INFOSEEKING

UNIVERSITY of WASHINGTON

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

Yunhe Feng, Chirag Shah
Information School
University of Washington, Seattle

Image Search

Source: www.aeroadmin.com

- Image search engines provide an important information-seeking interface for people to explore the world.
- Google processes more than 3.5 billion queries per day and 1.2 trillion searches per year[1].
- Image search results can significantly influence how people perceive and view the world.

[1] https://www.internetlivestats.com/google-search-statistics/

Gender Bias in Image Search

UW NEWS

ENGINEERING | NEWS RELEASES | RESEARCH | TECHNOLOGY

April 9, 2015

Who's a CEO? Google image results can shift gender biases

Jennifer Langston

UW News

Getty Images last year created a new online image catalog of women in the workplace that countered visual stereotypes on the Internet of moms as frazzled caregivers rathe powerful CEOs.

A new University of Washington study adds to those efforts by assessing how accurate gender representations in online image search results for 45 different occupations ma reality.

In a few jobs — including CEO — women were significantly underrepresented in Google image search results, the study found, and that can change searchers' worldviews. Across all the professions, women were slightly underrepresented on average.

Percentage of women in top 100 Google image search results for CEO: 11%
Percentage of U.S. CEOs who are women: 27%

Gender & Technology CHI 2015, Crossings, Seoul, Korea

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations

Matthew Kay
Computer Science & Engineering | dub, University of Washington
mjskay@uw.edu

Cynthia Matuszek
Computer Science & Electrical Engineering, University of Maryland Baltimore County
cmat@umbc.edu

Sean A. Munson
Human-Centered Design & Engineering | dub, University of Washington
smunson@uw.edu

[1] Kay M, Matuszek C, Munson SA. Unequal representation and gender stereotypes in image search results for occupations. InProceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 2015 Apr 18 (pp. 3819-3828).

Gender Bias in Image Search of *CEO* is Fixed

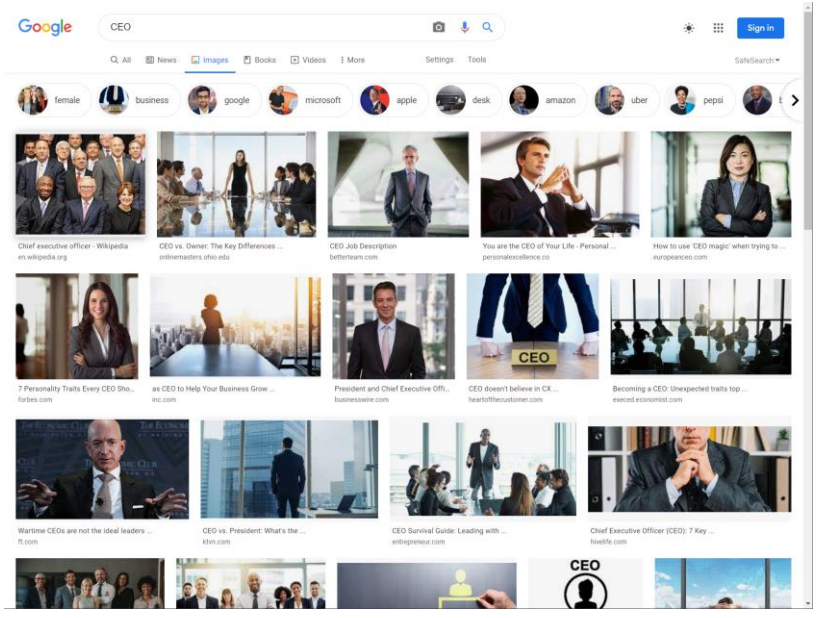


Image search results by Google (males and females)

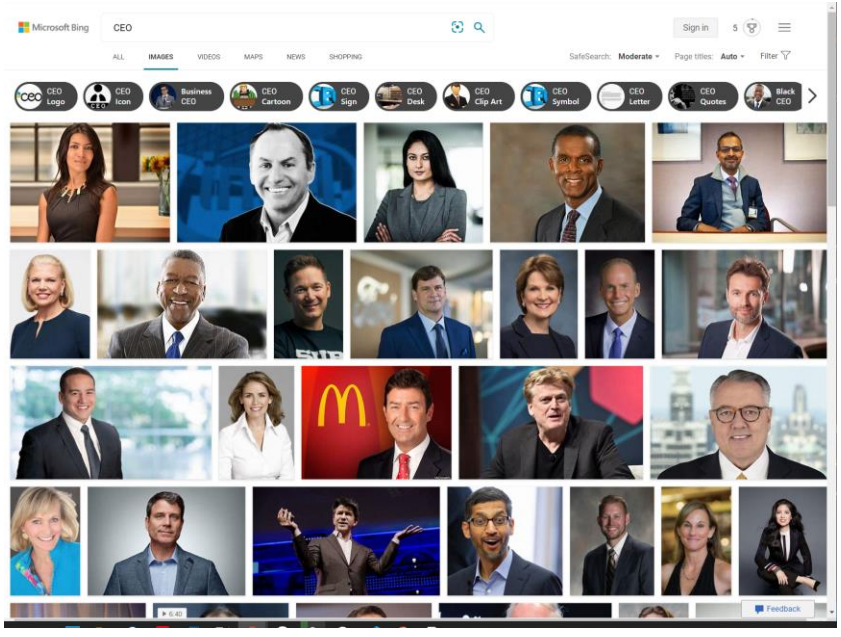


Image search results by Bing (males and females)

