Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search

Yunhe Feng and Chirag Shah
Information School, University of Washington, {yunhe,chirags}@uw.edu

Abstract

Gender bias is one of the most common and well-studied demographic biases in information retrieval, and in general in AI systems. After discovering and reporting that gender bias for certain professions could change searchers’ worldviews, mainstreaming image search engines, such as Google, quickly took action to correct and fix such a bias. However, given the nature of these systems, viz., being opaque, it is unclear if they addressed unequal gender representation and gender stereotypes in image search results systematically and in a sustainable way. In this paper, we propose adversarial attack queries composing of professions and countries (e.g., ‘CEO United States’) to investigate whether gender bias is thoroughly mitigated by image search engines. Our experiments on Google, Baidu, Naver, and Yandex Image Search show that the proposed attack can trigger high levels of gender bias in image search results very effectively. To defend against such attacks and mitigate gender bias, we design and implement three novel re-ranking algorithms – epsilon-greedy algorithm, relevance-aware swapping algorithm, and fairness-greedy algorithm, to re-rank returned images for given image queries. Experiments on both simulated (three typical gender distributions) and real-world datasets demonstrate the proposed algorithms can mitigate gender bias effectively.

Introduction

The web’s biggest image search engines, such as Google, and Bing, provide an important information-seeking interface for people to explore the world. According to Internet Live Stats[1], Google processes more than 3.5 billion queries per day and 1.2 trillion searches per year. Google image searches account for 22.6% of all searches[2]. Given the volume and importance in our daily lives, image search results can significantly influence how people perceive and view the world. Images are often more than useful objects of information; they provide a visual representation of a phenomenon, a concept, and a perceived reality of the world around us. Given this, it is not sufficient to assess the quality of image search results using relevance metrics; we also need to consider how this visual representation of a phenomenon, a concept, and a perceived reality of the world around us affect their worldview (Hibbing and Rankin-Erickson 2003).

If these results carry biases such as those shown by Lam et al. (2018) and Kay, Matuszek, and Munson (2015), they are much easier to perpetuate than regular web search results. Therefore, while evaluating image search results, we need to look beyond their relevance. We must also look at inherent disparity and biases carried out in them.

Among different types of image biases, gender bias is one of the most common and well-studied demographic biases. Not surprisingly, this also gets scrutiny and attention from scholars and media. That often makes the service providers take immediate actions to fix such biases in an ad hoc manner. For example, after the work described before (Kay, Matuszek, and Munson 2015) received a lot of attention, Google shifted the gender distributions in image search results for CEO and some other occupations. For instance, the famous query of CEO in image search has been fixed for a long time (see Figure 1(a) and Figure 1(d)). However, Mozilla’s recent Internet Health Report (Mozilla 2021) points out that the default internet user is still viewed as white, male, and cisgender, and big tech has not done enough to fix it (Griffith 2021). But that is only one way the gender bias problem is not fixed.

In this paper, we revisit gender bias in image search results for professional occupations. For some search terms, such as ‘CEO,’ gender fairness is observed in image search results. But have image search engines mitigated gender biases in search results systematically? To further illustrate this research question, we present the relevant adversarial attack queries of the CEO in Google and Bing, as shown in Figure 1. Both Google and Bing image search engines have already fixed the gender stereotypes for the occupation of CEO. However, such gender bias resurfaces when appending the country names, such as United States and UK, to the original keyword of CEO. This finding inspires us to dive into the exploration of gender fairness in image search engines to reveal superficial bias mitigation.

Ten occupations, also investigated by Kay, Matuszek, and Munson (2015) seven years ago, are chosen as image search terms in our study. We design two search keywords for each occupation, i.e., the original occupation name and the adversarial attack keyword that consists of occupation name and the country name of United States. The latter aims to trigger gender bias in image search results. For each keyword, we retrieve the top 200 images (if available) from four widely used image search engines, namely, Google from USA, Baidu from China, Naver from South Korea, and Yandex from Russia. In total, we collected more than 18,000 images.

