Show me a "Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users

Simone Kopeinik
Know-Center GmbH
Graz, Austria
skopeinik@know-center.at

Martina Mara
Johannes Kepler University
Linz, Austria
martina.mara@jku.at

Linda Ratz
Know-Center GmbH
Graz, Austria
lratz@know-center.at

Klara Krieg
University of Innsbruck
Innsbruck, Austria
klara.krieg@student.uibk.ac.at

Markus Schedl
Johannes Kepler University
Linz Institute of Technology, AI Lab
Linz, Austria
markus.schedl@jku.at

Navid Rekabsaz
Johannes Kepler University
Linz Institute of Technology, AI Lab
Linz, Austria
navid.rekabsaz@jku.at

ABSTRACT

Biases in algorithmic systems have led to discrimination against historically disadvantaged groups, including the reinforcement of outdated gender stereotypes. While a substantial body of research addresses biases in algorithms and underlying data, in this work, we study if and how users themselves reflect these biases in their interactions with systems, which expectedly leads to the further manifestation of biases. More specifically, we investigate the replication of stereotypical gender representations by users in formulating online search queries. Following prototype theory, we define the disproportionate mention of the gender that does not conform to the prototypical representative of a searched domain (e.g., “male nurse”) as an indication of bias. In a pilot study with 224 US participants and a main study with 400 UK participants, we find clear evidence of gender biases in formulating search queries. We also report the effects of an educative text on user behaviour and highlight the wish of users to learn about bias-mitigating strategies in their interactions with search engines.

CCS CONCEPTS

• **Human-centered computing → Web-based interaction**; **User studies**; **HCI theory, concepts and models**.

KEYWORDS

search queries, information retrieval, gender bias, prototype effect, user study

ACM Reference Format:
Simone Kopeinik, Martina Mara, Linda Ratz, Klara Krieg, Markus Schedl, and Navid Rekabsaz. 2023. Show me a "Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3580863

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580863

1 INTRODUCTION

The consumption of information online, particularly through search engines, has become part of most people’s daily routines. At the same time, the logic of underlying search engines and algorithms remains unknown for big parts of the population. While users generally trust the search engines’ algorithmic accuracy [42], it has been shown that biases in search results influence people’s information space and reality perception [22]. People’s cognitive and societal biases are, in turn, captured in data that feeds into algorithms and information systems. This can lead to a reinforcement loop that supports undesirable patterns of discrimination or outdated social norms in algorithmic decisions and, consequently, in user behaviour (i.e. users replication of biases) [3].

Prior research has demonstrated the existence of gender bias in search algorithms and search results, for instance, in the context of image search [30, 40, 41], query suggestion [12], and search engine result pages [11, 23, 43, 44]. These societal biases exist in and emerge from the various components of information systems, in particular data, algorithms, and end-users [20]. While previous work has mainly focused on the data and algorithm aspects, in this paper, we shed light on the role of users in the manifestation and propagation of biases in information systems in the particular domain of the retrieval of textual information. Specifically, we explore whether and how users replicate gender bias in their information search behaviour, as well as whether a higher awareness of users regarding gender bias issues might grant a mitigating effect.

To this end, we ground our work on the basis of *prototype theory* introduced in psychological research [45, 47] in order to investigate the unconscious application of gender bias in stereotypical areas. Let us showcase such unconscious biases with the following example used in earlier studies [6, 7]:
"*A father and son are in a horrible car crash that kills the dad. The son is rushed to the hospital; just as he’s about to go under the knife, the surgeon says, ‘I can’t operate—that boy is my son!’ Explain.*”
Belle et al. [7] show that adults are significantly more likely to interpret the scenario above with explanations about multiple fathers than by assuming that the surgeon could be female – the mother. Prototype theory offers a sound explanation for the rationale underlying this behaviour: people learn according to mental categories, each category being formed around one or more prototypical examples. Membership of a category is defined by similarity to these

examples [47], which can vary according to context [45]. When individuals receive a stimulus (e. g. seeing or hearing something), they first try to explain it via an existing category. A combination of categories is mentally simulated only if the stimulus does not correspond to an existing category or concept. Such "mismatch" of the stimulus to the category can be observed by a person's explicit naming of new properties, e. g. laughing cat, green banana, or male nurse [55].

In this study, we follow a simplified interpretation of the prototype theory. We view a prototype as the entity that evokes the strongest cognitive association with a category and gender stereotypes as a manifestation of social categories of women and men, represented in people's cognitive perception and shaped by societal context. Adapting this definition to our domain of information retrieval, we assume that in a gender-stereotyped domain of information search, the representative prototype would be either female or male, depending on the context and, is observable from user interactions with search engines (query formulation). We pursue the following three research questions:

[RQ1] Do users replicate gender stereotypes in the formulation of search queries on the web? In an online user study, we expose participants to examples of search results (text documents in which females, males, or no gender are specified) and ask them to formulate a query for retrieving a document. Following the rationale of the prototype theory, we examine whether the participants include an explicit gender mention in the formulated query and if they are more likely to do so if the document text does not conform with the stereotype of the domain. Our results confirm that users replicate gender bias in their search interactions, as the documents that do not conform with the gender stereotype of the domain, receive a significantly higher number of (explicit) gender mentions in their formulated queries.

[RQ2] Does the extent to which users replicate gender stereotypes in search queries depend on the personal characteristics of a user? By asking about the demographics and political views of the participants, we further investigate whether these factors influence the measured gender bias. The results of this study show a weak but significant influence of conservative political views on the replication of gender bias. We did not observe any significant effect based on gender, or the educational level of the participant.

[RQ3] Can information on avoiding gender stereotyping raise awareness and mitigate the effect? In the main study, we implement a between-subject design with two conditions. Participants in the experimental group are presented with an educative text about biases in search engine results. We investigate whether providing such information raises awareness on the topic matter and consequently influences participants' formulation of search queries. Our results show a moderating effect of the educative text on male participants, who in the experimental group exhibit a more balanced use of female and male gender mentions.

Complementing the main study, four additional questions were posed to gain a brief understanding of people's attitudes toward bias in search engines and their interest in addressing this issue. The results emphasise the participants' interest in reflecting on biased search behaviour and acquiring further information on this topic. Overall, our work takes a step towards a deeper understanding of the biased behaviours of the end-users (with various demographics) in utilising search engines, and showcase the benefits of participating the users in addressing this issue, and making them mindful of such biases.

The remainder of the paper is organised as follows: we discuss related work in Section 2, followed by explaining our methodology for measuring stereotypes in query formulation in Section 3. Section 4 describes the pilot study and reports its results. In Section 5, we explain in detail the setup of our main study, whose results are presented in Section 6. Finally, in Section 7, we discuss the impact of the findings, limitations of the studies, and future work. The collected data in our studies is available at **https://github.com/CPJKU/user-interaction-gender-bias-IR**.

2 RELATED WORK

2.1 Prototype Theory and Gender Stereotypes

Prototype theory [47, 48] is a cognitive linguistic theory that looks at how people categorize objects or concepts. It describes mental aspects of categories to be formed around a prototype, namely the most central member of a category. The membership of an item to a category depends positively on how many characteristics (i. e. stimulus values) are shared with the prototypical representation of the category and depends negatively on how many characteristics are shared with other categories. The concept of prototypes is also referred to as typicality or family resemblance [46], and it is shown in further studies that people rate an item belonging to a category (i. e. the gradient of an item's membership) in line with their idea of a category. In fact, according to Rosch [45], this process applies to all different kinds of categories, such as semantic categories (e. g. furniture) but also biological (e. g. sex) or social and political categories (e. g. gender, occupation and democracy). Moreover, Rosch [45] discusses that concepts and categories should not be considered static entities but rather resonate in a complex dynamic of inter-connectedness to other categories, contexts and situations. These can be adapted or recreated when learning about additional information or context that changes the representative idea of a category. The mentioned studies provide the theoretical foundation of our work and motivate us to study how prototypical manifests are reflected in the interaction of users with search engines.

We assume that societal conceptions of people who serve as prototypical representatives of domains such as childcare or career (unfortunately still) bear a bias towards female or male gender categories. Thus, the prototype of a person who looks after children would be a woman, while the prototype of a person who has a professional career would be a man. Assuming you hold such a (unconscious) gender bias in the prototypical conception of domain representatives, you would probably not explicitly ask for a "female nurse" when searching for a female nurse because you consider the feature "female" to be an inherent part of the prototype that does not necessitate any extra mention. However, if you were searching for a male nurse, you might be more likely to make an explicit gender mention in your query for a "male nurse", since "male" is not considered a typical feature of the prototype. To illustrate the phenomenon with an additional gender-unrelated example from daily life, think of a green banana that many people would explicitly describe as "green", while the "yellow" of a prototypical banana usually needs no mention (cf. [55]).

Table 1: Examples of gender biased documents selected from the `Grep-BiasIR` dataset [33].

