
Proceedings Paper:
Otterbacher, J., Bates, J. and Clough, P.D. (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. 2017 CHI Conference on Human Factors in Computing Systems, 06-11 May 2017, Colorado Convention Center, Denver, CO. Association for Computing Machinery , pp. 6620-6631. ISBN 978-1-4503-4655-9

https://doi.org/10.1145/3025453.3025727

Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results

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

Jo Bates
Information School
University of Sheffield, UK
jo.bates@sheffield.ac.uk

Paul Clough
Information School
University of Sheffield, UK
p.d.clough@sheffield.ac.uk

ABSTRACT
There is much concern about algorithms that underlie information services and the view of the world they present. We develop a novel method for examining the content and strength of gender stereotypes in image search, inspired by the trait adjective checklist method. We compare the gender distribution in photos retrieved by Bing for the query "person" and for queries based on 68 character traits (e.g., "intelligent person") in four regional markets. Photos of men are more often retrieved for "person," as compared to women. As predicted, photos of women are more often retrieved for *warm* traits (e.g., "emotional") whereas *agentic* traits (e.g., "rational") are represented by photos of men. A backlash effect, where stereotype-incongruent individuals are penalized, is observed. However, backlash is more prevalent for "competent women" than "warm men." Results underline the need to understand how and why biases enter search algorithms and at which stages of the engineering process.

Author Keywords
Algorithmic bias; "Big Two" dimensions of social perception; gender stereotypes; image search.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Networked information services, such as search engines, recommendation systems, and social media "feeds" make extensive use of algorithmic processes to guide users to interesting content, filtering out that which is likely of less value. Given the ever-growing volume of content and the diversity of available sources, such algorithms are a necessary part of our information eco-system. However, there is little doubt that they influence our view of the world, quite literally mediating social relations and our participation in public life [20]. Even when users are intimately familiar with a system, they are often unaware that algorithms filter their access to information [14] and users hold beliefs about algorithms, which, true or not, influence how they use systems [39].

Machines running algorithmic processes have become the new gatekeepers, largely determining what and whom we see, and do not see [8]. Given the power that algorithms exert, researchers must scrutinize these processes, their potential biases and social impact; in other words, we must work toward "algorithmic accountability" [11] and "algorithmic transparency" [10].

Bias in image search: perpetuating social stereotypes

That information systems bring a *slant* to the manner in which they present information is accepted, albeit not well understood. Given the hyper-personalized nature of modern information services, there is no "gold standard" against which we might compare what a given user sees [20]. Writing well before the rise of "Big Data" and online service giants, Friedman and Nissenbaum [18] explained that systems are *biased* when two conditions hold: 1) results are slanted in unfair discrimination against particular persons or groups, and 2) that discrimination is systematic, i.e., not just occurring in isolated cases. Our work focuses on one type of bias in search algorithms: the *perpetuation of gender stereotypes* in image search.

As a relatively mature technology, search continues to be the primary means to information access in networked systems (e.g., digital libraries) and the Web. The need to understand the values that search algorithms convey through the results they provide has not diminished over the years [1, 35]. In fact, as the use of complex personalization mechanisms increases, this need becomes more salient, given users' tendencies to approach complex systems at "interface value" [45] and the trust users place in search engines [37].

Search can perpetuate stereotypes in various ways. Baker and Potts [5] studied Google's auto-complete feature, designed to help users formulate queries by suggesting terms and phrases. They found that auto-complete associates questions about appearance, behavior and attitudes to particular social groups, stereotyping some more negatively than others. Kay and colleagues [26] considered the perpetuation of gender stereotypes through Google image searches on queries surrounding

professions. They showed that gender distributions in images retrieved for professions (e.g., doctor vs. nurse) reflect prevalent stereotypes and in fact, are exaggerated (e.g., the proportion of images of women retrieved for the query "doctor" is less than expected, compared to labor statistics). They also demonstrated the power of search on users' perceptions; viewing results for a given profession sways users' estimations of the actual gender distribution.

Recent coverage in the mass media also indicates that the public is actively questioning the role of search algorithms in reinforcing stereotypes. For instance, in April 2016, Twitter user 'BonKamona' discovered that in Google searches for "unprofessional" versus "professional hairstyles for work," the resulting images depicted primarily women of color and white women, respectively[1]. This led to a public discussion of search bias, its origins and role in the wider social landscape[2].

In such discussions, users inevitably end up questioning whether the algorithm could be considered "sexist" or "racist." In other words, users often apply social expectations to the behavior of the search tool, and when confronted with unexpected or undesirable results, they attribute human characteristics to them [34]. Inspired by these discussions, as well as Kay and colleagues' [26] call for further empirical approaches to studying the manner in which social groups are presented in online media, we adapt a method used for decades by social psychologists to evaluate the content and strength of stereotypes that people hold, to the digital context.

We develop an automated technique based on the "trait adjective checklist method." Previous studies of search engines' portrayal of people have either been conducted manually on a small set of queries (e.g., [5]), or within a particular context (e.g., professions [26]). Our method allows us to study more generalized stereotypes based on a large set of character traits (e.g., who is an "emotional person"?) and across geographic regions. As we will show, our results are consistent across regions and empirically demonstrate that image search results largely reinforce traditional gender stereotypes: images of woman are associated with warm character traits (e.g., expressive, emotional) whereas images of men represent agentic traits (e.g., competent, intelligent).

We also show that images of individuals who do not conform to these stereotypes exhibit a "backlash" effect; for instance, "competent women" are less likely to be portrayed in a positive way, in terms of an increase in perceived status or power, as compared to "competent men." We demonstrate the importance for designers, particularly those who build applications on top of search APIs, to consider the stereotypes conveyed through search results, and lay the groundwork for developing methods for the automatic detection of bias in image search.