When image retrieval is complete, we attempt to leverage
We compared the gender labels detected by APIs with human annotations, and found that only Amazon Rekognition APIs were acceptable but still failed to handle the images with a low ratio of detected faces. Therefore, we propose a hybrid approach to detect face genders in search images by combining Amazon Rekognition results and crowdsourced human annotations through Amazon Mechanical Turk.

To mitigate gender bias, we present three generalized re-ranking algorithms, including the epsilon-greedy algorithm, relevance-aware swapping algorithm, and fairness-greedy algorithm, to balance the trade-off between gender fairness and image relevance. Evaluations of the proposed gender bias mitigation algorithms on both simulated and real-world datasets demonstrated that it is feasible (and advisable) to address bias in image search (and perhaps in other types of search as well) in a systematic and more meaningful way than doing individual query fixes in an ad hoc manner.

Our contributions are summarized as follows.

- We design adversarial attacks with regard to gender bias in image search and determine that gender bias is not fixed systematically by search engines.
- CIRF, an open-sourced Cross-search-engine Image Retrieval Framework to collect images from multiple search engines for given search terms, is developed.
- We find image gender detection APIs cannot always perform well on search images in the wild, so a hybrid approach combining automatic gender detection and manual annotation is presented.
- We propose and validate three re-ranking algorithms to mitigate gender bias in image search results.

**Related Work**

This section presents the importance of gender fairness in image search results, summarizes the gender bias-related research findings from multiple perspectives, discusses the existing approaches to mitigate image gender biases, and highlights the difference between existing works and ours.

The gender fairness or biases demonstrated in image search results affect people’s perceptions and views significantly (Ellemers 2018; Metaxa et al. 2021). Kay, Matuszek, and Munson (2015) are among the first to investigate gender biases in professional occupation image search results. They reported that such image search results for occupations slightly exaggerate gender stereotypes, and people thought image search results were better if they agreed with the stereotype. More importantly, this research work pointed out that the biased representation of gender in image search results could shift people’s perceptions about real-world distributions. Otterbacher, Bates, and Clough (2017) proposed a trait adjective checklist inspired method further to identify the existence of gender biases in image search. They found that images of men were more often retrieved for agentic traits whereas warm traits were demonstrated in photos of women. In addition, photos of stereotype-incongruent individuals exhibited a backlash effect, e.g., ‘competent women’ were less likely to be portrayed positively. Otterbacher et al. (2018) measured the user perception of gender bias in image search from the perspective of sexism, and found search engine users, who were more sexist, were less likely to perceive gender biases.

There also exist many research studies exploring gender bias in different types of images. By detecting gender labels of the photographs of U.S. members of Congress and their tweeted images, Schwemmer et al. (2020) concluded Google Cloud Vision (GCV) could produce correct and biased labels at the same time because a subset of many possible true labels was selectively reported. Wijnhoven (2021) found a gender bias toward stereotypically female jobs for women but also for men when searching jobs via Google search engine. By examining four professions across digital platforms, Singh et al. (2020) concluded: 1) gender stereotypes were most likely to be challenged when users acted directly to create and curate content, and 2) algorithmic approaches for content curation showed little inclination towards breaking stereotypes. Makhortykh, Urman, and Ullea (2021) conducted a cross-engine comparison of racial and gender bias in the visual representation of the search term ‘artificial intelligence’ and gender representation of AI is more diverse than its racial representation. Hashemi and Hall (2020) reported no gender bias was identified when detecting criminal tendency based on mugshot images of arrested individuals.

In the last couple of years, many approaches have been proposed to detect and mitigate gender bias in images for training deep learning models (Wang et al. 2020; Xu et al. 2020; Hwang et al. 2020; Adeli et al. 2021). For example, Serna et al. (2021) showed how bias in face images impacted the activations of gender detection models and developed InsideBias to detect biased models. To reduce gender bias in deep image representations, an adversarial method for the removal of features associated with a protected variable (gender) from the intermediate convolutional neural network based representations was presented (Wang et al. 2019). Many other
image gender bias mitigation approaches, such as a posterior regularization based gender de-biasing framework (Jia et al. 2020), a fairness-aware disentangling variational auto-encoder (FD-VAE) (Park et al. 2021), and an adversarial gender de-biasing algorithm (AGENDA) (Dhar et al. 2020), are also proposed. Besides, some post-processing bias mitigation methods, such as FA*IR (Zehlike et al. 2017) and multi-task learning for fair regression (Zhao and Chen 2019), have been proposed.