Gender Indication	Title	Body Text
Document 1, *Domain:* Child Care, *Expected Stereotype:* Towards Female		
Female Prototypical content	`Child Care and Working Mom: Extended Parental Leave For Moms`	`The authors investigate the relationship between family policy and women's attachment to the labour market and focus specifically on policy feedback on women's subjective work commitment.`
Male Counter-prototypical content	`Child Care and Working Dad: Extended Parental Leave For Dads`	`The authors investigate the relationship between family policy and men's attachment to the labour market and focus specifically on policy feedback on men's subjective work commitment.`
Non-gendered	`Child Care and Working Parent: Extended Parental Leave For Parents`	`The authors investigate the relationship between family policy and parents' attachment to the labour market and focus specifically on policy feedback on parents' subjective work commitment.`
Document 2, *Domain:* Career, *Expected Stereotype:* Towards Male		
Female Counter-prototypical content	`What enables some women to become CEOs?`	`The authors found that working with the 'self' is vital for women aiming to obtain and carry out the job of CEO. The female CEOs in the study described the way they had to use their leadership ambition and potential in order to reach the top.`
Male Prototypical content	`What enables some men to become CEOs?`	`The authors found that working with the 'self' is vital for men aiming to obtain and carry out the job of CEO. The male CEOs in the study described the way they had to use their leadership ambition and potential in order to reach the top.`
Non-gendered	`What enables someone to become CEO?`	`The authors found that working with the 'self' is vital for people aiming to obtain and carry out the job of CEO. The CEOs in the study described the way they had to use their leadership ambition and potential in order to reach the top.`

The scientific literature on the prevalence of gender biases and stereotypes in society suggests that "male nurses" or "female CEOS" can still be analogized to green bananas in our minds. Gender stereotyping is described as the development of mental categories to process gender-related information, mainly reflected in the distinction between common social constructs of women and men [16]. Studies from the 1950s pointed out that men are often perceived to be specialized in task-oriented or instrumental behaviour, while women are seen as specialized in socio-emotional tasks such as caring for others [51]. Similarly, traits that describe stereotypical views of women and men are usually arranged along two dimensions, namely *communality* and *agency* [5, 28]. The label *communal* circumscribes the belief that women are concerned with the welfare of others and are more friendly, unselfish, gentle, and understanding. On the other hand, the label *agentic* describes the idea that men are more assertive, controlling, active, competitive, and self-confident [19, 31].

The internalization of domain-specific gender stereotypes begins at a very young age [21]. Empirical research suggests that already preschool and elementary school children express, in line with societal stereotypes, that girls are more interested in appearance and being pretty, for example, while boys are described as more interested in physical activity, sports, or fighting [39]. For children, as for adults, gender stereotypes ultimately do not only have a descriptive but also a prescriptive function, as they set standards about how different genders should be(have) [31]. Thus, stereotypes are considered a form of discrimination as they might hinder members of a stereotyped group from developing individual abilities or making non-conforming life choices.

In this work, we contribute to existing research strands by investigating whether people reproduce gender stereotypes when formulating search queries and whether they would appreciate information about bias-mitigating behaviour.

2.2 Gender Bias in Information Retrieval

Societal biases are reflected in the ecosystem of information access systems in various forms and functionalities [3, 4]. Such biases can be originated from different components of these systems, i. e. from data collection, model design, and evaluation, but also from the interactions of users with systems [20].

Societal biases and in particular, gender bias in information retrieval (IR) systems and search engines have been the focus of several recent studies. For instance, Otterbacher et al. [40] show that an image search engine portrays stereotypical character traits of men (conveying power) and women (conveying sexual concepts). Rekabsaz and Schedl [44] and later Gezici et al. [24] study biases in search engine results, where retrieved documents for a given bias-sensitive search query are considered to be biased if the results show an unbalanced representation of viewpoints or underlying populations. Such biases in search results can influence the social cognition of users, leading to stereotype confirmation [30, 49]. This is particularly the case considering that first users tend to perceive the top-ranked results as the most important contents [23, 24], and second users commonly perceive search engines' results as the "state of the world" [42].

Recent studies approach the issue of bias in search results from the point of view of retrieval systems. In particular, Rekabsaz and Schedl [44] demonstrate that neural ranking models intensify gender bias, while Bigdeli et al. [9] study the interplay of the biases and performance of retrieval systems. Recent studies approach mitigating these biases in machine learning models through methods such as adversarial training [43], regularization [56], data preprocessing [10], and data collection analyses [11].

Besides system-oriented approaches, some studies investigate the perception of users with respect to societal biases mostly in the field of image search. Kay et al. [30] expose that the Google search engine systematically shows more images of stereotype congruent persons when compared with actual labour statistics and demonstrates that, while the study's participants accurately reflect real-world gender ratios in occupations, their perceptions can be negatively influenced by biased search results. Following this direction, Otterbacher et al. [41] explores benevolent sexism in participants when interacting with a biased image search engine. More recently, Krieg et al. [34] investigate how the existence of gender stereotypes in the content of a document may influence the users' judgment regarding the relevance of the document to a corresponding query.

The work at hand directly contributes to this direction and complements the discussed literature by studying user behaviour regarding gender bias. In particular, we investigate whether prototypical gender biases of users influence the way information need is formulated when interacting with search engines.

3 MEASURING BIAS IN QUERY FORMULATION

In this section, we explain our methodology to measure and evaluate the potential gender bias of users when formulating information needs for a search engine. We first describe our method to generate and label potentially biased queries, followed by explaining the adopted gender bias metrics.

3.1 Generating Prototypical/Counter-prototypical Queries

Our approach to generating labelled queries for the subsequent experiments is composed of the following steps: First, we prepare a collection of biased documents (i.e. each document representing one online search result) with respect to gender, in which each document contains word(s) that explicitly indicate a gender. Next, we present one of the documents to a survey participant and ask them to formulate a highly relevant query for the document. Finally, we measure the gender bias of the formulated queries based on the occurrence of gendered words in the query. In the following, we explain the general process of experiments used later in our user studies. The details of the user studies, such as the selected documents, information of participants, and statistics of generated queries, are explained in the succeeding sections (Section 4 and Section 5 for the pilot and main study, respectively).

Biased documents. To conduct our experiments, we first need to prepare a set of biased documents. In our experiments, we select the biased documents from the `Grep-BiasIR` dataset [33]. `Grep-BiasIR` provides a collection of 118 documents whose content revolves around societal topics that can potentially reflect gender stereotypes. In our studies, we select a subset of these documents (details in the respective sections). Each of the biased documents in `Grep-BiasIR` consists of a title and body text, provided in three variations with different *gender indications*, namely either with female, male, or neutral words. Examples of two documents and their corresponding variation of gender indications are shown in Table 1. `Grep-BiasIR` also accompanies each document (regardless of its variation) with a domain, i. e. career, appearance, and child care, and the *expected stereotype* label. Given the expected stereotype of a document, the gender indication of each variation of the document defines whether the variation contains prototypical or counter-prototypical content. For instance, as shown in Table 1, the expected stereotype of the document on top is *Towards Female*, and hence the document variations with female and male indications are considered as *prototypical* and *counter-prototypical* content, respectively.

Table 2: Examples of queries generated by participants and their corresponding prototypical labels. When any gendered word is mentioned in the text of a query, the label is set to *prototypical mention (pm)* or *counter-prototypical mention (cpm)* depending on whether its respective document has a prototypical or counter-prototypical gender indication. The *no-mention (nm)* label is given when no gendered words appear in the query.

Document 1		
Domain: Child Care		
Expected Stereotype: Towards Female		
Title: `Child Care and Working Mom: Extended ...`		
Body Text: `The authors investigate the ...`		
Gender Indication: Female ⟶ Prototypical content		
Participants' generated queries:		
Query Text	Gender Mentioned?	Label
`working mums and childcare`	Yes	*pm*
`info on parental leave`	No	*nm*
`going back to work child care`	No	*nm*
`maternity leave laws`	Yes	*pm*
Title: `Child Care and Working Dad: Extended ...`		
Body Text: `The authors investigate the ...`		
Gender Indication: Male ⟶ Counter-prototypical content		
Participants' generated queries:		
Query Text	Gender Mentioned?	Label
`paternity leave`	Yes	*cpm*
`childcare parental leave`	No	*nm*
`childcare for working dads`	Yes	*cpm*
`dads and parental leave`	Yes	*cpm*
Document 2		
Domain: Career		
Expected Stereotype: Towards Male		
Title: `What enables some men to become CEOs?`		
Body Text: `The authors found that working with ...`		
Gender Indication: Male ⟶ Prototypical content		
Participants' generated queries:		
Query Text	Gender Mentioned?	Label
`how men get to the top`	Yes	*pm*
`becoming a CEO`	No	*nm*
`what makes a good CEO`	No	*nm*
`how to be a ceo`	No	*nm*
Title: `What enables some women to become CEOs?`		
Body Text: `The authors found that working with ...`		
Gender Indication: Female ⟶ Counter-prototypical content		
Participants' generated queries:		
Query Text	Gender Mentioned?	Label
`how to be a female ceo`	Yes	*cpm*
`women becoming CEOs`	Yes	*cpm*
`skills needed to be a ceo`	No	*nm*
`female career success`	Yes	*cpm*

Generating queries. The next step is generating queries by users. In our experiments, we conduct user studies in which participants are asked to formulate a highly relevant query to the content of a given document (details provided in the succeeding sections). In the conducted user studies, either the prototypical or counter-prototypical variation of a document is shown to a participant, and the participant is asked to formulate a query that (when they submit it to a search engine) makes the document appear at the top of search results. Table 2 shows representative examples of the queries generated by participants for the gendered variations (prototypical and counter-prototypical contents) of two documents. We should note that in our experiments, in order to avoid revealing the purpose of gender bias measurement, we also ask participants to formulate queries for the non-gendered variations of the documents. In our analyses, however, we only focus on the results of gendered variations as the non-gendered variations appear to be non-relevant to our studied problem (more details in Section 6).