Gender Bias in Image Search of *CEO* is Fixed

Ground Truth: 29.3% from www.bls.gov

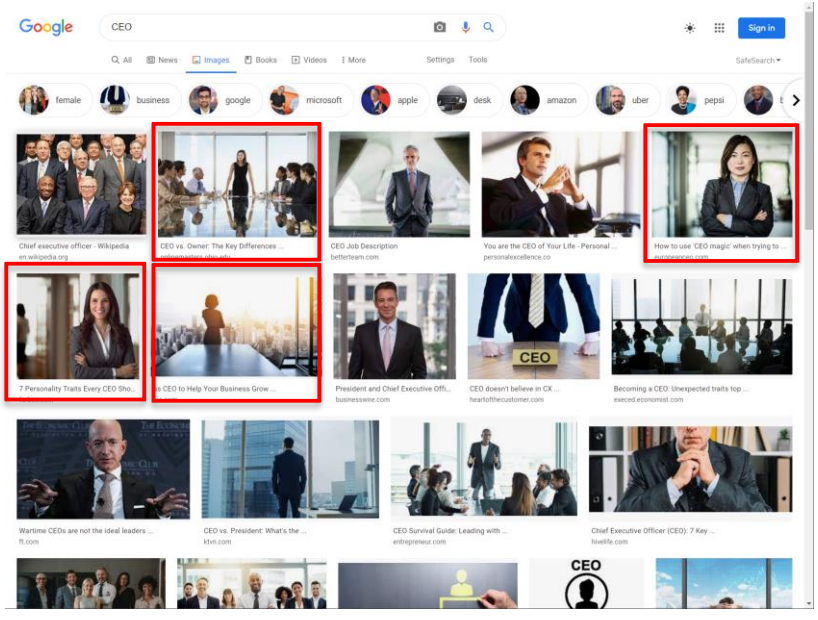


Image search results by Google (males and females)

Female ratio 4/14 = 28.57%

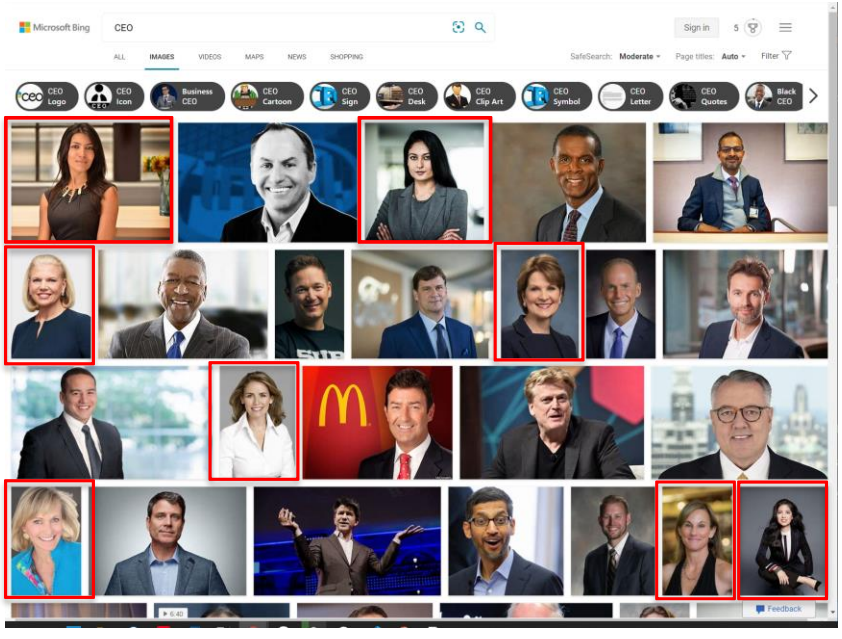


Image search results by Bing (males and females)

Female ratio 8/24 = 33.33%

Has CEO Gender Bias Really Been Fixed Systematically?

- **Adversarial Attacks**
- **Search Term + Location (Occupation + Country)**
  - CEO United States
  - CEO UK
  - CEO China
  - CEO South Korean
  - CEO Russia

Image Search Results of *CEO United States*

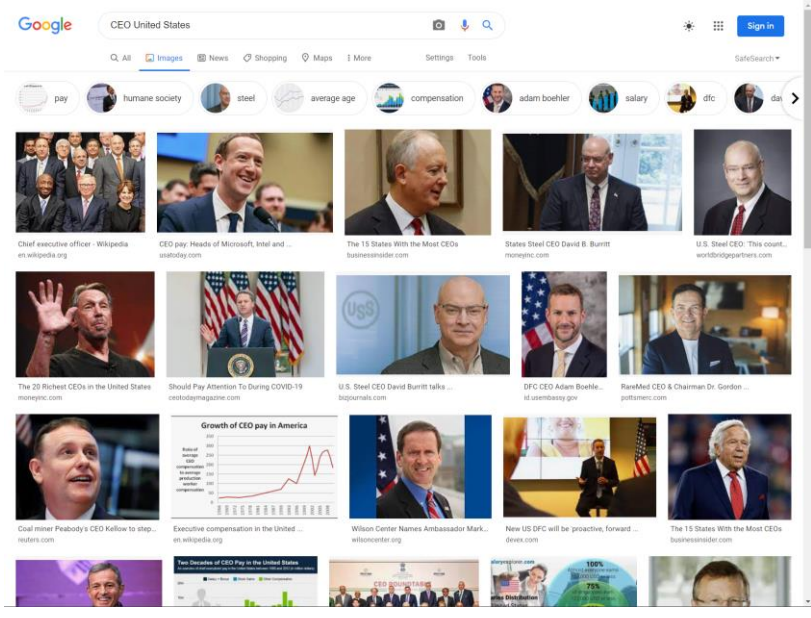


Image search results by Google (all males)