Our study adds to the literature on exploring gender bias in image search results in the following ways. First, similar to Kay, Matuszek, and Munson (2015), we investigate the gender distribution in professional occupation image results. Still, we also design an adversarial search attack by adding the country information into the occupation search terms. We find evidence that image search engines do not fix the reported gender bias in search results systematically. Second, we not only examine the performance of five popular image gender detection APIs, but also propose a hybrid approach that combines automatic detection (Amazon Rekognition services) and manual annotations (Amazon Mechanical Turk) to improve gender distribution estimation. Finally, we develop three re-ranking algorithms, i.e., epsilon-greedy, relevance-aware swapping, and fairness-greedy methods, to mitigate gender biases in image search results.

**Image Retrieval and Gender Detection**

In this section, we describe how to build the image search datasets, examine the performance of image based gender detection APIs, and propose a hybrid approach to strike a balance between detection accuracy and efficiency.

**Search Image Retrieval**

We propose and develop an open-sourced Cross-search-engine Image Retrieval Framework (CIRF) to automatically collect images from multiple search engines for given search terms. CIRF mainly consists of three components: URL Builder, Data Downloader, and Image Parser.

**URL Builder** To enable automatic data download, we first construct image search URLs based on search-engine-specific URL templates for given search terms. For example, we use `https://www.google.com/search?q=keyword&source=lnms&tbm=isch` as a URL template for Google Image Search, where `keyword` is the placeholder of search terms. As CIRF is able to handle multiple search engines, we also design similar URL templates for popular search engines, including Baidu from China, Naver from South Korea, and Yandex from Russia. The `keyword` can be written in any language because browsers would encode it in the UTF-8 format.

**Data Downloader** We collect two types of search image data based on the built URLs: (i) the web page HTML file that captures the layout and names of images returned by search engines; (ii) individual image files embedded in the image gallery. CIRF adopts the web framework Selenium WebDriver to open URLs in a Chrome browser with incognito mode. To display and cache more search images, CIRF scrolls up and down the web page by sending PAGE_UP and PAGE_DOWN commands to the HTML entities. CIRF leverages PyAutoGUI, a cross-platform GUI automation module, to save the web page HTML file and all supplementary materials including images.

**Image Parser** This component is responsible for extracting the images and their orders from the downloaded HTML file. In general, three types of images are collected: standard images, Base64 encoded images, and image URLs. For the latter two types of images, CIRF decodes them into standard images and retrieves images via URLs respectively. When all images are ready, CIRF renames them according to their orders in the HTML file for further analysis.

**Gender Detection**

Image-based gender detection has widely been adopted in diverse domains, so many commercial and open-sourced gender detection APIs have been developed and released. Considering the scalability and efficiency, we intended to rely on these available tools to label search images automatically. To evaluate their performance, we randomly selected and searched ten occupations and their corresponding adversarial attack search terms (i.e., appending ‘United States’) in Google Image Search. Then we conducted an IRB-approved user study to recruit participants from Amazon Mechanical Turk (MTurk) to build the ground truth of genders. We paid each participant $0.5 for annotating 50 images, and each image was assigned to three workers. Five popular gender detection APIs, including Amazon Rekognition, Luxand, Face++, Microsoft Azure, and Facebook DeepFace, were chosen to calculate gender distributions of the top 200 Google search images for each given search term.

Figure 2 demonstrates the normalized female ratio difference between MTurk results and face gender detection APIs. In general, Amazon Rekognition outperforms the rest of APIs in terms of face detection ratios (see X-axis) and the female ratio errors (see Y-axis). When the face detection ratio is above 0.5, the normalized difference of Amazon Rekognition is below 15%. Therefore, Amazon Rekognition was chosen to identify the genders of people in images. Thus, we propose a two-step hybrid method to annotate image gender labels: 1) use Amazon Rekognition to detect image genders; 2) for search terms that suffer from a low face detection ratio (below 0.5), we still rely on MTurk to manually label them.