[1]https://twitter.com/BonKamona/status/717457819864272896

[2]https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-

BACKGROUND AND RESEARCH QUESTIONS

Measuring stereotype content and strength

Our well-being relies on our ability to form impressions of others from many walks of life, both accurately and efficiently [9]. To this end, *social stereotypes* can be described as socio-cognitive heuristics that help us make sense of others. As early as the 1930s, social psychologists began to systematically study stereotype content and degree of consensus, to uncover the beliefs people hold about different social groups.

Katz and Braly's [25] study of racial stereotypes held by Princeton University students is arguably one of the most influential studies, having been replicated by many, most notably in the *Princeton Trilogy* studies. In these experiments, students were presented with a list of character traits and indicated which best describe different ethnic and racial groups. Consensus between participants was measured, to examine who endorses which stereotypes and for which groups. In addition, as in the Princeton Trilogy follow-ups [19, 24], the trait checklist method has been used to study stereotype changes over time. One point of contest with the method is the choice of traits to incorporate in the list. For instance, Madon and colleagues [30] argued that Katz and Braly's list needed updating, and added an additional 300 attributes, noting that this significantly increased the task's difficulty.

The "Big Two": warmth and competence

Given the emergence of new theories of social perception, the above problem is now much less of a concern. Cuddy and colleagues [9] explain that, regardless of the traits that study participants use to describe others, two underlying dimensions consistently emerge: *warmth* (also known as "communion") and *competence* (also called "agency"). Just as personality researchers have developed a "Big Five," the existence of a "Big Two" is now well accepted amongst researchers of social perception [2].

The *warmth dimension*, comprised by traits such as caring, moral, and kind, allows others to gauge one's intentions toward them. In contrast, the *competence dimension*, comprising traits such as intelligent, assertive, and creative, is an indication of one's ability to carry out his or her intentions. Fiske and colleagues [17] explain that competence has to do with perceived status (i.e., whether someone is perceived to be accomplished or capable of achievement); while warmth has to do with perceived competition (i.e., someone warm is not seen as posing a threat, while someone lacking in warmth is threatening). They argue that stereotypes are captured by combinations of the two dimensions. For instance, elderly

people are often seen as being a subordinate, non-competitive group (i.e., low competence, high warmth), while East Asians are stereotypically viewed in the West as high-status but competitive (i.e., high competence, low warmth). Although stereotypes serve as sense-making aids, they can also have negative consequences, when our cognitions (stereotypes) result in affective responses (emotional prejudices) that lead to negative behaviors, such as discrimination [9].

Returning to the problem of the choice of traits to include in a checklist, Abele and colleagues [3] conducted a cross-lingual study in five countries (Belgium, Germany, Italy, Poland and USA). They aimed to develop a standardized operationalization of the Big Two dimensions and derived a set of trait adjectives that do not differ with respect to valence or frequency of occurrence across languages. We use their list of traits in our work.

Gender stereotypes and backlash

Stereotypes represent normative expectancies; they describe people's beliefs about what a social group is or should be [22]. Research on descriptive gender stereotypes has found that women are perceived as being characteristically warm/communal, whereas men are perceived as being agentic/competent [6]. However, gender stereotypes also tend to have a strong prescriptive component – describing how women and men should or *should not* be [13] – and are therefore extremely influential. Furthermore, while descriptive stereotypes have changed over time to reflect women's more agentic roles (e.g., increasing participation in leadership), prescriptive stereotypes have largely remained constant [12]. In other words, beliefs about what women and men should be have remained rather traditional.

One mechanism that aids the perpetuation of stereotypes in society is *backlash*; people who are stereotype-incongruent, who do not conform to prevailing stereotypes, often experience negative consequences (i.e., social and/or economic reprisals). Backlash against agentic women has often been studied in organizational contexts. For instance, Rudman and Glick [40] found a bias against agentic women in hiring decisions. They explained that women are expected to be "warm" and thus, must soften their agentic traits. Rudman and Phelan [41] summed up the problem as a two-part impression-management dilemma. On the one hand, since women are generally considered less agentic, a woman desiring to excel as a leader must present herself as an atypical woman. However, to prevent backlash resulting from the prescriptive stereotype of women as warm (i.e., nonthreatening), she must temper her agency. The researchers also note that backlash against warm men in the workplace is common; compared to men with more agentic traits, those perceived as being warm are consistently rated as less suitable for leadership. More specifically, backlash against atypical men is likely to occur when they are perceived as being "too modest" (i.e., possessing warm traits linked to low status, and lacking in agentic traits of high status) [32].

Backlash effects have been extensively studied in the media, and it is widely observed that representation of gender impacts the social reproduction of inequalities in other areas of life. Templin [43], using then First Lady of the U.S., Hillary Clinton, as a case study, considered backlash against professional women in cartoons. She noted the extensive use of stereotypes and clichés in her corpus of cartoons depicting Clinton, as well as gender reversals, domestic imagery and sexualization.

Considering the impact of representation of women politicians, Bligh et al.'s [7] study concludes that the media have significant influence on voters' judgments regarding the likability and competence (and thus, electability) of politicians. Crucially, they observe that media discourses focusing on gender role incongruence impact negatively upon women politicians' likeability, and thus, generate a "double bind" for candidates who need to communicate both competence and likeability. In a similar vein, Mudrick's [33] study of sportscasters cited a double standard preventing women from gaining acceptance. While women sportscasters are often seen as likeable, and are increasingly being included in the profession, they are seldom given the opportunity to gain credibility, which requires the use of authoritative communication tactics seen as gender-congruent for men.

Finally, attesting to the significance of media portrayal of women on women's self-confidence, Simon and Hoyt [42] found that in an experimental setting, women exposed to counter-stereotypical images depicting women in leadership roles reported less negative self-perceptions and greater leadership aspirations, as compared to those exposed to stereotype-confirming images.

Research questions

We adapt the trait adjective checklist method to the study of search engine bias. In particular, we investigate the perpetuation of gender stereotypes in image search results. We address the following three research questions:

Representation bias (RQ1): In a search for images of a "person," which genders are depicted in the results?

