurement Model				
	Smart		Warm	
	Est.	z	Est.	z
→User characteristics				
Male	1.00	fixed	1.00	fixed
Benevolent	3.805	4.420**	3.629	4.492**
Hostile	4.878	4.203**	5.219	3.849**
→Perception				
Guess match	1.00	fixed	1.00	fixed
Social	0.850	7.428**	0.815	3.284*
Positive	0.527	5.490**	0.275	2.646*
→Evaluation				
Objectivity rating	1.00	fixed	1.00	fixed
Structural Model				
User→Perception	0.013	2.628**	0.009	2.155*
Perception→Evaluation	-0.025	-2.535*	-0.028	-2.677**

Table 3: Structural Equation Models.

more of the information collected during our experiment, into the model, to improve its predictive power. However, the evidence generally supports the conceptual model put forward in Figure 1. In line with what the theory predicts, users of search engines who are more sexist, perceive image results differently than non-sexist people, and are less likely to perceive gender-biased results sets. Furthermore, it is the benevolent component of sexism that appears to be the most important.

6 CONCLUSIONS

Increasingly, attention is being paid to identifying and highlighting sources of bias within search engines. In this paper we investigate the impact of personal traits on identifying gender-biased image search results. Understanding prejudices and beliefs is critical in better understanding how people engage with and evaluate search technologies and may influence future design. Our findings confirm that people who are rated as more sexist according to the Ambivalent Sexism Inventory measure are less likely to recognize gender biases in image search, thereby reinforcing social stereotypes.

ACKNOWLEDGMENTS

This project is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement Nos. 739578 and 732328.

REFERENCES

[1] Andrea E Abele and Susanne Bruckmüller. 2011. The bigger one of the "Big Two"? Preferential processing of communal information. *Journal of Experimental Social Psychology* 47, 5 (2011), 935–948.

[2] Andrea E Abele, Mirjam Uchronski, Caterina Suitner, and Bogdan Wojciszke. 2008. Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology* 38, 7 (2008), 1202–1217.

[3] Gordon W Allport. 1954. The nature of prejudice. (1954).

[4] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 561–570.

[5] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 40 (2008), 61–149.

[6] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.

[7] Peter Glick and Susan T Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of personality and social psychology* 70, 3 (1996), 491.

[8] Peter Glick, Susan T Fiske, Antonio Mladinic, José L Saiz, Dominic Abrams, Barbara Masser, Bolanle Adetoun, Johnstone E Osagie, Adebowale Akande, Amos Alao, et al. 2000. Beyond prejudice as simple antipathy: hostile and benevolent sexism across cultures. *Journal of personality and social psychology* 79, 5 (2000), 763.

[9] Ayse Göker, Richard Butterworth, Andrew MacFarlane, Tanya S Ahmed, and Simone Stumpf. 2016. Expeditions through image jungles: the commercial use of image libraries in an online environment. *Journal of Documentation* 72, 1 (2016), 5–23.

[10] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. 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*. ACM, 3819–3828.

[11] 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 (CSCW '17)*. ACM, New York, NY, USA, 417–432.

[12] Lori McCay-Peet and Elaine Toms. 2009. Image use within the work task model: Images as information and illustration. *Journal of the Association for Information Science and Technology* 60, 12 (2009), 2416–2429.

[13] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. 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 '17)*. ACM, New York, NY, USA, 6620–6631. https://doi.org/10.1145/3025453.3025727

[14] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication* 12, 3 (2007), 801–823.

[15] Laurie A Rudman and Peter Glick. 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues* 57, 4 (2001), 743–762.

[16] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[17] Myriam C Traub, Thaer Samar, Jacco Van Ossenbruggen, Jiyin He, Arjen de Vries, and Lynda Hardman. 2016. Querylog-based assessment of retrievability bias in a large newspaper corpus. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 7–16.

[18] Ryen White. 2013. Beliefs and Biases in Web Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 3–12.