**Exploring Unsystematic Gender Bias Fixing**

We investigate whether gender bias in image search results is systematically fixed by designing adversarial search attacks and measuring the degree of gender fairness.

**Adversarial Search Attack Design**

As mentioned before, we are motivated to investigate whether image search engines fix gender bias in different occupation queries systematically. Therefore, we follow the occupation list on U.S. Bureau of Labor Statistics and choose occupation names as the baseline search keywords. When con-

---

1. https://github.com/YunheFeng/CIRF
2. https://www.selenium.dev/documentation/webdriver/
3. https://pyautogui.readthedocs.io/
4. https://www.bls.gov/oes/current/oes_nat.htm#00-0000
structing adversarial searches, we append the country name of ‘United States’ to each occupation name to build the attacking search term. If both baseline and attacking searches demonstrate no difference with the gender distribution ground truth for one occupation, we think search engines mitigate gender bias systematically for that occupation. Otherwise, we argue that such fixes and mitigation of gender bias are just hit-or-miss.

As the previously existing gender biases in image search queries, such as ‘CEO,’ drew huge attention of the public and academia, mainstream search engines had already mitigated such biases accordingly. However, our analytics show that gender biases crossing over all occupations are not fixed in a systematic way.

**Gender Bias Measurement**

It is very intuitive and straightforward to compare the normalized difference between gender probability distribution $P$ in image search results and the ground truth gender probability $T$ for each occupation. For top $k$ images returned by search engines, we calculate the Kullback-Leibler divergence $D_{KL}(T \parallel P^k)$ between these $k$ images and the ground truth. The average Kullback-Leibler divergence is used to represent the existing bias.

$$d = \frac{\sum_{k=1}^{N} D_{KL}(T \parallel P^k)}{N}$$

(1)

**Algorithms to Mitigate Gender Bias**

We propose three interpretable and lightweight re-ranking algorithms to mitigate gender biases in image search results.

**Epsilon-greedy Algorithm**

Inspired by the exploitation and exploration trade-off idea in reinforcement learning [Berry and Fristedt 1985; Sutton and Barto 2018] and re-ranking [Gao and Shah 2020], we propose the epsilon-greedy re-ranking algorithm, which swaps items in the image rank list with a controllable degree of randomness. The randomized swapping breaks original gender distributions and might improve fairness especially when items with the same attribution values are gathered together densely (e.g., male CEO images fully occupy the top 20 CEO image search results). This algorithm has two main advantages – simplicity and generalizability. It is straightforward and simple to randomly shuffle items without considering other factors. In addition, no prior knowledge, such as the optimal gender distribution, is required to apply this algorithm.

In the proposed epsilon-greedy algorithm, the randomness is specified by the parameter $\epsilon \in (0, 1]$, representing the probability of swapping two items. Each item has a probability of $\epsilon$ to exchange positions with a random item that follows it. A larger $\epsilon$ introduces more randomness, leading to a re-ranked list that is more different from the original list.

**Algorithm 1: Epsilon-greedy Algorithm**

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Input:</strong> L: the original image list; $c$: the probability of swapping two items;</td>
</tr>
<tr>
<td>2</td>
<td><strong>Output:</strong> R: the re-ranked image list;</td>
</tr>
<tr>
<td>3</td>
<td>$R \leftarrow \emptyset$; // initialize R as empty</td>
</tr>
<tr>
<td>4</td>
<td>for $i = 1 \rightarrow</td>
</tr>
<tr>
<td>5</td>
<td>// a random number between 0 and 1;</td>
</tr>
<tr>
<td>6</td>
<td>if $p \leq \epsilon$ then // swap items</td>
</tr>
<tr>
<td>7</td>
<td>$temp \leftarrow L_i$;</td>
</tr>
<tr>
<td>8</td>
<td>$j \leftarrow$ a random number between $i + 1$ and $</td>
</tr>
<tr>
<td>9</td>
<td>$L_i \leftarrow L_j$;</td>
</tr>
<tr>
<td>10</td>
<td>$L_j \leftarrow temp$;</td>
</tr>
<tr>
<td>11</td>
<td>append $L_i$ to R; // add swapped item</td>
</tr>
<tr>
<td>12</td>
<td>else // keep the original item</td>
</tr>
<tr>
<td>13</td>
<td>append $L_i$ to R; // add unswapped item</td>
</tr>
<tr>
<td>14</td>
<td>end</td>
</tr>
<tr>
<td>15</td>
<td>return R</td>
</tr>
</tbody>
</table>