Stereotype content (RQ2): Do search engine results reflect the same gender stereotypes observed by social psychologists? Which traits are most characteristic of women versus men and which are gender-neutral?

Backlash effects (RQ3): Do we observe backlash effects in images that depict stereotype-incongruent individuals?

METHODOLOGY

Data collection

We used the Bing Image Search API[3] from Microsoft Cognitive Services to build a corpus of image search results. Bing is an exemplar Web search engine, having a market share second only to Google[4]. However, its API is arguably more flexible for researchers looking to build a corpus of search results, as it allows requests of large numbers of results, as well as control of search parameters (e.g., desired language and regional server).

We requested the most relevant 1,000 images for the general query "person," for each of four search markets - UK, US, India (IN) and South Africa (ZA) - as our intent was to study relatively large, Anglophone markets. Next, for each of the 68 trait adjectives listed in Tables 5 to 8, we submitted the query "X person" to the API. We requested the most relevant 1,000 images from each regional server, using the API's search market parameter. For the query "person," Bing provided 1,000 images for all four regions; however, for the character trait queries, Bing often provided slightly fewer images. Over the 272 queries (68 traits * 4 regions), the mean/median number of images returned was 979/990.

	1-10	1-20	1-100	1-500	1-1,000
UK-US	7	14	67	249	451.5
UK-IN	7	13	50.5	180	311.5
UK-ZA	10	19	97	421.5	797.5
US-IN	7	14	67	257	459.5
US-ZA	7	14	65	236	420
IN-ZA	7	12.5	50.5	171	288

Table 1: Median # images in common across 68 trait queries.

Table 1 shows the pairwise overlap between regions, detailing the median number of images in common across the 68 trait queries by rank (i.e., by first 10 images retrieved to all 1,000 images retrieved). Like Hannak and colleagues [23], we find that results vary by region, but that top results are similar. For instance, on average for all pairs, other than UK-ZA, seven of the top-10 results are identical; however, as we consider more results, there is more variation. In comparing the results of UK versus those of ZA, we observe a very high degree of overlap. For this particular pair of regions, while there are more differences as additional results are considered, there is nearly 80% overlap in the set of 1,000 images.

[3]https://www.microsoft.com/cognitive-services/en-us/bing-image-search-api

[4]https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

Pilot study: recognizing photos and gender

Since we needed to analyze large sets of images, we developed a method to automatically infer, from an image retrieved from Bing, the genders of depicted person(s). We recognize that our present treatment of gender is as a binary construct. As will be shown, the task of inferring depicted person's genders is not trivial and therefore, a more sophisticated treatment of gender was not possible.

To understand the complexity of the task, we conducted a pilot study via Crowdflower[5]. Our procedure was approved by the University of Sheffield Information School's Ethics Committee. Anglophone participants were recruited from our four regions of interest and passed a quality control task consisting of 10 test images. We collected three responses per image. Participants were compensated $0.20 for every five images, which on average took two minutes. Crowdflower's contributor survey indicated participants were satisfied with our instructions, task and pay, with a satisfaction score of 4/5.

Using 1,000 images retrieved for the "person" query from Bing's UK-English language search market, we designed the task using Crowdflower's image classification template. Specifically, participants were presented with an image and answered two questions: 1) Is the image a photograph, a sketch or illustration, or some other type of image? 2) Does the image depict only women/girls, only boys/men, a mixed gender group, gender ambiguous person(s), or no persons at all? Participants also indicated if the image did not display properly.

Table 2 presents the distribution of "person" images that are photos, sketches/illustrations, or other. The interjudge agreement (IJA) column shows the mean agreement between all three annotators. More than half of the images are photos, and there is very high agreement with respect to which is a photo versus a sketch. However, there is lower agreement with respect to "other" types of images. Manual inspection revealed that a wide variety of images are returned for the query "person," including quotes or jokes that feature primarily words. Annotators had trouble classifying such images as being "sketches" or "other."

	# Images	IJA
Photos	576	0.97
Sketches	346	0.96
Other	22	0.74
Not accessible	56	1.00

Table 2: Manual recognition of photos and sketches.

Table 3 presents the distribution of the genders of people in photos versus sketches. Most of the photos and sketches in our "person" corpus depict people; only 1% of

[5] https://www.crowdflower.com

photos and 4% of sketches do not depict people. Next, it can be observed that in more than half of the sketches, the gender of depicted persons cannot be determined ("unknown"). The quintessential example is a stick figure that has too little detail to convey gender. Clearly, both in photos as well as sketches, there are more images of men/boys as compared to girls/women. Finally, the mean agreement between judges on the five-way classification of gender was 0.94 in photos, and 0.91 in sketches.

	Women	Men	Mixed	Un-known	No person
Photo	0.27	0.55	0.10	0.07	0.01
Sketch	0.08	0.28	0.05	0.55	0.04

Table 3: Gender distribution by image type.

Based on the pilot, we decided to concentrate on inferring the gender of people in *photographic images* retrieved by Bing. Even for annotators, determining the gender of persons depicted in sketches is difficult. Therefore, we leave this challenge for future work, and focus on analyzing the most realistic depictions of people in image search results. The mean/median number of photos retrieved across the 68 trait-queries is: 306.6/311.5 (UK), 303.6/295.5 (US), 305.7/303 (IN) and 308.0/310.5 (ZA).

Automatic detection of gender in photos

We used the Clarifai API[6] to glean information on the content of each image retrieved. There are many image recognition tools on the market, each with its own limitations concerning the types (format, size) of images it will process. In testing various tools, our experience has been that those specific to facial recognition are quite sensitive, requiring depicted people to be positioned in a particular manner. Such restrictions would severely limit our ability to process images retrieved from the Web.