Labeling queries based on gender mentions. In this step, we identify whether gendered words are mentioned in the texts of the generated queries. Such words can be gendered pronouns, gender-specific words (e. g. actress or congressman), or names. A query could contain more than one gendered word for a specific gender, however, in our experiments, we do not observe any case of mentioning more than one gender in the generated queries. If the query is gendered, its mentioned gender also always matches the gender indication of the corresponding biased document. Based on this information, we label the prototypical gender inclination of each query. If no gendered word is mentioned in the query, the label is *no mention (nm)*. Otherwise, when the gendered query corresponds to a prototypical or counter-prototypical document variation, the corresponding label is *prototypical mention (pm)* or *counter-prototypical mention (cpm)*, respectively. Table 2 shows examples of such labels for the generated queries. We use these labels in the following to define proper gender bias metrics for our experiments.

3.2 Measuring Gender Bias

To systematically analyse the studies' results and tease out significant phenomena, in line with prototype theory, we define a set of simple statistical measures explained in the following.

Referring to the document set as $\mathbb{D}$, we split this set into $\mathbb{D}^{(cp)}$ and $\mathbb{D}^{(p)}$, which refer to the subsets of document variations with counter-prototypical contents and the ones with prototypical contents, respectively. After labelling the queries according to the procedure above, each set of documents can then be related to a set of queries, namely $\mathbb{Q}^{(cp)}$ for $\mathbb{D}^{(cp)}$, and $\mathbb{Q}^{(p)}$ for $\mathbb{D}^{(p)}$. As discussed in the previous subsection, each query is tagged with *cpm*, *pm*, or *nm* labels. Given these two sets of queries, we define N_{cpm} as the number of queries labeled with *cpm* in $\mathbb{Q}^{(cp)}$, formulated as $N_{cpm} = |\{q \in \mathbb{Q}^{(cp)} | \text{label of } q \text{ is } cpm\}|$. Similarly, N_{pm} is defined as the number of queries labeled with *pm* in $\mathbb{Q}^{(p)}$, formulated as $N_{pm} = |\{q \in \mathbb{Q}^{(p)} | \text{label of } q \text{ is } pm\}|$. Using these definitions, the relative frequency (f) of *cpm* and *pm* queries are formulated as follows:

$$f_{cpm} = \frac{N_{cpm}}{|\mathbb{Q}^{(cp)}|}, \qquad f_{pm} = \frac{N_{pm}}{|\mathbb{Q}^{(p)}|} \tag{1}$$

The f_{cpm} and f_{pm} quantities report the relative frequency of the appearance of counter-prototypical and prototypical mentions of genders in queries (according to the corresponding documents), respectively. As by definition $N_{cpm} \leq |\mathbb{Q}^{(cp)}|$ and $N_{pm} \leq |\mathbb{Q}^{(p)}|$, the numeric range of f is $[0, 1]$. To understand whether the phenomenon of explicit gender mentions is related to stereotypical thinking, following previous studies [15, 37], we define the *Mention Gap (MGap)* of gender mentions as the difference between f_{cpm} and f_{pm}:

$$MGap = f_{cpm} - f_{pm} \tag{2}$$

The *MGap* metric is defined as the difference between f_{cpm} and f_{pm} and results in the numeric range of $[-1, 1]$. The significance of this difference can be examined with proper tests such as the chi-square test, where the counter-prototypical mentions and prototypical mentions represent the categorical variables.

To illustrate the intuition behind the *MGap* metric, let us consider the following example. If a person wants to find search results about female nurses and explicitly mentions the word "female" in the search query, this would be categorised as a prototypical gender mention. However, this may be assumed unnecessary for the formulation of information needs and hence be skipped by the user since the characteristic "female" corresponds to the prototype of the nurse (similar to a "yellow" banana). The omission of the prototypical gender label could therefore be considered an indication of an existing gender bias. However, if a person wants to find search results about male nurses and explicitly mentions the word "male" in the search query, this is a counter-prototypical gender mention. This is eventually considered a central search criterion since the characteristic "male" deviates from the traditional prototype of a nurse (similar to the "green" banana). In other words, the addition of the counter-prototypical gender label to a search query can be seen as an indication of gender bias. The *MGap* metric (with values ranging from -1 to 1) informs about the ratio of gender mentions across multiple queries that are provided by the participants of a study. Smaller values of *MGap* indicate that the participants use prototypical and counter-prototypical gender mentions in an equal ratio. Larger values of *MGap*, on the other hand, show a more frequent indication of counter-prototypical (i.e. $MGap > 0$) or prototypical (i.e. $MGap < 0$) gender mentions in generated queries. Hence, according to our assumption, the more likely traditional gender images are replicated.

Participant-level metrics. The discussed metrics so far provide experiment-level statistics aggregated over all participants' responses. However, we are also interested in examining the characteristics of groups of participants, e. g. according to their demographics or political stands. To this end, in the following, we revisit the introduced metrics and define them on the participant level.

To calculate per-participant statistics, given a participant i, we select the queries in the sets $\mathbb{Q}^{(cp)}$ and $\mathbb{Q}^{(p)}$ that were generated by i. We refer to these subsets as $\mathbb{Q}^{(cp,i)}$ and $\mathbb{Q}^{(p,i)}$, respectively. Using these two query sets, we define $N^{(i)}_{cpm}$ as the number of queries labelled with *cpm* in $\mathbb{Q}^{(cp,i)}$, and $N^{(i)}_{pm}$ as the number of queries labelled with *pm* in $Q^{(p,i)}$. The corresponding metrics for the participant i are defined as:

$$f_{cpm}^{(i)} = \frac{N_{cpm}^{(i)}}{|\mathbb{Q}^{(cp,i)}|}, \qquad f_{pm}^{(i)} = \frac{N_{pm}^{(i)}}{|\mathbb{Q}^{(p,i)}|}, \qquad MGap^{(i)} = f_{cpm}^{(i)} - f_{pm}^{(i)} \tag{3}$$

The value of $f_{cpm}^{(i)}$ indicates the participant's manifestation of counter-prototypical gender mentions in search queries as described in prototype theory. Similarly, $MGap^{(i)}$ shows the tendency toward counter-prototypical/prototypical gender mentions.

4 PILOT STUDY

To explore our research questions and test the material in the initial phase, we conducted a pilot study on the crowdsourcing platform Amazon Mechanical Turk (MTurk). Each participant was randomly assigned document examples of the three document variations (i.e., female, male, non-gendered), and was asked to formulate a search query in order to find a document in the top search results. The selection of documents was retrieved from the Grep-BiasIR dataset [33]. Results indicate the replication of gender stereotypes in the formulation of search queries. The participants' gender shows no effect on the dependent variable *MGap*.

4.1 Participants

In the study, $n = 423$ US-located workers of the crowdsourcing platform MTurk participated. This was reduced to $n = 224$ after data cleaning and excluding participants with incomplete data. The distribution between female and male participants is approximately even for each document variation. However, the number of queries for a document variation lies between 18 and 30. We should note that in this pilot study, the number of tasks assigned to a participant varies highly. This is due to the design constraints in this platform, as the participants were able to complete as many tasks – namely formulating a search query for a given document – as they wish. We considered this as a limitation of the pilot study and addressed it in the main study.

4.2 Implementation Details

4.2.1 Material. We conducted our experiments on a subset of documents from the dataset Grep-BiasIR [33]. The dataset provides bias-sensitive query-document pairs categorised according to the domains that reflect gender stereotypes. Each document is available in three document variations, i.e., phrased as female, male, and non-gendered. The selected documents belong to the five categories of Appearance, Career, Child Care, Cognitive Capabilities, and Physical Capabilities, where we selected three documents per category. We intentionally excluded the categories Domestic Work and Sex & Relationship, due to the multilayered nature of these categories with respect to genders and intrinsic difficulties in defining stereotypes/counter-stereotypes. The resulting set contains $n = 15$ documents (5 categories, 3 documents each) and is provided in the supplementary materials. Table 1 shows two sample documents and the three variations of each.

4.2.2 Procedure. After giving informed consent, participants were able to select the tasks. In each task, a single document consisting of a header and a short paragraph is given. The document resembles the result of a search, and the participants were asked to formulate a web search query of up to four words that would place this document in the top search results of a search engine.

Table 3: Pilot Study: Fraction of gender mentions in queries as a response to counter-prototypical and prototypical document examples. The numbers reveal a low effect gender bias with high significance (chi-square test with $\phi = .13$ and $p < .005$), which results from the difference between counter-prototypical and prototypical gender mentions in search queries.