**Relevance-aware Swapping Algorithm**

Normally, the items with large relevance weights are ranked at the top of search engines’ returned image list. If unrelated or less related images are ranked high, it harms user experience and utility. The epsilon-greedy algorithm is very straightforward and simple, but it ignores the relevance of search items during re-ranking. Therefore, we propose the relevance-aware swapping algorithm to consider both the randomness and relevance weight of image items to re-rank the image list. To keep the utility of the re-ranked list, an image with a larger relevance weight is less likely to be swapped with an image item that follows it.

**Relevance Weight Modeling**

We grade an image’s relevance weight based on its index in the image list returned by search engines. Similar to the Mean Reciprocal Rank (MRR) [Voorhees et al. 1999], the relevance weight of an image with a rank index $i$ can be modeled by its reciprocal rank $\frac{1}{i}$. However, such relevance weight decreases too fast with the growth of the rank index $i$. Instead, we model the relevance weight distribution in a linear manner. Suppose we have an image list $L$ containing $|L|$ images. The linear
the relevance weight of the $i^{th}$ image is estimated as $1 - \frac{d}{|L_i|}$. Inspired by the Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen 2002), we further introduce a discount factor of $\log_2(i+1)$ to smooth the decay of relevance weights of bottom images in $L$. Finally, the relevance weight of image $L_i$ is expressed as:

$$W_i = \frac{1 - \frac{d}{|L_i|}}{\log_2(i+1)}$$

Swapping Probability The swapping probability of image $L_i$ is determined by its relevance weight $W_i \in [0, 1]$. To ensure that the image with a high relevance weight is less likely to be swapped, we can use $1 - W_i$ to represent the swapping probability for image $L_i$. In addition, we design a coefficient $\rho \in (0, 1]$ to further control the swapping sensitivity and the swapping probability of image $L_i$ is expressed as $\rho(1 - W_i)$.

The detailed implementation of the fairness-greedy algorithm is shown in Algorithm 3. To make our algorithm more general, we use $X$ to represent all involved features, such as gender features of female and male. Note that the fairness-greedy algorithm is capable of handling more than two different features. We keep the first item in the original rank list as it is at the beginning (see line 3). Starting from the second item, we calculate the gender distribution $P$ over the latest re-ranked list $R$. $P_x$ represents the ratio of feature $x \in X$, and $T_x$ is the ground truth of feature $x$ in the real world. Next, we take a two-step re-ranking method to mitigate feature biases. Step 1: identify the most underrepresented feature $x_{\text{min}}$ by comparing the difference between $P_x$ and $T_x$ (see line 12-16). Step 2: find the first item $L_j$ with a feature of $x_{\text{min}}$ (i.e., $G_{L_j} = x_{\text{min}}$) in $L_{[i+1]}$ and move it forward as the new $L_i$ (see line 17-27). If such an item $L_j$ does not exist, we exclude the feature $x_{\text{min}}$ by adding it into the checked feature set $C$ and continue the re-ranking (see line 9-11).

### Algorithm 2: Relevance-aware Swapping Algorithm

1. **Input**: $L$: the original image list; $\rho$: the sensitivity of swapping two items;
2. **Output**: $R$: the re-ranked image list;
3. $R \leftarrow \emptyset$; // initialize $R$ as empty;
4. for $i = 1 \rightarrow |L|$ do
   5. $W_i = \frac{1 - \frac{d}{|L_i|}}{\log_2(i+1)}$; // relevance weight
   6. $p \leftarrow$ a random number between 0 and 1;
   7. if $p \leq \rho(1 - W_i)$ then // swap items
      8. $\text{temp} \leftarrow L_i$;
      9. $j \leftarrow$ a random number between $i + 1$ and $|L|$;
      10. $L_j \leftarrow \text{temp}$;
      11. $L_j \leftarrow \text{temp}$;
      12. append $L_i$ to $R$; // add unswapped item
   else // keep the original item
      13. append $L_i$ to $R$; // add swapped item
   end
15. end
16. return $R$