While Clarifai is a general image recognition tool, it provided good coverage, processing 95% of the images in our corpus. Clarifai uses deep learning to infer the content of a given image. For an input image, the 20 most likely concept tags are provided, along with their associated probabilities. As a proprietary tool, details of its algorithm are not available. However, an earlier version of Clarifai won the 2013 ImageNet Large Scale Visual Recognition Challenge task[7], in which algorithms must produce a list of relevant object category labels for an image from a set of labels. An error rate of less than 12% was reported.

We use the Clarifai tag, "portrait," to disambiguate photos versus other images. On our "person" corpus, we estimate Clarifai's recall to be 75% and precision to be 91%. In other words, 91% of images in our pilot corpus labeled as "portraits" are confirmed as being photographs by our human annotators. Clarifai identified three-quarters of the photos identified by our annotators as "portraits."

Clarifai's tags are then subjected to analysis via the Linguistic Inquiry and Wordcount tool (LIWC) [38], which has gained wide acceptance among researchers who need to infer meaning from short social media texts, as well as image metadata [36]. We rely on two LIWC categories, *female* and *male references*. These categories consist of 124 and 116 words, respectively, such as "girl" and "mom" for females, and "boy" and "dad" for males. LIWC provides a score representing the percentage of tags referencing females and males. We use these scores to label photos; if the *female* score is positive and the *male* score zero, we label the photo as depicting women/girls, and vice versa. Photos not labeled as depicting women/girls or men/boys are labeled "other." "Other" photos include those depicting mixed-gender groups and individuals for whom gender is ambiguous. Very few "other" photos (~1%) depict no persons at all. Figure 1 depicts the entire process with an example.

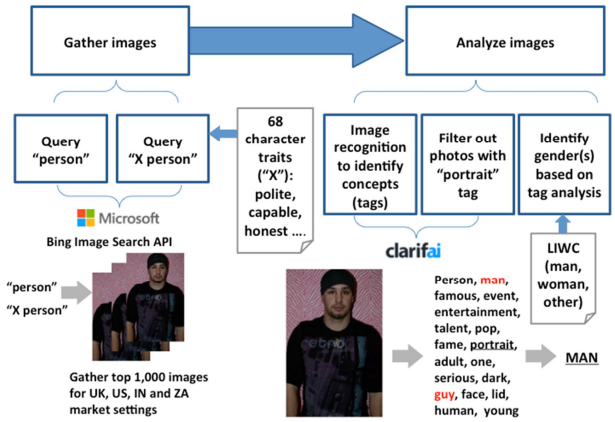

Figure 1: Image collection and gender detection process.

Table 4 presents precision, recall and F_1-measure for classification on the 473 "person" images labeled as photos. The method has better precision than recall; while Clarifai missed some photos identified by annotators, the vast majority labeled as such are indeed photos. We confirm high precision for the women and men categories. The method does not introduce many false positives for detecting photos of women and men, and misses only slightly more images of women as compared to men (recall of 0.60 versus 0.67).

	N	Precision	Recall	F_1
Women	130	0.89	0.60	0.717
Men	282	0.95	0.67	0.786
Other	61	0.68	0.82	0.743

Table 4: Performance on gender classification in photos.

[6]https://developer.clarifai.com/guide/

[7]http://www.image-net.org/challenges/LSVRC/2013/results.php#cls

ANALYSIS

Who represents a "person"?

To answer RQ1, which concerns representational bias, we considered the gender distribution across photos for the query "person." Figure 2 shows gender(s) in photos retrieved for the first 100, 200, 500 and 1,000 results. A Chi-square test of independence within each set reveals no significant differences between regions. For all regions, the gender gap is largest in the top-ranked results - the set of images users are most likely to view [37]. This gap progressively reduces as we consider additional images, reaching its minimal point of 42% men / 18% women. In other words, among photos presented to users looking for images of a "person," there are over twice as many photos of men/boys as compared to women/girls.

Here, the question of a baseline to which we might compare these results arises. For instance, in Kay and colleagues' work [26], labor statistics were used as a baseline to which they compared observed gender distributions in image searches for professions. In our work, which concerns searches of a general nature, there is no clear choice of baseline. For this reason, we focus on presenting a reproducible method for examining gender distributions in retrieved photos, such that we can expose potential biases to users and developers.

We avoid proposing a normative baseline, and instead draw on the notion of *retrievability* in interpreting our results. Information retrieval researchers have coined the term *retrievability*, which refers to how accessible a document is in a system [4]. When a system systematically favors documents with particular characteristics, such that their retrievability is higher than that of others, the system exhibits a *retrievability bias* [44]. In a search for "person" images, it is clear that photos of men have significantly greater retrievability as compared to photos of women. Even when users are willing to consider a large set of images, men are still more representative of a "person."

Stereotype content and strength

RQ2 considers whether results perpetuate the prescriptive stereotype of men as "competent" and women as "warm." Creative professionals, including journalists and marketers, rely on images as data sources as well as objects that illustrate concepts they are trying to communicate [31]. To this end, they often formulate search queries that express their needs for an image that conveys abstract concepts, moods or inspirations [27]. Our approach, in which we formulated queries based on trait adjectives, not only allows us to replicate a psychological test in the digital context, it also represents a genuine, non-trivial search task, often performed on a "tight, commercially driven timescale" [21].

For each query executed to the API with the four search market parameters, we found the proportion of photos retrieved that depict men/boys, women/girls, and others. We then performed a cluster analysis in order to better understand which character traits are similar, in terms of the gender of individuals in the photos chosen by Bing to depict those traits, and the extent to which the same traits cluster together across regions / search markets.

We first plotted the within-clusters sum of squares by the number of clusters, which helps the analyst choose the optimal number of clusters [16]. In the plot, one looks for the "elbow," the point at which adding a cluster does not significantly reduce the within-group variance. The marginal improvement drops rapidly after three clusters, and thus, we settle on the three-cluster solution.