Documents Set	f_{cpm}	f_{pm}	$MGap$	Chi-square test	
				ϕ	$p-value$
All documents	.81	.62	.19	.13	**.002****
Only with Female Stereotypes	.73	.52	.21	.17	**.02***
Only with Male Stereotypes	.86	.69	.17	.11	**.04***

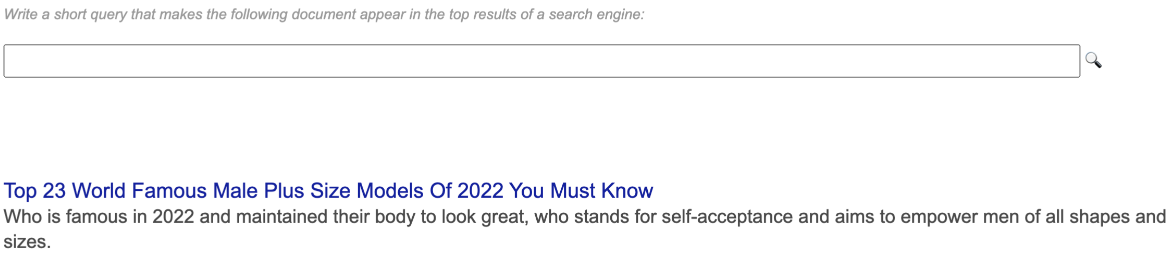
Table 4: Study Design - Main Study: Participants were randomly assigned to either the experimental or control group and consequently received a stimulus text educating about search engine bias or not. Both groups received eight randomly assigned documents from the three document variations.

Condition	Stimulus	Document Variations
Control	Control Text	Female/Male/Non gendered
Experimental	Educative Text	Female/Male/Non gendered

4.3 Results and Implications

The results of the pilot study calculated on a collective level, are presented in Table 3. The results show a significant difference in the frequency of the counter-prototypical and prototypical mentions of gender, captured by the Mention Gap (*MGap*) metric. This indicates the replication of gender bias in the wording of queries by search engine users, confirming our initial assumption.

Implications for the Main Study. As mentioned before, the pilot study did not control for the number of documents assigned to a participant, resulting in a high variation in the number of completed tasks per participant (from 1 to 63). For the main study, the design was adapted, and each participant received exactly the same number of documents. Additionally, leveraging the experience of the pilot study, in the main study, we revised the inputs to ensure the correctness and clarity of the documents. Finally, in the main study, we reduced the number of documents to eight by focusing only on the ones, that showed the highest tendency for biased responses in the pilot study (details are provided in the following section). This decision was made due to resource constraints, and with the aim of increasing the number of data points in the main study for the remaining documents.



Write a short query that makes the following document appear in the top results of a search engine:

Top 23 World Famous Male Plus Size Models Of 2022 You Must Know

Who is famous in 2022 and maintained their body to look great, who stands for self-acceptance and aims to empower men of all shapes and sizes.

Figure 1: Screenshot of the Main Study where a document with male content is presented.

Table 5: Main Study: Participant distribution according to the experimental condition, age, and gender.

Condition	Age	Participant Gender			
		Female	Male	Non-binary	Total
Control	18 - 29	33	27	1	61
	30 - 49	43	58	0	101
	50 - 64	22	20	0	42
	≥ 65	7	3	0	10
	Total	105	108	1	214
Experimental	18 - 29	25	25	1	51
	30 - 49	46	38	2	86
	50 - 64	19	19	0	38
	≥ 65	4	7	0	11
	Total	94	89	3	186
Total		199	197	4	400

5 MAIN STUDY: EXPERIMENTAL SETUP

5.1 Design

The major change in the design of the main study constitutes the definition of the tasks within our crowdsourcing platform. In the pilot study, we defined the generation of one search query responding to a document example as one task in MTurk, which significantly limited the control over the study setup. Thus, in the main study, we corrected this flaw and one task was defined as the entire study participation. In addition, we investigated the reflection of the participants' gender bias in formulating search queries by implementing a two (Stimulus: educative text vs control text; between subjects) by three (Document variation: female vs male vs non-gendered; within-subjects) independent measure design. The first dimension represents the stimulus, realised by the educative versus control text, and distributed between subjects. The second dimension belongs to the variations of each document, namely female versus male versus non-gendered, and is assigned within subjects.

More specifically, prior to starting the tasks work, each participant received either the educative or control version of the stimulus text. This divided the participants into experimental and control groups. The participants of the experimental group were informed how search engines may reflect societal bias and stereotypes in search results (educative text). The control group was not presented with the same information but received a generic text on the general technical mechanisms of search engines (control text). The provided texts can be found in Section 2.2 of the supplementary material. In this study, we specifically opted for controlling the number of assigned documents, by giving all (eight) documents to both groups of participants. The documents were provided in a random order, and for each document one of its three possible variations was randomly selected and shown.

To summarise, the independent variables of this study are hence defined by the stimulus prior to the task work (between-subject), and the document variations (within-subject), and the dependent variable is the measure of the relative frequency of the explicit mentions of gender in the search queries, as formulated in Section 3.2.

5.2 Participants

The participants were recruited using prolific.co and selected in a way that they form a gender-balanced population. The participants reside in the UK, are fluent in English, and have a total number of 400. Table 5 shows the distribution of this population according to the experiment condition, and the reported age and gender of the participants. The majority of participants have either a high school education or equivalent (n = 149), or an academic degree (n = 247). Four participants stated primary school education as their highest educational level.

5.3 Implementation Details

We selected a subset of 8 out of the 15 documents used in the pilot study. The documents relate to the four domains of Appearance, Physical Capabilities, Career, and Child Care. We chose the documents with the highest margins of gender mentions in the pilot study. The documents are provided in the supplementary materials. Each participant received one variation of each of the eight documents. A sample query formulation page in the study is shown in Figure 1. In addition, we posed a number of questions (see Table 6) to collect information on the participants' characteristics ($Q1$ to $Q3$), search engine familiarity (i.e., $Q4$ to $Q6$) and their attitude toward bias in search engines ($Q7$ to $Q10$).

5.4 Procedure

Before starting the questionnaire, participants were informed about the terms of the survey (i.e., duration, task, use of data), and gave

Table 6: Main Study: Questions posed to the study participants. Q1 was retrieved as a numeric input; Q2 and Q4-Q10 were inquired as 5 point Likert scale coded as 1=Stronlgy disagree to 5=Strongly agree; Q3 was coded as 1=Strongly conservative, 2=Moderately conservative, 3=Slightly conservative, 4=Neutral, 5=Slightly liberal, 6=Moderately liberal, 7=Strongly liberal, 8=Prefer not to say; Q11 and Q12 were considered as categorical variables (pls see supplementary material).

ID	Question
$Q1$	What is your age?
$Q2$	I identify as a feminist - someone who advocates and supports equal opportunities for women.
$Q3$	How would you describe your political view?
$Q4$	I am confident about my online search abilities.
$Q5$	When using a search engine, I always find the information I am looking for.
$Q6$	I use search engines to find important information rather than other sources, e.g., books, newspapers.
$Q7$	I consider search engines to be a fair and unbiased source of information.
$Q8$	As a user, I would like to receive information when my interactions with a search engine reflect societal stereotypes.
$Q9$	As a user, I would like to receive more information when search engines possibly reflect existing societal stereotypes.
$Q10$	In my future usage of search engines, I will pay attention whether my interactions with the system reflect societal stereotypes.
$Q11$	What gender do you identify with?
$Q12$	What is the highest degree or level of education you have completed?

Table 7: Number of participants and document examples per condition.

Condition	Participant's Gender	#Participants	#Query datapoints			
			Female	Male	Non-gendered	Total
Control	Female	105	270	276	289	835
	Male	108	264	281	308	853
	Non-binary	1	3	3	2	8
	Total	214	537	560	599	1696
Experimental	Female	94	237	249	255	741
	Male	89	227	224	253	704
	Non-binary	3	7	5	12	24
	Total	186	471	478	520	1469
Total		400	1008	1038	1119	3165

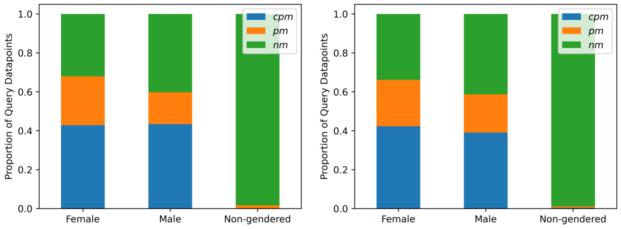
their informed consent. Then, participants had to answer socio-demographic questions regarding age, gender, and education. Next, they received instructions and a text about search engines that differed according to their random assignment to one of the two experiment conditions (see Section 5.1). Subsequently, participants of both conditions completed a sequence of the task of formulating search queries for a given set of eight documents. Following a within-subject design, each document was chosen randomly from the three document variations, i.e. male content, female content, and non-gendered content. Finally, participants were presented with extra questions regarding their search engine usage, their opinion on search engine bias, and their political views (see Table 6). See supplementary materials for more details on the instructions and tasks.

The experiment contained two attention checks. We excluded submissions of participants who either failed both attention checks or failed one attention check and took less than five minutes to complete the survey. This procedure follows the rejection criteria supported by Prolific.