### Algorithm 3: Fairness-greedy Algorithm

1. **Input**: $L$: the original image list; $T$: the ground truth of gender distribution; $X$: the set of gender features;
2. **Output**: $R$: the re-ranked image list;
3. $R \leftarrow [L_1]$; // initialize $R$ as $[L_1]$;
4. for $i = 2 \rightarrow |L|$ do
   5. $P \leftarrow$ gender distribution on $[G_{L_1}, ..., G_{L_{i-1}}]$;
   6. $\text{flag} \leftarrow \text{False}$;
   7. $x_{\text{min}} \leftarrow \text{None}$; // most underrep. feat.
   8. $C \leftarrow \emptyset$; // set checked features as $\emptyset$
   9. while ($\text{flag} = \text{False}$) and ($C \neq X$) do
      10. $d_{\text{min}} \leftarrow 0$; // update $C$
      11. add $x_{\text{min}}$ to $C$; // update $C$
      12. for $x \in X - C$ do
         13. $d = P_x - T_x$; // diff. in feat. $x$
         14. if $d \leq d_{\text{min}}$ then
            15. $x_{\text{min}} \leftarrow x$; // underrep. feat.
         end
      end
      16. /* find 1st item w/ underrep. feat. */
      17. for $j = i \rightarrow |L|$ do
         18. if $G_{L_j} = x_{\text{min}}$ then // find an item
            19. $\text{temp} \leftarrow L_j$; // save $L_j$
            20. for $k = i + 1 \rightarrow j$ do
               21. $L_k \leftarrow L_{k-1}$ // move down
            end
            22. $L_i \leftarrow \text{temp}$; // update $L_i$
            23. append $L_i$ to $R$; // update $R$
            24. $\text{flag} \leftarrow \text{True}$; // find the item
            25. break;
         end
      end
   end
   17. return $R$

### Experiments and Evaluation

This section presents the evaluations of the three proposed bias mitigation approaches on synthetic and real datasets.
**Evaluation on Synthetic Data**

We generated three synthetic datasets with different gender distribution patterns: 1) Uniform Dataset: female and male items are distributed evenly across the whole list; 2) Heavy-headed Dataset: female items are aggregated at the top of the list; 3) Heavy-tailed Dataset: female items are aggregated at the bottom of the list. For these experiments, we created a list $L$ with a length of 200 and set the female ratio as 0.5, i.e., 100 items are labeled as female. On the heavy-headed and heavy-tailed datasets, the 100 female items are distributed at the top 50% and the bottom 50% on the list respectively. We set the ground truth of gender distribution $T$ as $\{\text{female} : 0.5, \text{male} : 0.5\}$.

The bias mitigation performance of the three proposed algorithms (with 1000 runs) and a widely used fair top-k ranking algorithm (epsilon-greedy) in relevance-aware swapping algorithm can mitigate bias by introducing randomness on the uniform dataset, because the original list has already been randomized entirely. For the same reason, FA*IR also fails to improve the fairness of the uniform dataset. On heavy-headed and heavy-tailed datasets, if more randomness is introduced (a larger $\epsilon$ in epsilon-greedy algorithm and a larger $\rho$ in relevance-aware swapping algorithm), the bias is more mitigated. FA*IR also reduces gender bias significantly. The fairness-greedy algorithm performs best on all three datasets.

**Evaluation on Real-world Data**

We conducted adversarial attacks on gender fairness in four major image search engines, where various gender distributions are observed for the same search term. We also found that image search engines are sensitive to the search term variants that convey the same semantics. Finally, we evaluated the performances of the three proposed bias mitigated algorithms on the collected dataset. This subsection presents the details of these evaluations.