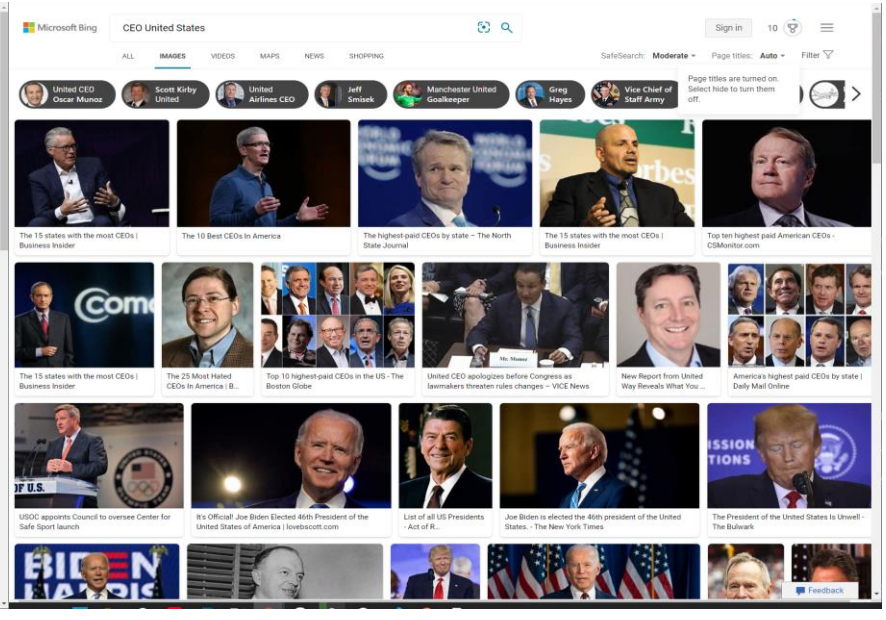


Image search results by Bing (all males)

Image Search Results of *CEO United States*

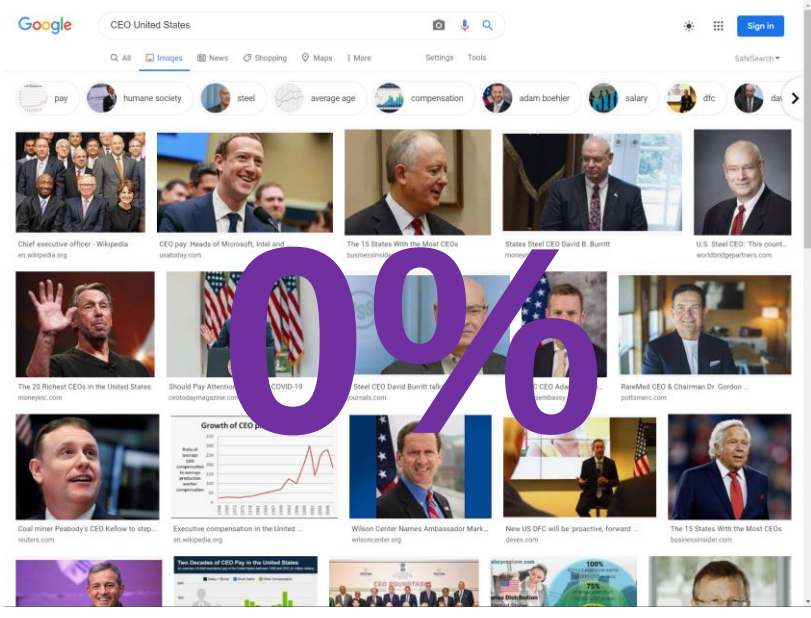


Image search results by Google (all males)

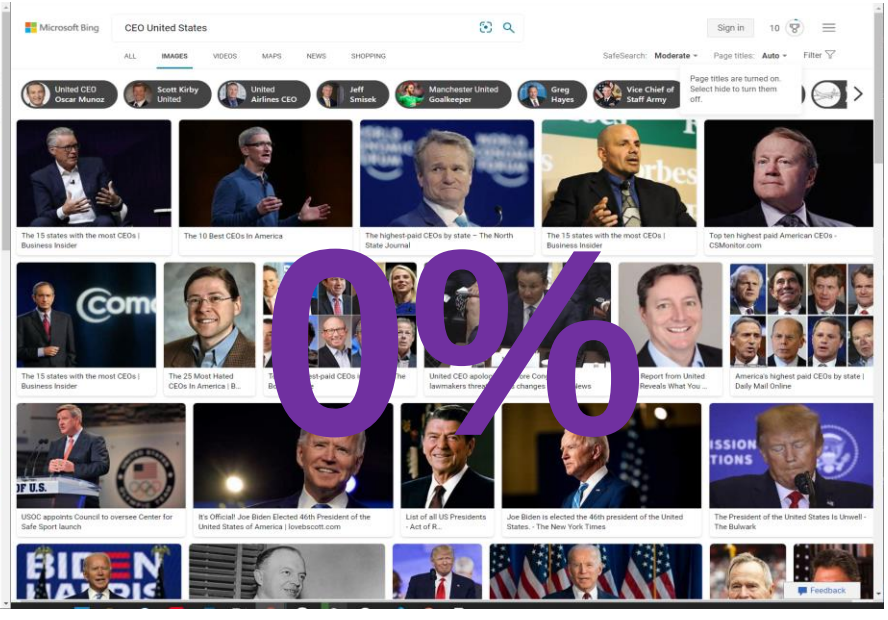


Image search results by Bing (all males)

Image Search Results of *CEO UK*

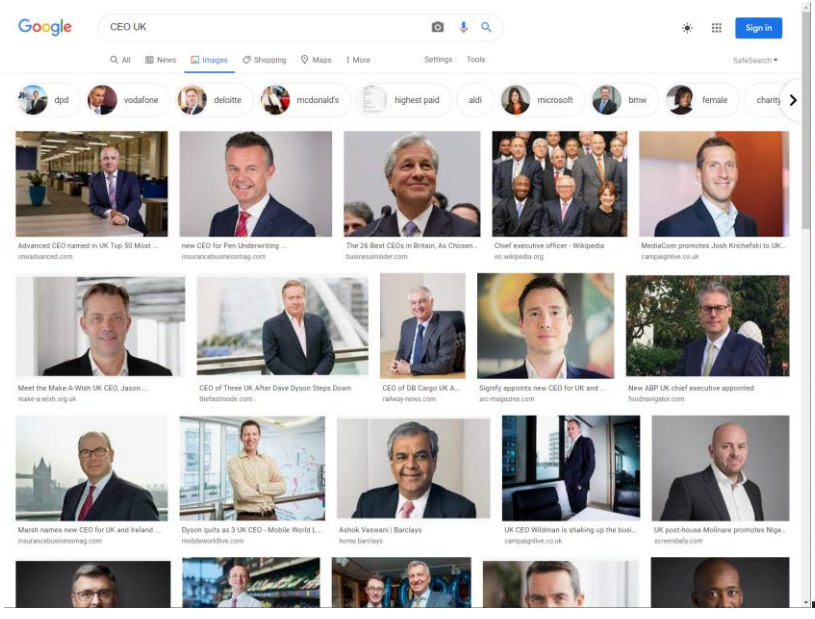


Image search results by Google (all males)