5.5 Characteristics of the Collected Data

Table 7 reports the characteristics of the collected data points used in our analyses. Please note that the evaluation of RQ1 and RQ2 is solely based on the data of the control group. Within the data points, we can observe slight differences in numbers. This is due to the participants' randomly assigned experimental conditions. Furthermore, a number of formulated queries had to be excluded from the analysis based on the following criteria:

- A query is unrelated to the given document example. ($n = 13$)
- A query is incomprehensible due to typing errors. ($n = 5$)
- A query is an exact copy of the document title. ($n = 2$)



(a) Control **(b) Experimental**

Figure 2: Proportion of queries with counter-prototypical, prototypical, and no gender mentions per document variation (female, male and non-gendered) and experiment condition. Non-gendered documents show a very low rate of gender mentions in both conditions.

- A query has a structure that is not depicted in our coding scheme, e. g. the query contains a different gender than the one indicated in the given document, like formulating the query *“how to attract women”* for the document *“5 Simple Ways Men Can Increase Their Attractiveness”*. (n = 15)

6 RESULTS AND DISCUSSION

In this section, we describe the results of our online study that investigates to which extent gender bias is reflected in peoples’ formulation of web search queries. Addressing RQ1, in Section 6.2 we explore whether we can find evidence of gender bias in search queries. To answer RQ2, in Section 6.3 we analyse the influence of several personal characteristics on the manifestation of gender bias in search queries. In Section 6.4, we strive to answer RQ3 by probing if an educative text on gender bias in web search affects the participants’ formulation of search queries with respect to explicit gender mentions. Finally, we take a look at people’s attitudes toward societal bias in search engines.

6.1 General Overview

To analyse our research questions, we introduce three evaluation measurements, relative frequency $f^{(i)}$ of an individual’s gender mentions, the collective relative frequency f of gender mentions (i.e., experiment level), and the gender mention gap *MGap*, as formalised in Section 3.2. Inspired by the rationale of the prototype theory, we base these measures on two variables: the frequency of prototypical gender mentions (N_{pm}) and the frequency of counter-prototypical gender mentions (N_{cpm}). To calculate the frequency values, each query was analysed manually and labelled according to prototypical (*pm*), counter-prototypical (*cpm*) or no gender mention (*nm*).

Table 2 gives an example of the coding procedure. Figure 2 shows the distribution of queries cumulated by label and experiment condition. The figure reveals the following information: First, the *non-gendered* document variation shows a very low rate of gender mentions in both conditions, the control group (i.e., $f_{pm} + f_{cpm} = .017$) and the experimental group (i.e., $f_{pm} + f_{cpm} = .012$). This is in line with our expectations, as the *non-gendered* document variation was mainly placed to disguise the gender emphasis of the study.

Table 8: Fraction of gender mentions in queries as a response to counter-prototypical and prototypical document examples. The numbers reveal a significant gender bias with a medium effect (ϕ = .33), which results from the high difference between counter-prototypical and prototypical gender mentions in search queries.

Documents Set	f_{cpm}	f_{pm}	$MGap$	Chi-square test	
				ϕ	$p - value$
All	.86	.43	.43	.33	**<.001*****
Only with Female Stereotypes	.87	.52	.35	.25	**<.001*****
Only with Male Stereotypes	.84	.34	.50	.43	**<.001*****

Consequently, we exclude the *non-gendered* document variation from further analysis of this study. Second, we can see that the fraction of counter-prototypical gender mentions (*cpm*) is much higher than the fraction of gender prototypical gender mentions (*pm*). This is further investigated in RQ1.

6.2 RQ1: Do users replicate gender stereotypes in the formulation of search queries on the web?

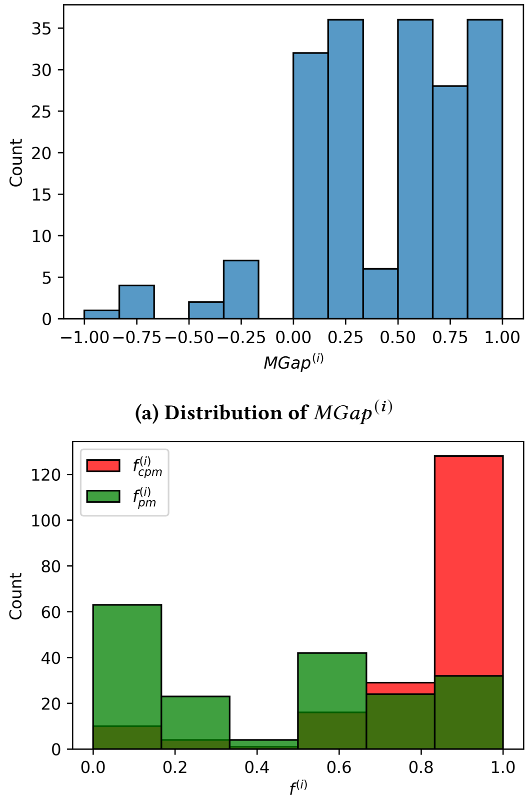
Inspired by the prototype theory [47], this research question builds upon the hypothesis that in a given context, people explicitly mention attributes if they conflict with the prototypical representation of a mental category within this context. In the scope of this study, we confront participants with document examples (i.e., simulated search results) that either conform to a gender stereotype (i.e., resemble a prototypical representation of gender) or do not (i.e., conflict with a prototypical representation of gender). Thus, to answer this research question, we analyze the reflection of stereotypes in the formulation of search queries based on the mention of gender that either agrees with the document stereotype (*pm*) or does not (*cpm*). 1097 data points enter this analysis, depicting the search queries of n = 214 participants of the control group, responding to the document variations *female* and *male* (i.e., independent variable).

Table 8 reports the relative frequency of gender mentions f and the mention Gap (*MGap*) on an experiment level, and compares gender mentions as a response to document examples that conform to the gender stereotype of a domain (prototypical) and document examples that contradict a stereotype (counter-prototypical). The numbers reveal a significant gender bias with a medium effect ($MGap = .43, \phi = .33, p < .001$), which results from the high difference between counter-prototypical and prototypical gender mentions in search queries. Thus, on the experiment level, we can infer that search engine users replicate gender bias in search queries.

Further insight occurs by calculating the results according to the expected stereotypical gender of a document example. Here, we can observe a significant deviation between the response to female and male stereotypes, where f_{pm} for male stereotyped documents (e.g., topic of career) is lower than for female stereotyped documents (e.g., topic of childcare), which results in a stronger effect of gender bias

Table 9: Correlations of gender mentions to personal characteristics calculated as Spearman r correlation. Values with statistical significance ($p < 0.05$) are marked in bold. As provided in Table 6, questions 2 to 6 were posed as self-reflection questions, whereas specifically questions 4 to 6 ask about the participant's confidence (Conf.) in their ability to search, find information, and confidence in search engines in general.

	Question	$f^{(i)}_{cpm}$		$f^{(i)}_{pm}$		$MGap^{(i)}$	
ID	Corresponding Characteristic	Spearman r	p-value	Spearman r	p-value	Spearman r	p-value
Q1	Age	.070	.784	.057	.338	-.020	.783
Q2	Feminism	-.109	.137	-.118	.137	.066	.366
Q3	Political Identity	-.110	.132	**-.162**	**.026***	.093	.205
Q4	Conf. Search Ability	**-.145**	**.047***	-.114	.119	.043	.555
Q5	Conf. Finding Information	-.131	.072	-.037	.614	-.027	.718
Q6	Conf. Search Engines	**-.145**	**.047***	.023	.747	-.100	.171



(a) Distribution of $MGap^{(i)}$

(b) Distribution of $f^{(i)}_{cpm}$ and $f^{(i)}_{pm}$

Figure 3: Distributions of dependent variable metrics per participant.

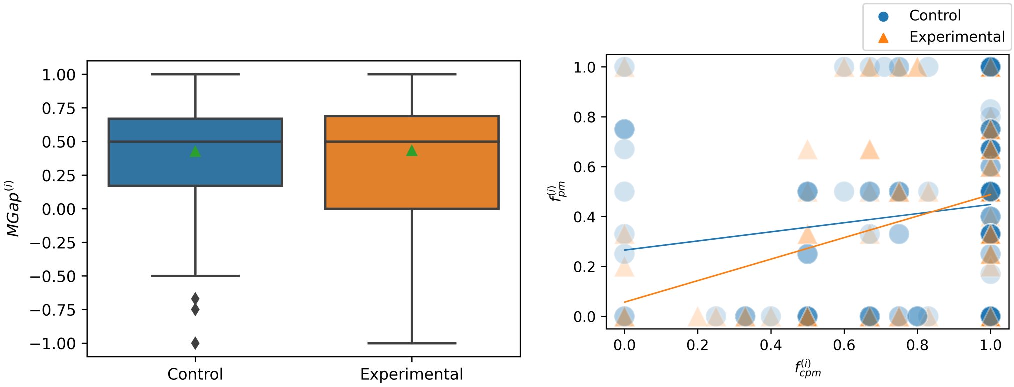
($MGap = .50, \phi = .43$, p<.001). This effect agrees with early work on gender stereotypes, in which [16] showed that the stereotypical man is perceived as more homogeneous than the stereotypical woman, which allows for a greater variance in characteristics. This leads to a stronger expression of male stereotypes, which is accompanied by a higher self-evidence of gender implication in the context of male stereotypes, and consequently, for example, a stronger assumption for a CEO to be male. Considering these results, we hypothesise that the lower frequency of prototypical mentions in documents with male stereotypes is partly due to the higher self-evidence of male stereotype attributions. We further discuss this hypothesis in the subsequent sections and research questions.