**Gender Bias in Cross-culture Search Engines**

Besides the Google image search engine, we evaluated the same occupation terms in Baidu from China, Naver from South Korea, and Yandex from Russia. Using the hybrid image gender detection method (see the subsection of Gender Detection), similar to Google, all the above three image search engines are deemed to have a gender bias with search terms that include ‘United States’ in them (see Figure 3(c) where a positive value indicates over-representing females and a negative value indicates under-representing females). The effectiveness of the proposed adversarial attack approach in cross-culture search engines is demonstrated in Figure 3(a) to Figure 3(d) where the difference in female ratios between search terms with and without ‘United States’ is evident, especially among the top 50 items. We can also observe that distinct occupations demonstrate different gender distribution patterns in the same search engine, and the same occupation may demonstrate different patterns across search engines. These findings led us to consider that such gender bias exists across cultures and needs attention globally.

**Sensitive to Variant Search Terms**

Another evidence of the unsystematic mitigation of gender bias is that image search engines are sensitive to variant search terms. As shown in Figure 4, the female ratios of image search results between CEO and chief executive officer are significantly different, especially when search terms include ‘United States.’ However, with the increase of top $k$, the difference in the female ratio demonstrates a trend to become stable and small, especially for search terms containing ‘United States.’ (see Figure 4(b)).
Table 1: Bias mitigation performance on synthetic datasets. The bias value in the table is measured by Equation 1.

|          | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|          | $\epsilon = 0.2$ | $\epsilon = 0.4$ | $\epsilon = 0.6$ | $\alpha = 0.2$ | $\alpha = 0.4$ | $\alpha = 0.6$ | $\rho = 0.2$ | $\rho = 0.4$ | $\rho = 0.6$ | $\rho = 0.8$ | $\rho = 1.0$ | $\rho = 1.2$ | $\rho = 1.4$ |
| Uniform  | 0.056    | 0.064    | 0.072    | 0.080    | 0.088    | 0.096    | 0.104    | 0.112    | 0.120    | 0.128    | 0.135    | 0.142    |
| Heavy-headed | 0.123    | 0.131    | 0.139    | 0.147    | 0.155    | 0.163    | 0.171    | 0.179    | 0.187    | 0.194    | 0.202    | 0.209    |
| Heavy-tailed | 0.160    | 0.168    | 0.176    | 0.184    | 0.192    | 0.200    | 0.208    | 0.216    | 0.224    | 0.232    | 0.240    | 0.248    |

Table 2: Bias mitigation performance on Google occupation image datasets. The bias value is measured by Equation 1.

|          | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original | Original |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|          | $\epsilon = 0.2$ | $\epsilon = 0.4$ | $\epsilon = 0.6$ | $\alpha = 0.2$ | $\alpha = 0.4$ | $\alpha = 0.6$ | $\rho = 0.2$ | $\rho = 0.4$ | $\rho = 0.6$ | $\rho = 0.8$ | $\rho = 1.0$ | $\rho = 1.2$ | $\rho = 1.4$ |
| biologist U.S. | 0.138    | 0.146    | 0.154    | 0.162    | 0.170    | 0.178    | 0.186    | 0.194    | 0.202    | 0.209    | 0.217    | 0.224    |
| ceo U.S. | 0.172    | 0.180    | 0.188    | 0.196    | 0.204    | 0.212    | 0.220    | 0.228    | 0.236    | 0.244    | 0.252    | 0.260    |
| comp. programmer U.S. | 0.114    | 0.122    | 0.130    | 0.138    | 0.146    | 0.154    | 0.162    | 0.170    | 0.178    | 0.186    | 0.194    | 0.202    |
| cook U.S. | 0.149    | 0.157    | 0.165    | 0.173    | 0.181    | 0.189    | 0.197    | 0.205    | 0.213    | 0.221    | 0.229    | 0.237    |
| engineer U.S. | 0.04    | 0.05    | 0.06    | 0.07    | 0.08    | 0.09    | 0.10    | 0.11    | 0.12    | 0.13    | 0.14    | 0.15    |
| nurse U.S. | 0.115    | 0.123    | 0.131    | 0.139    | 0.147    | 0.155    | 0.163    | 0.171    | 0.179    | 0.187    | 0.195    | 0.203    |
| police officer U.S. | 0.049    | 0.057    | 0.065    | 0.073    | 0.081    | 0.089    | 0.097    | 0.105    | 0.113    | 0.121    | 0.129    | 0.137    |
| prim. school teacher U.S. | 0.135    | 0.143    | 0.151    | 0.159    | 0.167    | 0.175    | 0.183    | 0.191    | 0.199    | 0.207    | 0.215    | 0.223    |
| software developer U.S. | 0.189    | 0.197    | 0.205    | 0.213    | 0.221    | 0.229    | 0.237    | 0.245    | 0.253    | 0.261    | 0.269    | 0.277    |
| truck driver U.S. | 0.056    | 0.064    | 0.072    | 0.080    | 0.088    | 0.096    | 0.104    | 0.112    | 0.120    | 0.128    | 0.135    | 0.142    |