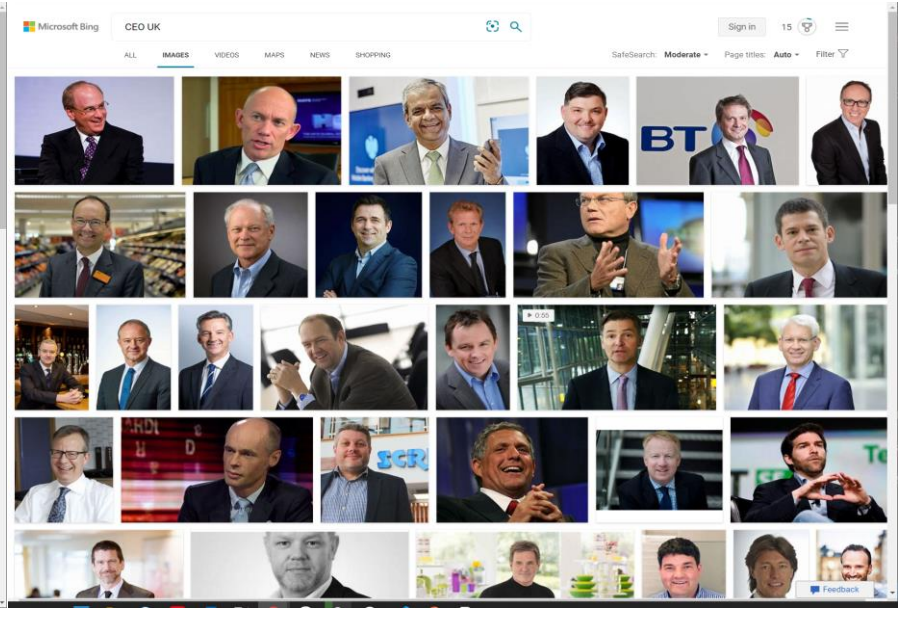


Image search results by Bing (all males)

Image Search Results of *CEO UK*

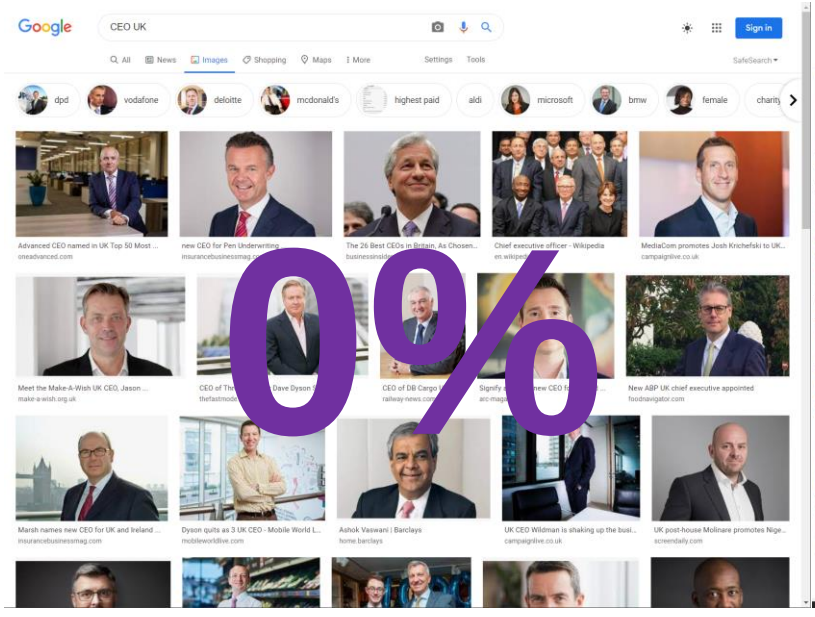


Image search results by Google (all males)

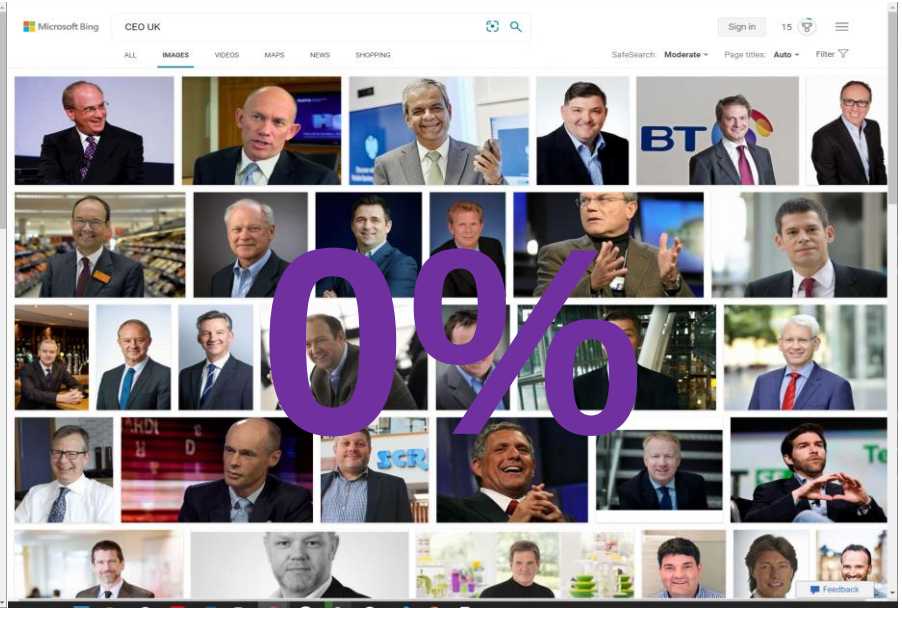


Image search results by Bing (all males)

Scalable Investigation of Gender Bias in Image Search

- **More Occupation Keywords and Search Engines**
- **Automatic Data Collection Framework**
  - Cross-search-engine Image Retrieval Framework (CIRF)
- **Gender Detection**
  - IRB-approved Amazon Mechanical Turk (MTurk) user study
  - Image-based gender detection APIs