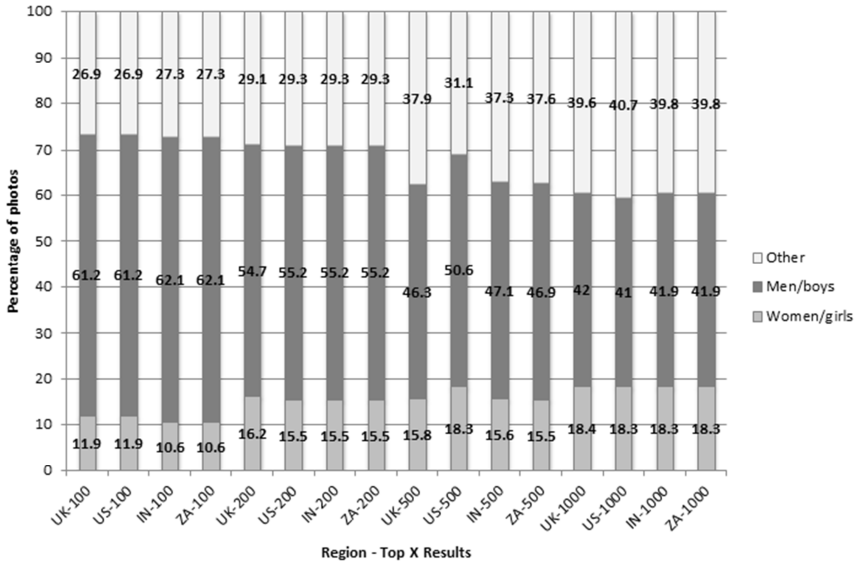

Figure 2: Gender distribution in photos retrieved for "person."

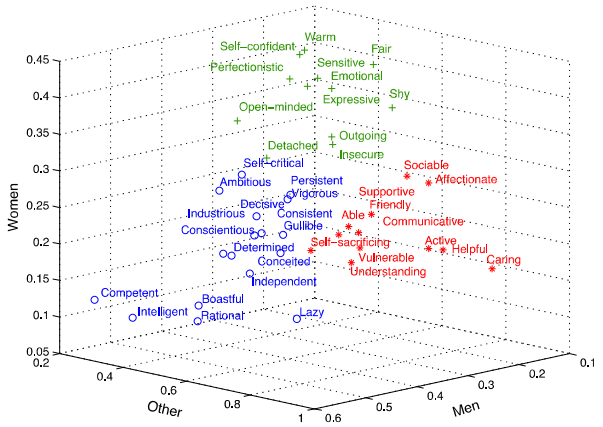

Figure 3: Three-cluster solution - feminine (green cross), masculine (blue circle), gender-neutral (red asterisk) traits.

We used R's *kmeans* routine to cluster the 272 traits (68 traits * four regions). The input for each trait consisted of the proportion of photographic images retrieved that depicted women/girls, men/boys, and others. Figure 3 depicts the clusters, while Table 5 describes cluster centroids. For 41 of 68 traits, the manner in which the trait was depicted in Bing results was consistent across all four regions. For clarity, Figure 3 and Table 5 represent these traits; the remaining 27 traits, which exhibit regional differences, are explored in Tables 6 to 8.

Clusters of character traits	Women	Men	Other
Gender-neutral: Able, Active, Affectionate, Caring, Communicative, Competitive, Friendly, Helpful, Self-sacrificing, Sociable, Supportive, Understanding, Vulnerable	0.19	0.23	0.58
Masculine: Ambitious, Boastful, Competent, Conceited, Conscientious, Consistent, Decisive, Determined, Gullible, Independent, Industrious, Intelligent, Lazy, Persistent, Rational, Self-critical, Vigorous	0.19	0.38	0.43
Feminine: Detached, Emotional, Expressive, Fair, Insecure, Open-minded, Outgoing, Perfectionistic, Self-confident, Sensitive, Shy, Warm	0.37	0.23	0.39

Table 5: Cluster centroids -Mean proportion of images depicting each gender.

Considering the 41 traits, we observe that characteristics such as "ambitious," "intelligent," and "rational," describing *competence*, are more often depicted in search results with photos of men as compared to women. In contrast, in searches for photos depicting *warm* traits, such as "emotional," "expressive" and "sensitive," photos depicting women are more frequent. It can also be noted that several gender-neutral traits are other-oriented. Characteristics such as "helpful" and "caring" describe relations with and behavior toward others. However, these traits are represented by images of both women and men. In a study of masculine norms and stereotypes, Moss-Racusin and colleagues [32] differentiated warm characteristics of low status (e.g., "insecure," "shy") from those that are status-neutral (e.g., "cooperative," "friendly," "supportive"). Indeed, 12 of the 13 gender-neutral traits are high-status; being an "able" and "sociable" person is beneficial, both in terms of getting along with others but also in achieving goals. In contrast, many feminine traits (e.g., "insecure," "shy") arguably hinder one's abilities.

As mentioned, 27 traits exhibited variation across Bing's regional search markets, in terms of the gender distributions of depicted persons in the photos retrieved. Tables 6 to 8 detail these differences. For instance, Table 6 indicates the search regions in which the listed character traits were depicted primarily with images of persons classified as "other," thus presenting as gender-neutral traits in the given region(s). As an example, "altruistic" is gender-neutral in three search markets (UK, IN, ZA), while "trustworthy" is deemed as a gender-neutral trait in only one search market (IN). Tables 7 and 8 make the analogous comparisons for traits deemed as being more masculine and feminine, respectively. Observations can be made with respect to the regional/cultural differences in image search results.

Character traits	UK	US	IN	ZA
Altruistic	X		X	X
Chaotic	X	X	X	
Loyal, Self-reliant	X	X		X
Assertive, Capable, Considerate, Sympathetic	X			X
Harmonious			X	X
Honest	X	X		
Open		X		
Reliable, Tolerant				X
Hardhearted	X			
Dogmatic, Moral, Obstinate, Strong-minded, Trustworthy			X	

Table 6: Gender-neutral traits with regional differences.