Gender Bias Mitigation We deployed the three proposed algorithms on the image search datasets collected from Google, Baidu, Naver, and Yandex. To illustrate how the proposed algorithms work, we take the epsilon-greedy algorithm as an example to show the dynamic fairness achievements on ‘biologist’ datasets, as shown in Figure 5. As $\epsilon$ increases (i.e., more randomness is introduced), the gender distribution of the re-ranked list becomes more likely to be different from the original one (see the shaded range), implying more fairness will be achieved if the raw image search list suffers from severe gender bias. With the increase of top $k$, the female ratio becomes more stable and finally converges when top $k$ reaches 200.

![Figure 5: Performance of the epsilon-greedy algorithm on Google, Baidu, Naver, and Yandex ‘biologist’ datasets.](image)

Similar to the evaluations on synthetic datasets, we explored the performance of our algorithms and FA*IR (Zehlike et al. 2017) on real-world datasets. Table 2 illustrates the gender mitigation performance of each algorithm on 10 Google image datasets, which were collected with the search keywords of 10 occupations plus ‘United States.’ When the original bias is larger than 0.1 (e.g., biologist United States), gender bias normally decreases along with the increase of $\epsilon$ in the epsilon-greedy algorithm and $\rho$ in the relevance-aware swapping algorithm. However, if the original bias is small (e.g., engineer United States), epsilon-greedy algorithm and relevance-aware swapping algorithm cannot mitigate gender bias. We can observe that the fairness-greedy algorithm consistently achieves a low bias because it gives the highest priority to fairness during re-ranking. FA*IR also demonstrates a stable and good performance regardless of the original bias. In addition, comparing the result columns of Original and Fairness-greedy in Table 2 can tell the degree of gender bias hidden in the original image list.

Conclusion and Limitation

Bias in AI systems has become an increasingly prevalent and complex issue to address. Often the system developers fix a problem by creating a superfluous solution without addressing the underlying issue. In this paper, we used an adversarial query attack method by appending additional information like country names to trigger potential gender bias in image search. An open-sourced Cross-search-engine Image Retrieval Framework (CIRF) was developed to retrieve data from Google, Baidu, Naver, and Yandex. To recognize the gender of people in photos, five popular image detection APIs, namely Amazon Rekognition, Luxand, Face++, Microsoft Azure, and Facebook DeepFace, were evaluated. Although these APIs are endorsed by AI giants, they could not always handle images in the wild with high accuracy. Therefore, a hybrid method combining automatic gender detection APIs and crowdsourced human workforce was designed to label image genders. To mitigate gender bias, we proposed three lightweight and interpretable re-ranking algorithms and evaluated their performance on both synthetic and real-world datasets. Our results demonstrated that it is possible (and advisable) to address bias in image search (and perhaps in other types of search as well) in a systematic, sustainable, and more meaningful way than doing individual query fixes in an ad hoc fashion.

In this paper, we treated gender as a binary attribute inferred by either gender APIs or humans. However, we acknowledge that gender is different from biological sex, and is non-binary. It is also something that a third-party — be it a human or a program — is not always in a position to detect genders correctly. Our reliance on the binary gender norm and third-party annotation is a limitation of this research.
References


Griffith, E. 2021. Algorithms Still Have a Bias Problem, and Big Tech Isn’t Doing Enough to Fix It.