Image Collection Using Occupation Keywords

- biologist U.S.
- biologist
- chief executive officer U.S.
- chief executive officer
- computer programmer U.S.
- computer programmer
- cook U.S.
- cook
- engineer U.S.
- engineer
- nurse U.S.
- nurse
- police officer U.S.
- police officer
- primary school teacher U.S.
- primary school teacher
- software developer U.S.
- software developer
- truck driver U.S.
- truck driver

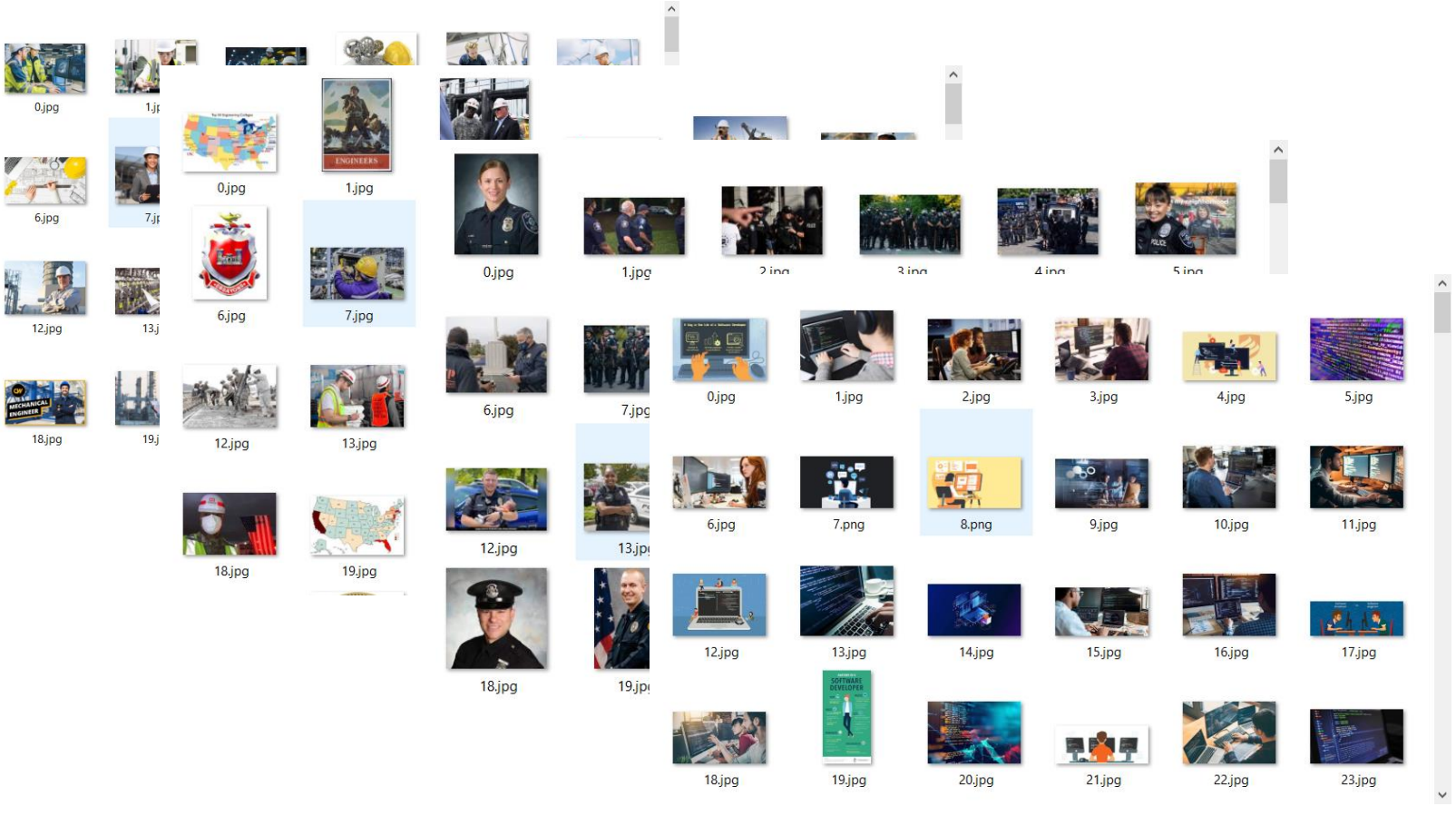
Cross-search-engine Image Retrieval Framework (CIRF)

- **URL Builder**
  - Google - https://www.google.com/search?q=**keyword**\&source=lnms\&tbm=isch
  - Baidu - https://image.baidu.com/search/index?tn=baiduimage&word=**keyword**
  - Naver - https://search.naver.com/search.naver?where=image&sm=tab_jum&query=**keyword**
  - Yandex - https://yandex.com/images/search?text=**keyword**

- **Data Downloader**
  - Selenium WebDriver – open URLs in Chrome with incognito mode
  - PyAutoGUI – save the HTML file and supplementary materials