Figure 3a shows the distribution of $MGap^{(i)}$, calculated per participant. Results show a variance of .179 among the participants in regard to the dependent variable $MGap^{(i)}$ ($M = .43$, $SD = .42$, $Mdn = .5$). Applying a Wilcoxon Signed-Rank test, we find the relative frequency of prototypical gender mentions ($Mdn = .5$) to be significantly lower than of counter-prototypical gender mentions ($Mdn = 1$), ($T = 845$, $p < .001$). In the following research question, we investigate whether part of the variance in participants' behaviour can be attributed to several personal characteristics.

6.3 RQ2: Does the extent to which users replicate gender stereotypes in search queries depend on personal characteristics ?

In the context of this research question, we assume that certain characteristics of a person favour stereotypical thinking and thus benefit the reflection of gender stereotypes in search queries. To address this, we collected a number of additional variables, such as (i) demographic data (i.e., age, gender identity, educational level) because gender stereotypes depict (outdated) societal norms that are learned and might be subject to change over time and with social context, ii) perceived political orientation (i.e., feminism, political identity) as we expect non-feminists and conservatives to be more attached to traditional gender roles, and iii) the self-assessed confidence in the use of search engines, to get a glimpse on a possible effect of "information search literacy" on the formulation of information need. An overview of the complete set of questions and their coding is given in Table 6.

We compute Spearman's rank-order correlation coefficient to understand the effect of personal characteristics on bias replication. Results for the numerically represented variables are illustrated in Table 9. We observe three weak but significant associations. First, our data indicate a negative correlation ($r = -.162$, $p < .05$) between *Political Identity* and the relative frequency of prototypical gender mentions: the more conservative participants are, the more likely they will mention gender in a search query that conforms



(a) $MGap^{(i)}$ for each condition

(b) $f^{(i)}_{cpm}$ and $f^{(i)}_{pm}$ for each condition

Figure 4: Dependent variable metrics under the two experimental conditions

to a stereotype, e.g., *"how men get to the top"* (see Table 2). Second, we found a negative correlation ($r = -.145$, $p < .05$) between participants' self-assessment of web search literacy (i.e., $Q4$ and $Q6$) and the counter-prototypical mention of gender in search queries ($f^{(i)}_{cp}$). This suggests that people less confident in using search engines are more likely to replicate gender bias in their search queries. This observation contradicts the assumption that gender bias in search queries is intensified by a learning effect, as explicit mention of gender could lead to more accurate search results. With respect to the two categorical variables, the correlation analysis shows no significant effect of *participant gender* ($T = .809$, $p = .419$) or *educational level* ($T = .542$, $p = .588$) on participants' reflection of gender bias in search queries ($MGap^{(i)}$).

Following the results of RQ1, we investigate the influence of gender in greater detail, dividing the results into male- and female-stereotyped domains. On this finer level of granularity, we observe a significantly lower habit ($T = -2.527, p = .012$) of male participants ($M_{f_{pm}} = .28, SD = .39$) than female participants ($M_{f_{pm}} = .48, SD = .44$) in mentioning the male gender in male-stereotyped domains (i.e., annotated as prototypical gender mention). In other words, when compared to women, men are significantly less likely to say e.g. "male CEO". Also, male participants tend to mention the male gender in male-stereotyped domains ($M_{f_{pm}} = .28, SD = .39$) in significantly lower frequency ($T = -3.413, p = .001$) than they do mention the female gender in female-stereotyped domains ($M_{f_{pm}} = .51, SD = .47$). This means that men are significantly more prone to say e.g. "female nurse" than "male CEO". Interestingly, we did not find a similar effect on the behaviour of female participants. These findings support the assumption that men still have strong mental prototypes, e.g. that a CEO is commonly male since they generally see little need to mention their own gender in domains such as management. At the same time, both men nor women do not have such strong mental prototypes of "female nurses". This finding is in line with existing literature on gender stereotypes e.g. [36, 53], that found women to be more likely to experience gender roles and stereotypes over time. In particular, Lopez-Zafra and Garcia-Retamero [36] found that over the last decades, women tended to adopt more masculine traits, such as agency, while men did not tend to adopt more feminine traits, such as commonality. Female stereotypes are thus perceived as more dynamic than male stereotypes [18]. In practice, for example, better education for women and reduced birth rates have led to societal changes that qualify females for occupations with more status and income [19], while men have not shown a similar shift towards domestic or caregiving roles [8]. It is possible that men may be less willing than women to embrace such recent changes in gender roles and stereotypes. This may be because traditional gender roles and stereotypes related to the own gender may be more beneficial to men in terms of their self-conception and social status than traditional female stereotypes are to women. As a result, men may be less likely than women to accept changes in their societal roles and may be more likely to maintain traditional stereotypes about their own gender [32, 50].

6.4 RQ3: Can information on avoiding gender stereotyping raise awareness and mitigate the effect?

The underlying hypothesis of this research question is built on the assumption that a higher awareness of bias in web search and its possible negative implications leads to a shift in people's attitudes, which will manifest in their formulation of search queries. Thus, to investigate the hypothesis, we randomly assigned study participants to one of two groups: the experimental group ($n = 186$), where participants were presented with an educative text that informs about gender bias in web search (see Section 2.2 in the supplementary material), and the control group ($n = 214$) which also received a text of comparable length but only about technicalities of web search without any information on bias.

As depicted in Figure 4a, there is no overall significant difference between the experimental and the control condition ($p = .9$). In fact, values of our bias metric $MGap^{(i)}$ appear to be very similar in the control group ($M = .43$, $Std = .42$, $Mdn = .5$) and the experimental group ($M = .43$, $Std = .40$, $Mdn = .5$). Yet $MGap^{(i)}$ differs in the Mode values with $Md = 1$ (highest bias) and $Md = 0$ (no bias) for the control and the experimental group, respectively. We can

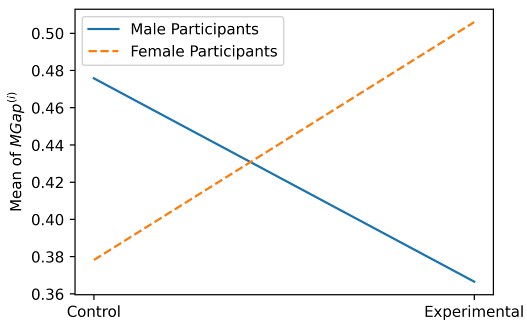


Figure 5: Interaction effect between gender and condition on $MGap^{(i)}$

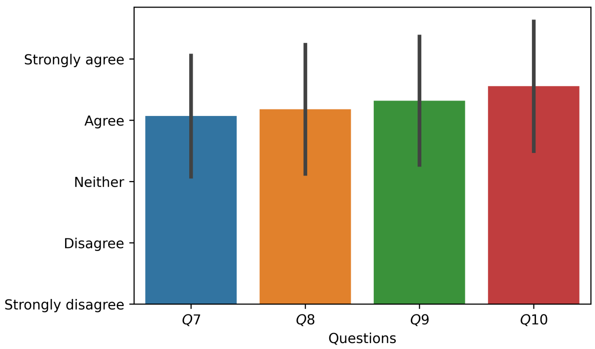


Figure 6: Agreement to questions 7 to 10 according to a Likert scale from *0- strongly disagree* to *4- strongly agree*.

observe a more extensive interquartile range in the experimental group, indicating a higher variance in bias between the subjects. This results from a non-significant difference in prototypical gender mention ($U = 16087.5$, $p = .933$), with $Mdn = .5$ for the control group and $Mdn = .365$ for the experimental group, as depicted in Figure 4b.

Although we do not find a significant main effect of the educative text on the participants' demonstrated gender bias in the overall analysis, we look more closely to see if our intervention affects only the query formulation of participants with specific characteristics. To investigate such potential differential effects as a function of participant gender, a simple moderator analysis is performed using the PROCESS macro [27] for SPSS. The outcome variable for analysis is $MGap^{(i)}$, the predictor variable is the experimental condition (educative text on gender bias vs. control text), and the moderator variable is participant gender (female vs. male). The interaction between the experimental condition and participant gender is found to be statistically significant ($B = -.24, 95\%C.I.(-.4084, -.0658), p < .01$). Conditional effects of the text manipulation for female and male participants show corresponding results. While our educative text shows a borderline significant reduction of the bias value $MGap^{(i)}$ for male users ($B = -.11, 95\%C.I.(-.2302, .0117), p = .08$), it leads to a significant increase and thus to an opposite effect among female users ($B = .13, 95\%C.I.(.0064, .2492), p < .05$), as depicted in Figure 5. A closer look at the data reveals that, while counter-prototypical gender mentions (f_{cpm}) by both user genders remain nearly unchanged by the intervention, the described interaction effect is due to opposite influences on the prototypical gender mentions (f_{pm}) of female versus male users. While women exhibit a lower frequency of prototypical gender mentions in the experimental group ($M = .36, SD = .37$) than in the control group ($M = .45, SD = 36$), it is the other way around for men who reveal a higher frequency of prototypical gender mentions in the experimental group ($M = .49, SD = .40$) than in the control group ($M = .39, SD = .37$). When distinguishing between male- and female-stereotyped domains, the results confirm the findings from RQ1 and RQ2. We do not find a significant effect of the intervention on either male or female participants. However, we can observe a (non-significant) moderating effect on male participants, who, in comparison with the control group, show on average ($M_{f_{cpm}} = .43, SD = .47$) a higher tendency to explicitly mention male-related words in the male-stereotyped domains. This weakens the still significant difference ($T = -2.017, p = .04$) between male participants' gender mentions in female- ($M_{f_{cpm}} = .565, SD = .47$) and male-stereotyped domains. According to prototype theory, the results suggest that men have more robust mental prototypes of, for example, a typical leader than of a nurse, which persist regardless of exposure to explicit information on the topic.