In particular, there are interesting differences in the photos retrieved for Bing's India search market. The IN results deem traits such as "dogmatic," "obstinate" and "strong-minded" as gender-neutral, while in other regions these terms tend to be depicted with photos of men. Likewise, "creative," "dominant" and "energetic" are often represented by photos of women in the US and IN markets, whereas these are presented as being more masculine traits in the UK and ZA.

In sum, it is clear that image search results reinforce the "offline" stereotypes of women being warm and other-oriented, and men as being competent individuals who can achieve their goals and aspirations. However, as discussed, we do observe regional differences for a number of traits.

Character traits	UK	US	IN	ZA
Dogmatic, Moral, Obstinate, Striving, Trustworthy	X	X		X
Polite	X		X	X
Egoistic, Tolerant	X		X	
Creative, Dominant, Energetic	X			X
Capable		X	X	
Altruistic, Sympathetic		X		
Chaotic, Honest				X
Considerate, Open, Self-reliant			X	

Table 7: Masculine traits with regional differences.

Character traits	UK	US	IN	ZA
Reliable	X	X	X	
Hardhearted		X	X	X
Reserved	X			X
Harmonious	X	X		
Assertive, Creative, Dominant, Energetic		X	X	
Egoistic		X		X
Polite		X		
Open				X
Striving, Sympathetic			X	

Table 8: Feminine traits with regional differences.

Backlash in stereotype-disconfirming photos

RQ3 asks whether there is evidence of a backlash effect against stereotype-incongruent individuals. We considered whether photos of men and women differed with respect to two characteristics: whether they convey themes of power and sex. To this end, we examined the gendered traits in Table 5, which carry positive or neutral valance (i.e., are not associated with low status).

Using photos retrieved for the UK search market, we compared the Clarifai tags for photos of men versus women. We analyzed tags with LIWC, relying on two categories: *power* (including words such as "success" and "superior") and *sexual* (including words such as "love" and "incest"). Tables 9 and 10 compare the proportion of photos of men versus women with tags conveying these concepts.

We compared the differences between men and women in terms of the proportion of photos conveying power, via the *z*-test for two population proportions. Since we examine multiple character traits, we use the appropriate Bonferroni correction in all our comparisons [29]. Across all traits, we find significantly more photos of men conveying power. Likewise, we find significantly more photos of women conveying sexual concepts, for all traits in both Big Two dimensions. This result is consistent with previous findings that images of women on the Web tend to be associated with metadata containing sexual language [36].

	Power		*Sex*	
	M	F	M	F
Person	*.64*	*.35*	*.02*	*.35*
Ambitious	.68	.28	0	.36
Competent	.71	.26	0	.24
Conscientious	.70	.21	.01	.29
Consistent	.76	.24	.01	.24
Decisive	.58	.26	.02	.28
Determined	**.79**	.30	.01	.27
Independent	**.80**	.33	0	.17
Industrious	.70	.30	.01	.32
Intelligent	.73	.36	0	.26
Persistent	.60	.35	.02	.21
Rational	.70	.24	0	.28
Vigorous	.78	.21	.01	.43
Wilcoxon rank sum test	W-value: 1* mean *d*: 0.17		W-value: 10 mean *d*: 0.10	

Table 9: *Agentic traits* – proportion of photos conveying themes of power and sex. (*$p < .01$)

	Power		*Sex*	
	M	F	M	F
Person	*.64*	*.35*	*.02*	*.35*
Emotional	.59	.29	.02	.27
Expressive	**.49**	**.20**	0	.50
Fair	.63	.23	.03	.29
Open-minded	**.48**	.23	.02	.37
Outgoing	.73	.23	.01	.22
Sensitive	.70	.29	.01	.31
Warm	**.45**	**.20**	**.10**	.41
Wilcoxon rank sum test	W-value: 5 mean *d*: 0.11		W-value: 8 mean *d*: -0.14	

Table 10: *Warm traits* – proportion of photos conveying themes of power and sex.

Given these differences, we also compared photos retrieved for each gender and trait against those retrieved for our baseline query, "person." The bold entries in Tables 9 and 10 indicate a significant difference ($p<p_{critical}$) with respect to the appropriate baseline, per the *z*-test for two population proportions. For example, photos of "determined" and "independent" men (i.e., competent/agentic men) more often convey power, as compared to photos of men retrieved for the query "person." Likewise, photos of "expressive," "open-minded" and "warm" men are less often associated with concepts of power, as compared to the baseline. Finally, photos of "warm" men are more likely to convey sexual themes as compared to baseline.

In contrast, photos of competent/agentic women never enjoy a boost in their association with concepts of power. Few photos of women that depict warm character traits are associated with tags suggesting power, with "warm" and "expressive" women having significantly fewer power tags as compared to baseline. In addition, half of the photos of women retrieved on the query for "expressive person" conveys sexual themes, while none of the photos of men retrieved for "expressive" do so.

In considering evidence for backlash, we observed that photos of *warm* people of both genders are depicted as less powerful as compared to baseline. However, with respect to agentic traits, we observe a difference across genders. While photos of agentic men gain a boost in their perceived power, women are penalized, being perceived as less powerful as compared to baseline. We used the Wilcoxon rank sum test, a non-parametric alternative to the paired t-test, to compare the magnitude of the deviations in Table 9 and Table 10. As shown, we find a statistically significant backlash effect only for agentic traits, with respect to power. Figure 4 visually depicts this backlash effect.