- **Image Parser**
  - Standard images
  - Base64 encoded images
  - Image URLs

Image Collection Using Occupation Keywords

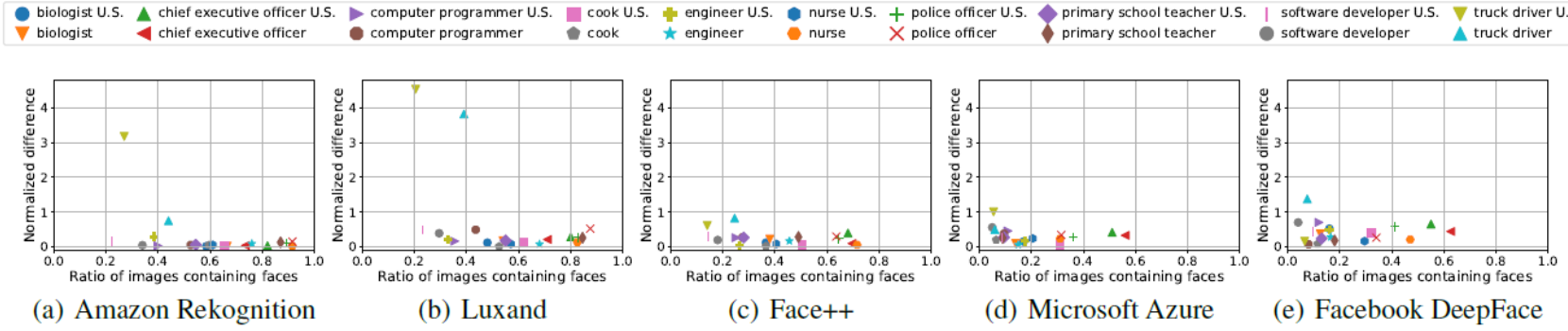


Gender Detection

- **IRB-approved MTurk user study**
  - Paid each participant $0.5 for annotating 50 images
  - Each image was assigned to three workers.

- **Five popular image-based gender detection APIs/models**
  - Amazon Rekognition APIs
  - Luxand APIs
  - Face++ APIs
  - Microsoft Azure Face APIs
  - Facebook DeepFace

Normalized Female Ratio Difference btw. MTurk and APIs



(a) Amazon Rekognition (b) Luxand (c) Face++ (d) Microsoft Azure (e) Facebook DeepFace

Figure 2: Normalized female ratio difference (compared with MTurk results) vs. the ratio of detected faces in images.

- **When the face detection ratio is above 0.5, the normalized female ratio difference between MTurk results and Amazon Rekognition is below 15%.**

- **A two-step hybrid method to annotate image gender labels**
  - Use Amazon Rekognition to detect image genders
  - For search terms that suffer from a low face detection ratio (below 0.5), we still rely on MTurk to manually label them.

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.

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

For top *k* images returned by search engines, we calculate the Kullback-Leibler divergence $D_{KL}(T \,||\, P^k)$ between these *k* images and the ground truth. The average Kullback-Leibler divergence is used to represent the existing bias.

Algorithms to Mitigate Gender Bias

- **Epsilon-greedy Algorithm**
  - Simplicity and Generalizability.
- **Relevance-aware Swapping Algorithm**
  - Consider the relevance of search items during re-ranking.
- **Fairness-greedy Algorithm**
  - Considering more than 90% of users do not go past the first page of the Google search results and the first three items displayed in Amazon search results account for 64% of all clicks.
  - Narrow the difference in gender distributions between top-ranked images and the ground truth by moving images up and down.

Epsilon-greedy Algorithm

Algorithm 1: Epsilon-greedy Algorithm

Input: $\mathbf{L}$: the original image list; ϵ: the probability of swapping two items;
Output: $\mathbf{R}$: the re-ranked image list;
$\mathbf{R} \leftarrow \varnothing$; // initialize $\mathbf{R}$ as empty
for $i = 1 \rightarrow |\mathbf{L}|$ **do**
  $p \leftarrow$ a random number between 0 and 1;
  if $p <= \epsilon$ **then** // swap items
    $temp \leftarrow \mathbf{L}_i$;
    $j \leftarrow$ a random number between $i + 1$ and $|\mathbf{L}|$;
    $\mathbf{L}_i \leftarrow \mathbf{L}_j$;
    $\mathbf{L}_j \leftarrow temp$;
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add swapped item
  else // keep the original item
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add unswapped item
  end
end
return $\mathbf{R}$

Inspired by:

[2] Gao R, Shah C. Toward creating a fairer ranking in search engine results. Information Processing & Management. 2020 Jan 1;57(1):102138.

The ε-greedy exploration in Reinforcement Learning.

[3] Berry DA, Fristedt B. Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability). London: Chapman and Hall. 1985 Oct;5(71-87):7-.

[4] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018 Nov 13.

Relevance-aware Swapping Algorithm

Relevance Weight Modeling

$$\mathbf{W}_i = \frac{1 - \frac{i}{|\mathbf{L}|}}{\log_2(i+1)}$$

Swapping Probability

$$\rho(1 - \mathbf{W}_i)$$

Algorithm 2: Relevance-aware Swapping Algorithm

Input: $\mathbf{L}$: the original image list; ρ: the sensitivity of swapping two items;
Output: $\mathbf{R}$: the re-ranked image list;
$\mathbf{R} \leftarrow \varnothing$; // initialize $\mathbf{R}$ as empty
for $i = 1 \rightarrow |\mathbf{L}|$ **do**
  $\mathbf{W}_i = \frac{1 - \frac{i}{|\mathbf{L}|}}{\log_2(i+1)}$; // relevance weight
  $p \leftarrow$ a random number between 0 and 1;
  if $p <= \rho * (1 - \mathbf{W}_i)$ **then** // swap items
    $temp \leftarrow \mathbf{L}_i$;
    $j \leftarrow$ a random number between $i + 1$ and $|\mathbf{L}|$;
    $\mathbf{L}_i \leftarrow \mathbf{L}_j$;
    $\mathbf{L}_j \leftarrow temp$;
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add swapped item
  else // keep the original item
    append $\mathbf{L}_i$ to $\mathbf{R}$; // add unswapped item
  end
end
return $\mathbf{R}$

Fairness-greedy Algorithm

- Need to know the ground truth of gender distribution *T* (i.e., the gender distribution of search terms in real life) and a list of gender labels *G* for retrieved images.
- Step 1: identify the most underrepresented feature x_{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_{min} in $L_{i \to |L|}$ and move it forward as the new L_i(see line 17-27).