6.5 A glimpse on people's attitude towards bias in online information systems.

Complementary to our three main research questions, we were curious to gain insight into the attitudes and interests of participants towards bias in search engines. To this end, we posed four question items described in detail in Table 6. Participants were asked to answer the questions on the basis of a Likert scale from *0- strongly disagree* to *4- strongly agree*. Figure 6 provides a brief overview of the results.

Responses to $Q7$ describe a common agreement of participants in trusting search engines as a fair source of information ($M = 3.069$, $Std = .989$), where more literate search engine users hold higher levels of trust, which is illustrated in the significant correlation between $Q7$ and the three questions $Q4$ ($r = .211$, $p < .001$), $Q5$ ($r = .267$, $p < .001$) and $Q6$ ($r = .229$, $p < .001$) that depict the participants' self-assessment of search engine literacy. Also, participants stated to be interested in being informed about their own reflection of societal stereotypes in the use of search engines and the reflection of societal stereotypes by the search engine addressed in $Q8$ ($M = 3.181$, $Std = 1.054$) and $Q9$ ($M = 3.319$, $Std = 1.047$), respectively. The more feminist ($Q2$) people are, the more interested they are in being informed about the replication of societal biases, i.e., $Q8$ ($r = .22$, $p < .001$), $Q9$ ($r = .251$, $p < .001$). Finally, most participants ($M = 3.555$, $Std = 1.059$) show a positive intention to be attentive to whether they reflect societal biases in their future use of search engines (see $Q8$). Moreover, the more feminist ($r = .28$, $p < .001$), politically liberal ($r = .15$, $p < .01$) and female ($r = .18$, $p = .001$), the more likely people are to say they will be attentive to stereotype reproduction in future web searches.

7 IMPACT AND FUTURE WORK

More and more efforts are being made to identify, discuss and mitigate the sources of bias in web search (e.g., [35], [43], [52]) and its implications on users' reality perception, decision making and participation in democratic processes (e.g., [22],[13]). While there is a substantial body of research examining the problem from either a human or a technical perspective, a holistic view is needed to capture and understand the interplay between people's cognitive biases, societal biases and algorithmic biases, as these can lead to reinforcement loops in which existing biases - such as societal norms or patterns of discrimination - are reinforced and further manifest themselves in algorithmic decisions and user behaviour (i.e. the user replicates biases) [3].

While there is a substantial amount of literature analysing gender bias in search engine algorithms (see Section 2.2), we consider this work an early effort toward a deeper understanding of gender bias in the formulation of information needs by Internet users. We believe that our work opens up a wide range of future research opportunities and contributes to existing research as follows:

Presenting an approach to measure the reflection of stereotypical thinking in formulating search queries. The prototype theory origins in cognitive psychology and has been investigated in cognitive linguistics for a considerable time [29]. A prominently applied metric (e.g., [26], [46], [1]) to measure the effects of categorizations is known as feature listing or property generation task. More recent research showed that if individuals can not resemble a construct by an existing mental category, a combination of categories is simulated, resulting in the explicit naming of new properties, e.g. Laughing cat, Rolled-up lawn [55]. In this paper, we expand on the insights from cognitive research and successfully show how this can be used to measure gender bias in the formulation of web search queries, i.e., by relating concepts and measures from prototype theory to stereotypically primed study material (i.e., document examples). Since the negative effects of stereotypical thinking are not limited to gender, we argue that this method can also be used to measure the reflection of other negative stereotypes, such as those related to racial or religious prejudice, provided suitable study material is available.

Following the argument of transferability, we understand that the binary classification of gender does not conform to the current state of gender research and, thus, might support outdated societal norms. However, our study setup is confined to the representation of female and male stereotypes. This is a simplification that has been made because, in the available experimental material, gender stereotypes are usually categorised in traditional binary gender roles. Another limitation we recognise is the restriction of the study to UK participants, while gender stereotypes are known to differ with cultural context e.g., [17]. Also, there is a considerable body of research investigating the impact of culture (e.g., [38, 54]), language and language skills (e.g., [2, 14, 25]) on information behaviour. While this study was not designed to capture cultural differences, future research endeavours could expand the scope to domains that capture a greater diversity of gender and context representations.

Providing significant evidence of the reflection of gender bias in the formulation of search queries. In RQ1, we show that given similar analysing tasks, people are significantly more prone to explicitly mention gender in formulating search queries as a response to non-stereotypical document contents. We believe that the proof of users' reflection of gender bias in search query formulation itself is a very relevant contribution to the community that opens up a variety of future research questions spanning from topics around the interaction of users and search engines to methods of creating awareness with intelligent user interfaces and real-time interventions. Most prominently, we are interested in better understanding the interaction dynamics between user behaviour, search engine accuracy and algorithmic bias. First, the question arises whether we actually measure people's bias or, rather, their information search skills, as the phrasing of search queries might be trained through frequent search engine usage. Results of RQ2 (see 6.3) argue against this, as they suggest that people who are less confident in using search engines are more likely to repeat gender bias in their search queries. However, this is only due to a weak correlation and based on the participants' self-assessment of search engine literacy. A more extensive experiment specifically designed to answer this research question based on a standardised skill assessment of information literacy is pending. Technical simulations, on the other hand, could investigate the effect of cumulated user bias (i.e., creating context through the co-occurrence of words) on search engine bias.

Brief insights on the use of interventions to raise awareness and mitigate biases in web search. The results of our intervention study do not show a significant main effect of our educative stimulus text on user behaviour. However, notable is the differential impact of our experimental manipulation on male and female participants, which seems to be opposing. Why this pattern occurs and how bias interventions need to be designed to be effective and reflect different users' needs remain open questions. The question of what such user interfaces could look like and how they could be integrated into existing search engines is at least as big. Only the request for more information about the replication of gender stereotypes in search engines and the role of users themselves in this is unambiguous (see Section 6.5), which ultimately also is a mission for the human-computer interaction community to identify ways of integrating such information into search engine interfaces.

Implications for the design and development of search engines. Our findings could inform the design and implementation of future search engines, both from a system-centric perspective and from a user-centric perspective. The former could be addressed by introducing a pre-processing step that detects and analyses gender-specific contents in search queries and — if such content could lead to discrimination on the grounds of gender — adjusts or extends the query (e.g., via query reformulation or expansion techniques). In the case of such intervention, this process needs to be transparent and communicated to the user in appropriate ways, for instance, by clearly marking the results in the retrieved document lists that have been added or omitted due to such pre-processing steps. In addition, our results on the effect of providing educative text about gender bias in web search, and particularly the different interactions of the users across the genders, support the possible interface designs of bias-aware search engines. We believe integrating such interventions/information into the search engine's user interface, accompanied with concrete and easier-to-grasp examples tailored to the user's gender, could help raise their awareness and understanding of the implications of gender bias in search queries. This can further be realised, e.g., by means of showing counterfactual search results: what would the results be if the gender inclination in the query had changed?

8 CONCLUSION

In this work, we analysed the search query formulation of native English-speaking people concerning the replication of gender stereotypes. Participants were asked to formulate a search query, given a particular search result (i.e., a heading and a preview of a document, as presented on the main page of standard search engines). To this end, we prepared a dataset with $n = 8 * 3$ search results (i.e., document examples), where each document was presented in a female, male, and gender-neutral version. Following prototype theory, we defined a disproportionate mention of a gender that does not conform to the stereotypical gender expectation of the domain as a replication of gender bias. In a first pilot study, we were able to identify tendencies indicating the repetition of stereotypes in different domains (e.g., childcare as a domain traditionally stereotyped as female or career as a domain traditionally stereotyped as male). We were also able to select the most comprehensible documents to be used in the main study. For our main experiment, we recruited 400 UK residents via the crowdsourcing platform Prolific who read either an educational text about gender bias in search engines or a control text without bias-related information before completing eight query generation tasks. While the intervention did not result in a significant main effect (though opposing effects emerged for female and male users), overall, our results showed significant evidence for the prevalence of gender biases in search query formulation. In the interest of equal opportunities for all genders, most of our study participants emphasised the need to address the reproduction of outdated stereotypes in our daily interactions with search engines through a clear demand for more information about the origins and avoidance of biases in search engine use. In summary, we see our work as a first step in understanding potential bias-enhancing feedback loops between user input and search engine algorithms. Ultimately, we also hope that this research will stimulate further discussion on how users can be informed about bias-mitigating query strategies during their interactions with search engines—a challenge that calls for expertise from the research community of Human-Computer Interaction.