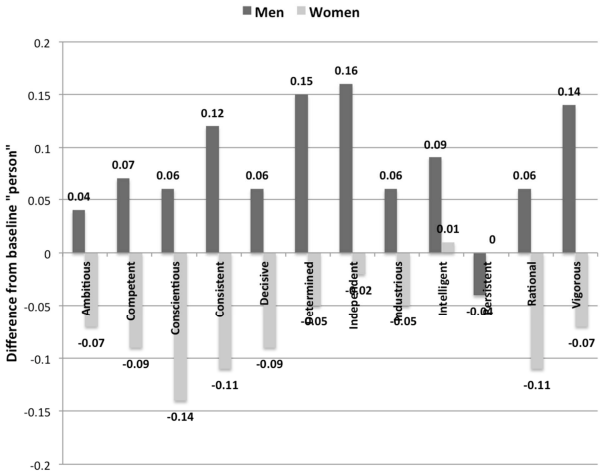

Figure 4: Proportion of agentic images conveying power (difference from respective "person" baseline).

In conclusion, we find a backlash effect against images of agentic women, mirroring that described in other domains (e.g., hiring and promotion in organizations [41], media depictions of women politicians [7]). The photos of men retrieved by Bing to represent agentic traits convey more power, as deemed by the Clarifai tags, as compared to the photos of men retrieved in the baseline query "person." However, the reverse is true for the photos retrieved by Bing for agentic women, which convey even less power than photos retrieved to the baseline query.

DISCUSSION

Algorithms are playing an increasing role in our information eco-system. To this end, they have increasing influence over our ability to be informed and engaged citizens. The algorithmic processes behind modern search engines are complex, proprietary, and difficult to study since there is no "ground truth," given the use of extensive machine learning and personalization. Yet despite this complexity, users approach systems at "interface value" [45]; people equate a simple interface with simple, trustworthy results, with some users remaining completely unaware that algorithms mediate their access to information [14].

While our work does not evidence that the Bing algorithm itself is gender biased (i.e., biases are equally, if not more, likely to be rooted in the underlying training data, media and metadata), it does demonstrate empirically that explicit and implicit gender biases that are widely observed in the social world (e.g., [13, 32]) are reproduced in image search results. Specifically, the current work evidences that the following gender biased phenomena are surfacing in results presented to users: gendered perceptions of personhood that work to undermine women's agency, a gendering of character traits that relate agency and competence to men, and warmth and communality to women, and a backlash effect that generates a status penalty for gender-incongruent women displaying "competency" related traits.

Further research is required to understand specifically how and why such subtle forms of gender bias are present in the results. However, given that subtle and implicit forms of gender bias do exist, it is clear that any solution must not simply correct the proportion of men and women represented in image searches. Rather, it will be important to dig deeper in order to thoroughly examine how and why subtle and implicit biases are entering the system at various stages of the engineering process. Only with such thorough understanding can meaningful socio-technical solutions be proposed and explored. Until then, designers of applications using images from APIs, such as Bing's, must be cognizant that any inherent biases will be carried into their own tools. The methodology presented here is reproducible and usable by others wanting to detect such biases.

In the bigger picture, though, to understand when gender bias emerges in search results is important because, although the nature of the relationship between sex and gender identity is not yet fully understood [15], what is understood is that gender expression (the ways we express gender through dress, behavior, speech, etc.) is a socially constructed phenomenon that has significant consequences for gender equality. Although many people act in gender stereotypical ways, this is largely because they are socially conditioned to do so. The gendering of particular character traits as masculine and feminine varies over time and place.

Even in our own findings it was interesting to observe that some character traits (e.g., creative, dominant and energetic) took on different gendered meanings in different countries, and that some traditionally masculine traits (e.g., assertiveness and competitiveness) emerged as gender-neutral. This means that there is no essential "truth" of gendered character traits, only historically constituted perceptions about what forms of gender expression are, and are not, seen to be culturally appropriate. It is widely recognized that these types of social constructs are deeply embedded into social institutions, media products, cultural norms, language use, etc., thus providing resources for individuals to learn and "do" gender [46]. In the current work, we observed these same constructs clearly emerging in the results of image searches.

Algorithmic outputs, such as search results, mediate how we perceive the nature of social reality [20]. They act as a lens through which we come to understand the world, and act as a means of circulation for media objects – endowing some images with greater retrievability than others, allowing them to be accessed and reproduced more freely. As such they – as with all media – implicitly inform our understanding and practice of gender expression.

Continuing gender inequality across a variety of social contexts is dependent upon the types of gendered representations of personhood and character traits that are observed in our findings. When search results uncritically reproduce bias in the images they present to users, they contribute to the reinforcement and stabilization of gender bias. This is the case whether the bias is a phenomenon of the algorithm or the underlying data. If the underlying data is biased then a "neutral" algorithm – if this is indeed a possibility – becomes a force for conservation of the status quo. History tells us that social change occurs when we bias our perception towards our vision of the future and generate new imaginaries for how things might be – and that we need to sow the seeds of that future in the present. To take such a stance would mean engineers critically and actively biasing their algorithms towards gender equality, rather than aiming to uncritically and objectively represent the gender bias that emerges in the images they make retrievable. Neither option is a neutral act; both have significant social consequences.

We recognize a number of limitations of our work. Firstly, we use a simplistic and binary view of gender rather than a more sophisticated construct. Secondly, the study relies on results provided by the Bing API. However, the Bing documentation does not fully describe how results are generated and results are similar, but not identical, to results produced using Bing Web search. Finally, we recognize that the resources and tools used in the automated methodology may themselves introduce bias. For example, similar to any other image recognition tool, Clarifai has been trained on specific datasets that may introduce other nuances into the investigation of algorithmic bias.

CONCLUSION

Increasingly, information services such a search systems utilize algorithms for filtering and selecting content that can influence users' views of the social world. Calls to identify and make such biases clear to users are being made that may affect the design of future services. In this paper, we present a reproducible methodology for studying the perpetuation of gender stereotypes within image search. Results show the existence of gender biases, as well as evidence of backlash effects. In future work we plan to test other visual recognition tools, examine other aspects of stereotypes (e.g., age, race and ethnicity), compare results for image databases beyond the Web and carry out a deeper qualitative analysis of results, particularly for understanding backlash effects.

REFERENCES

1. Colin Allen, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems* 21, 4, 12-17.

2. 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, 935-948.