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 ← [L_1] ;                          // initialize R as [L_1]
4  for i = 2 → |L| do
5      P ← gender distribution on [G_{R_1}, ..., G_{R_{i-1}}];
6      flag ← False;
7      x_min ← None ;    // most underrep. feat.
8      𝒞 ← ∅ ;    // set checked features as ∅
9      while (flag = False) and (𝒞 ≠ 𝒳) do
10         d_min ← 0;
11         add x_min to 𝒞;                  // update 𝒞
           /* select most underrep. feature  */
12         for x ∈ 𝒳 − 𝒞 do
13             d = P_x − T_x ;   // diff. in feat. x
14             if d ≤ d_min then
15                 x_min ← x ;   // underrep. feat.
16         end
           /* find 1st item w/ underrep. feat.  */
17         for j = i → |L| do
18             if G_{L_j} = x_min then // find an item
19                 temp ← L_j ;              // save L_j
20                 for k = i + 1 → j do
21                     L_k ← L_{k−1} // move down
22                 end
23                 L_i ← temp ;            // update L_i
24                 append L_i to R ;        // update R
25                 flag ← True ;  // find the item
26                 break;
27         end
28     end
29 end
30 return R
```

Evaluation on Synthetic Data

- **Uniform Dataset**
  - female and male items are distributed evenly across the whole list
- **Heavy-headed Dataset**
  - female items are aggregated at the top of the list
- **Heavy-tailed Dataset**
  - female items are aggregated at the bottom of the list

Evaluation on Synthetic Data

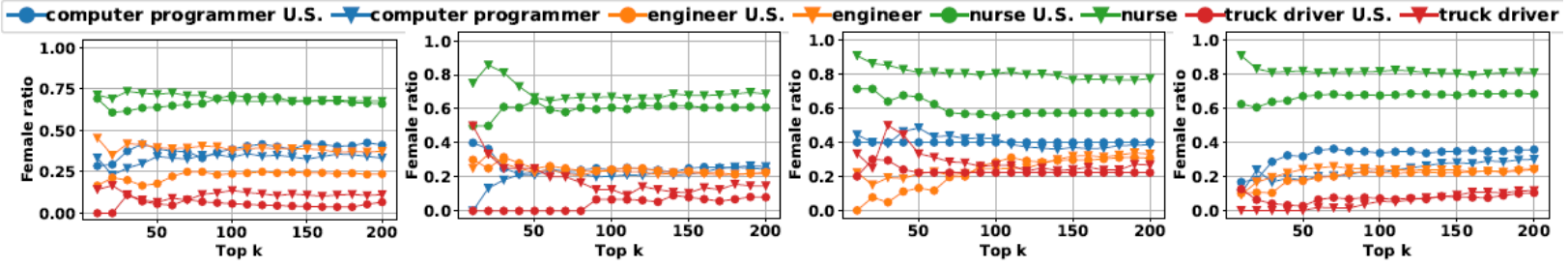
Table 1: Bias mitigation performance on synthetic datasets. The bias value in the table is measured by Equation 1.

	Original	Epsilon-greedy			Relevance-aware Swapping			Fair-greedy	FA*IR
		ϵ=0.2	ϵ=0.4	ϵ=0.6	ρ=0.2	ρ=0.4	ρ=0.6		p=0.5 α=0.1
Uniform	0.066	0.059±0.019	0.055±0.025	0.052±0.028	0.065±0.013	0.064±0.018	0.063±0.022	**0.020**	0.066
Heavy-headed	2.046	0.426±0.189	0.203±0.107	0.105±0.063	0.553±0.222	0.316±0.143	0.198±0.095	**0.020**	0.142
Heavy-tailed	2.046	0.423±0.199	0.194±0.096	0.102±0.061	0.548±0.219	0.312±0.136	0.198±0.098	**0.020**	0.142

- **As female and male items are evenly distributed in the Uniform dataset, epsilon-greedy, relevance-aware swapping, and FA*IR [5] algorithms can not mitigate bias.**
- **As female and male items are aggregated at the top (bottom) in the Heavy-headed (Heavy-tailed) dataset, the bias is more mitigated when more randomness is introduced.**

[5] Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R. Fa* ir: A fair top-k ranking algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management 2017 Nov 6 (pp. 1569-1578).

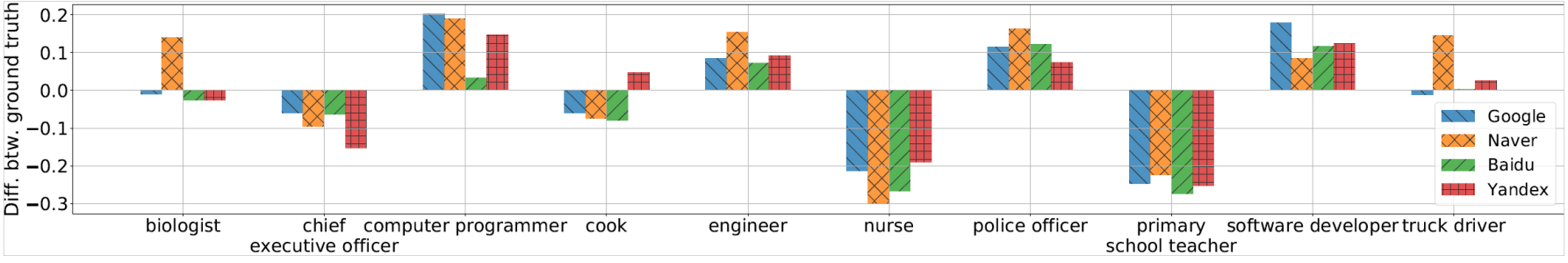
Female Ratio of *Occupation* VS *Occupation + United States*