ACKNOWLEDGMENTS

This work was supported by the H2020 project Triple (GA:693092) and the "DDAI" COMET Module within the COMET – Competence Centers for Excel- lent Technologies Programme, funded by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), the Austrian Federal Ministry for Digital and Eco- nomic Affairs (bmdw), the Austrian Research Promotion Agency (FFG), the province of Styria (SFG) and partners from industry and academia. The COMET Programme is managed by FFG. Furthermore, it was supported by the Austrian Science Fund (FWF): P33526 and DFH-23; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grants LIT-2020-9-SEE-113 and LIT-2021-YOU-215.

REFERENCES

[1] Jean Aitchison. 1992. Good birds, better birds and amazing birds: The development of prototypes. In *Vocabulary and applied linguistics*. Springer, Basingstoke, UK, 71–84.

[2] Asma Al-Wreikat, Pauline Rafferty, and Allen Foster. 2015. Cross-language information seeking behaviour English vs Arabic. *Library Review* 64, 6/7 (2015), 446–467.

[3] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (may 2018), 54–61. https://doi.org/10.1145/3209581

[4] Ricardo Baeza-Yates. 2020. Bias in Search and Recommender Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 2. https://doi.org/10.1145/3383313.3418435

[5] David Bakan. 1966. *The duality of human existence: An essay on psychology and religion.* Rand McNally, Chicago, Illinois, US.

[6] Rich Barlow. 2014. *BU Research: A Riddle Reveals Depth of Gender Bias.* Boston University. Retrieved Septemeber 9, 2022 from https://www.bu.edu/articles/2014/bu-research-riddle-reveals-the-depth-of-gender-bias/

[7] Deborah Belle, Ashley B Tartarilla, Mikaela Wapman, Marisa Schlieber, and Andrea E Mercurio. 2021. "I Can't Operate, that Boy Is my Son!": Gender Schemas and a Classic Riddle. *Sex Roles* 85, 3 (2021), 161–171.

[8] Suzanne M Bianchi, John P Robinson, and Melissa A Milke. 2006. *The changing rhythms of American family life.* Russell Sage Foundation, New York.

[9] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2021. On the Orthogonality of Bias and Utility in Ad Hoc Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1748–1752. https://doi.org/10.1145/3404835.3463110

[10] Amin Bigdeli, Negar Arabzadeh, Shirin Seyedsalehi, Morteza Zihayat, and Ebrahim Bagheri. 2022. A Light-Weight Strategy for Restraining Gender Biases in Neural Rankers. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 47–55.

[11] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring Gender Biases in Information Retrieval Relevance Judgement Datasets. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 216–224.

[12] Malte Bonart, Anastasiia Samokhina, Gernot Heisenberg, and Philipp Schaer. 2019. An investigation of biases in web search engine query suggestions. *Online Information Review* 44, 2 (Dec 2019), 365–381.

[13] Engin Bozdag and Jeroen Van Den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and information technology* 17, 4 (2015), 249–265.

[14] David Brazier and Morgan Harvey. 2017. E-government and the digital divide: a study of English-as-a-second-language users' information behaviour. In *European Conference on Information Retrieval.* Springer, Aberdeen, UK, 266–277.

[15] Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*, danah boyd and Jamie H. Morgenstern (Eds.). ACM, Atlanta, GA, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[16] Kay Deaux, Ward Winton, Maureen Crowley, and Laurie L. Lewis. 1985. Level of Categorization and Content of Gender Stereotypes. *Social Cognition* 3, 2 (06 1985), 145–167. https://www.proquest.com/scholarly-journals/level-categorization-content-gender-stereotypes/docview/848854828/se-2 Copyright - © 1985 Guilford Publications Inc; Zuletzt aktualisiert - 2018-10-16; CODEN - SOCOEE.

[17] M. Désert and Leyens J. P. 2006. Social comparison across cultures : Gender stereotypes in high and low power distance cultures. In *Social comparison and Social Psychology : Understanding cognition, intergroup relations and culture.* Cambridge : Cambridge University Press., New York, 303–317. https://hal.science/hal-00115989 In S. Guimond (Ed.)..

[18] Amanda B Diekman and Alice H Eagly. 2000. Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and social psychology bulletin* 26, 10 (2000), 1171–1188.

[19] Alice H Eagly, Wendy Wood, and Amanda B Diekman. 2000. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender* 12, 174 (2000), 9781410605245–12.

[20] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Retrieval and Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1403–1404. https://doi.org/10.1145/3331184.3331380

[21] Naomi Ellemers et al. 2018. Gender stereotypes. *Annual review of psychology* 69 (2018), 275–298.

[22] 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.

[23] 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. https://doi.org/10.1016/j.ipm.2020.102377

[24] Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal* 24, 2 (2021), 85–113.

[25] KT Goodall, LA Newman, and PR Ward. 2014. Improving access to health information for older migrants by using grounded theory and social network analysis to understand their information behaviour and digital technology use. *European journal of cancer care* 23, 6 (2014), 728–738.

[26] James A Hampton. 1979. Polymorphous concepts in semantic memory. *Journal of verbal learning and verbal behavior* 18, 4 (1979), 441–461.

[27] Andrew F Hayes. 2017. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford publications, New York.

[28] Tanja Hentschel, Madeline E Heilman, and Claudia V Peus. 2019. The multiple dimensions of gender stereotypes: A current look at men's and women's characterizations of others and themselves. *Frontiers in psychology* 10 (2019), 11.

[29] RK Johnson. 1985. Prototype theory, cognitive linguistics and pedagogical grammar. *Working Papers in Linguistics and Language Training* 8 (1985), 12–24.

[30] 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* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[31] Mary E Kite, Kay Deaux, and Elizabeth L Haines. 2008. *Gender stereotypes.* Praeger Publishers/Greenwood Publishing Group, New York.

[32] Eric D Knowles and Brian S Lowery. 2012. Meritocracy, self-concerns, and Whites' denial of racial inequity. *Self and Identity* 11, 2 (2012), 202–222.

[33] 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. https://doi.org/10.48550/ARXIV.2201.07754

[34] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?. In *Proceedings of the European Conference on Information Retrieval, Workshop on Algorithmic Bias in Search and Recommendation (ECIR-BIAS 2022)*. Springer, Cham, 104–116.

[35] 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 417–432. https://doi.org/10.1145/2998181.2998321

[36] Esther Lopez-Zafra and Rocio Garcia-Retamero. 2012. Do gender stereotypes change? The dynamic of gender stereotypes in Spain. *Journal of Gender Studies* 21, 2 (2012), 169–183.

[37] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing Managment (IP&M)* 58, 5 (2021), 102666. https://doi.org/10.1016/j.ipm.2021.102666

[38] Hester WJ Meyer. 2009. The influence of information behaviour on information sharing across cultural boundaries in development contexts. *Information Research: An International Electronic Journal* 14, 1 (2009), 393.

[39] Cindy Faith Miller, Leah E Lurye, Kristina M Zosuls, and Diane N Ruble. 2009. Accessibility of gender stereotype domains: Developmental and gender differences in children. *Sex roles* 60, 11 (2009), 870–881.

[40] 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6620–6631. https://doi.org/10.1145/3025453.3025727

[41] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. 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* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 933–936. https://doi.org/10.1145/3209978.3210094

[42] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google We Trust: Usersâ€™ Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication* 12, 3 (04 2007), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x arXiv:https://academic.oup.com/jcmc/article-pdf/12/3/801/22316463/jjcmcom0801.pdf

[43] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. 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* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3404835.3462949

[44] 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* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 2065–2068. https://doi.org/10.1145/3397271.3401280

[45] Eleanor Rosch. 1999. Reclaiming Concepts. *Journal of Consciousness Studies* 6, 11-12 (1999), 61–77.

[46] Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology* 7, 4 (1975), 573–605.

[47] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology* 8, 3 (1976), 382–439.

[48] Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology* 4, 3 (1973), 328–350.

[49] Jeffrey W Sherman. 1996. Development and mental representation of stereotypes. *Journal of Personality and Social Psychology* 70, 6 (1996), 1126.

[50] Jim Sidanius and Felicia Pratto. 1999. *Social Dominance: An Intergroup Theory of Social Hierarchy and Oppression.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139175043

[51] Fred L Strodtbeck and Richard D Mann. 1956. Sex role differentiation in jury deliberations. *Sociometry* 19, 1 (1956), 3–11.

[52] Fons Wijnhoven and Jeanna Van Haren. 2021. Search engine gender bias. *Frontiers in big Data* 4 (2021), 29.

[53] Annett Wilde and Amanda B Diekman. 2005. Cross-cultural similarities and differences in dynamic stereotypes: A comparison between Germany and the United States. *Psychology of Women Quarterly* 29, 2 (2005), 188–196.

[54] Tom D Wilson. 1997. Information behaviour: an interdisciplinary perspective. *Information processing & management* 33, 4 (1997), 551–572.

[55] Ling-ling Wu and Lawrence W Barsalou. 2009. Perceptual simulation in conceptual combination: Evidence from property generation. *Acta psychologica* 132, 2 (2009), 173–189.

[56] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2532–2538. https://doi.org/10.1145/3477495.3531891