3. 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, 1202-1217.

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* (CIKM '08). ACM, New York, NY, USA, 561-570.
5. Paul Baker and Amanda Potts. 2013. "Why do White People have Thin Lips?" Google and the Perpetuation of Stereotypes via Auto-complete Search Forms. *Critical Discourse Studies 10,* 2: 187-204.
6. S.A. Basow. 1986. *Gender Stereotypes: Traditions and Alternatives.* Monterey, CA: Brooks/Cole.
7. Michelle C. Bligh, Michele M. Schlehofer, Bettina J. Casad, and Amber M. Gaffney. 2012. Competent enough, but would you vote for her? Gender stereotypes and media influences on perceptions of women politicians. *Journal of Applied Social Psychology* 42, 3: 560-597.
8. Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3: 209-227.
9. Amy J. 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 Psychology* 40, 61-149.
10. Anupam Datta, Shayak Sen and Yair Zick. 2016. Algorithmic transparency via quantitative input influence. Proceedings of the 37th IEEE Symposium on Security and Privacy, IEEE Computer Society, 598-617. DOI 10.1109/SP.2016.42
11. Nicholas Diakopoulos. 2015. Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3: 398-415.
12. 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, 1171-1188
13. Alice H. Eagly, and Steven J. Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological review* 109, 3, 573-598.
14. Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 153-162.
15. Randi Ettner and Antonio Guillamon. 2016. Theories of the Etiology of Transgender Identity. In R. Ettner, S. Monstrey, and E. Coleman (eds.), *Principles of Transgender Medicine and Surgery* (pp. 3-15). New York: Routledge.
16. Brian S. Everitt and Torsten Hothorn. 2009. *Statistical Analysis Using R.* Chapman and Hall / CRC.
17. Susan T. Fiske, Amy J.C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology 82, 6, 878-902.
18. Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3: 330-347.
19. G.M. Gilbert. 1951. Stereotype persistence and change among college students. *Journal of Abnormal and Social Psychology* 46, 245-254.
20. Tarleton Gillespie. 2014. The Relevance of Algorithms. *Media technologies: Essays on communication, materiality, and society*: 167.
21. Ayşe Göker, Richard Butterworth, Andrew MacFarlane, Tana Ahmed, and Simone Stumpf. 2016. Expeditions through image jungles: the commercial use of image libraries in an online environment. *Journal of Documentation* 72,1, 5-23.
22. David L. Hamilton and Tina K. Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach. In J. Dovidio and S.L. Gaertner (eds.), *Prejudice, Discrimination, and Racism* (pp. 127-163). New York: Academic Press.
23. Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 527-538. International World Wide Web Conferences Steering Committee.
24. Marvin Karlins, Thomas L. Coffman and Gary Walters. 1969. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology* 13, 1-16.
25. Daniel Katz and Kenneth Braly. 1933. Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology* 28, 3, 280-290.
26. 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* (CHI '15). ACM, New York, NY, USA, 3819-3828. DOI=http://dx.doi.org/10.1145/2702123.2702520

27. Elena Konkova, Andrew MacFarlane and Ayşe Göker. 2016. Analysing creative image search information needs. *Knowledge Organization* 43, 1: 13-21.
28. John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (Spring 2004): 50-80.
29. John Ludbrook. 1998. Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology*, *25*, 12, 1032-1037.
30. Stephanie Madon, Max Guyll, Kathy Aboufadel, Eulices Montiel, Alison Smith, Polly Palumbo, and Lee Jussim. 2001. Ethnic and national stereotypes: The Princeton Trilogy revisited and revised. *Personality and Social Psychology Bulletin* 27, 8, 996-1010.
31. Lori McCay-Peet and Elaine Toms. 2009. Image use within the work task model: Images as information and illustration. *J. Am. Soc. Inf. Sci. Technol.* 60, 12 (December 2009), 2416-2429.
32. Corinne A. Moss-Racusin, Julie E. Phelan, and Laurie A. Rudman. 2010. When men break the gender rules: Status incongruity and backlash against modest men. *Psychology of Men & Masculinity* 11, 2: 140-151.
33. Michael Mudrick. 2015. Pervasively offside: A gendered analysis of sportscasting. Doctoral dissertation, Paper 722. http://digitalcommons.uconn.edu/dissertations/772
34. Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1, 81-103.
35. Safiya Umoja Noble. 2012. Missed connections: What search engines say about women. *Bitch: Feminist Response to Pop Culture* 54 (Spring 2012), 37-41.
36. Jahna Otterbacher. 2015. Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via GWAP. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 1955-1964.
37. 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: 801-823.
38. James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. https://utexas-ir.tdl.org/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?sequence=3
39. Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 173-182.
40. Laurie A. Rudman and Peter Glick. 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues* 57, 4, 743-762.
41. Laurie A. Rudman and Julie E. Phelan. 2008. Backlash effects for disconfirming gender stereotypes in organizations. *Organizational Behavior* 28, 61-79.
42. Stefanie Simon and Crystal L. Hoyt. 2012. Exploring the effect of media images on women's leadership self-perceptions and aspirations. *Group Processes & Intergroup Relations* 16, 2, 232-245.
43. Charlotte Templin. 1999. Hillary Clinton as threat to gender norms: Cartoon images of the first lady. *Journal of Communication Inquiry* 23, 1, 20-36.
44. Myriam C. Traub, Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Argen de Vries, and Lynda Hardman. 2016. Querylog-based assessment of retrievability bias in a large newspaper corpus. In *Proccedings of the Joint Conference on Digital Libraries* (JCDL '16), ACM, Newark, NJ, 7-16.
45. Sherry Turkle. 1997. *Life on the Screen: Identity in the Age of the Internet*. New York: Simon & Schuster.Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & society* 1.2, 125-151.
46. Candace West and Don H. Zimmerman. 1987. Doing gender. *Gender & society* 1.2, 125-151