(a) Google (w/ and w/o U.S.) (b) Baidu (w/ and w/o U.S.) (c) Naver (w/ and w/o U.S.) (d) Yandex (w/ and w/o U.S.)

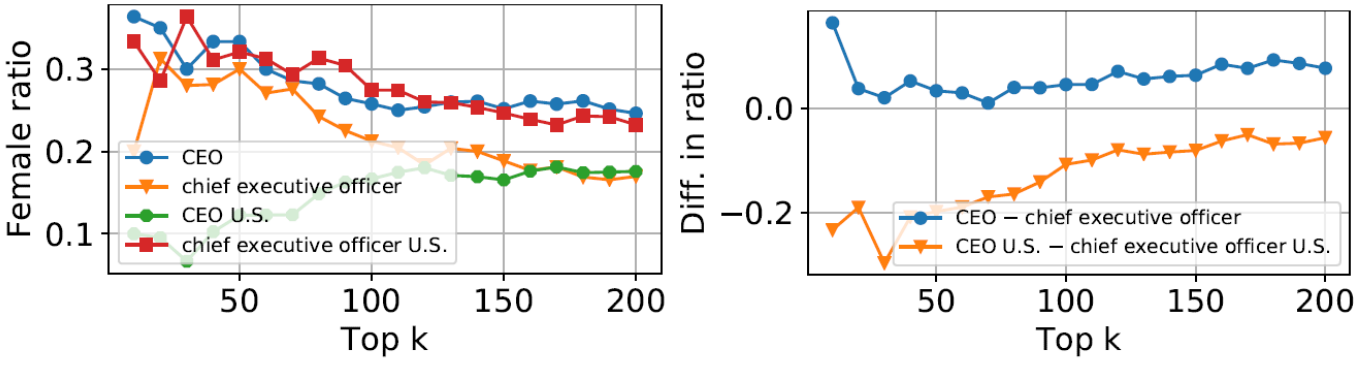
- **The difference in female ratios between search terms with and without 'United States' is evident, especially among the top 50 items.**
- **Distinct occupations demonstrate different gender distribution patterns in the same search engine.**
- **The same occupation may demonstrate different patterns across search engines.**

Female Ratio of *Occupation + United States* VS Ground Truth



- **+positive value indicates over-representing females**
- **-negative value indicates under-representing females**

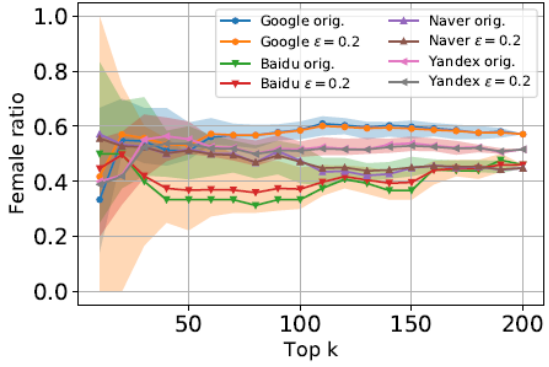
Sensitive to Variant Search Terms



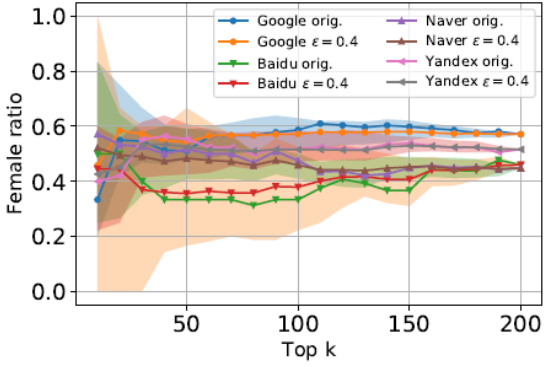
(a) CEO vs Chief Exec. Officer (b) Difference in female ratio

- **Female ratios between CEO and chief executive officer are significantly different, especially when search terms include 'United States.'**
- **With the increase of top k, the difference in female ratio demonstrates a trend to becomes stable and small.**

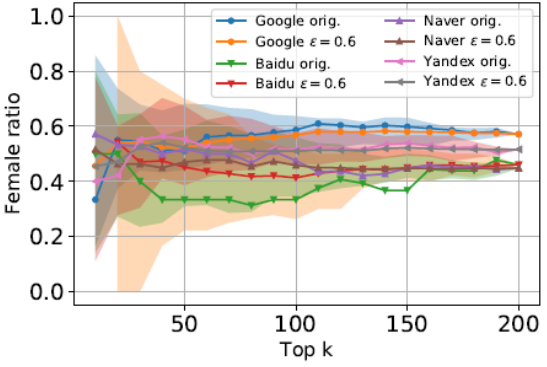
Performance of Epsilon-greedy Algorithm



(a) epsilon = 0.2



(b) epsilon = 0.4



(c) epsilon = 0.6

- **With the increase of *epsilon*, the gender distribution of the re-ranked list becomes more likely to be different from the original one.**
- **With the increase of top *k*, the female ratio becomes more stable and finally converges when top *k* reaches 200.**

Evaluation on Real-world Data

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

	Original	Epsilon-greedy			Relevance-aware Swapping			Fair-greedy	FA*IR
		ϵ=0.2	ϵ=0.4	ϵ=0.6	ρ=0.2	ρ=0.4	ρ=0.6		p=0.5 α=0.1
biologist U.S.	0.138	0.102±0.044	0.087±0.046	0.071±0.049	0.128±0.032	0.108±0.046	0.114±0.046	**0.018**	0.072
ceo U.S.	0.172	0.175±0.055	0.160±0.082	0.144±0.087	0.169±0.048	0.167±0.052	0.160±0.054	**0.021**	0.084
comp. programmer U.S.	0.114	0.119±0.027	0.120±0.030	0.135±0.062	0.113±0.021	0.114±0.030	0.120±0.035	**0.034**	0.071
cook U.S.	0.149	0.131±0.051	0.109±0.064	0.101±0.070	0.148±0.049	0.133±0.052	0.128±0.064	**0.017**	0.102
engineer U.S.	0.04	0.044±0.011	0.053±0.019	0.063±0.036	0.045±0.022	0.048±0.016	0.052±0.022	**0.02**	0.027
nurse U.S.	0.115	0.119±0.011	0.119±0.015	0.128±0.023	0.118±0.009	0.121±0.015	0.124±0.017	**0.066**	0.076
police officer U.S.	0.049	0.053±0.015	0.054±0.016	0.055±0.018	0.048±0.008	0.047±0.011	0.046±0.013	**0.015**	0.088
prim. school teacher U.S.	0.135	0.136±0.007	0.136±0.010	0.137±0.011	0.137±0.006	0.136±0.008	0.137±0.009	0.1	**0.085**
software developer U.S.	0.189	0.193±0.066	0.171±0.078	0.156±0.082	0.193±0.035	0.180±0.061	0.184±0.067	**0.055**	0.094
truck driver U.S.	0.056	0.067±0.044	0.088±0.062	0.088±0.067	0.070±0.044	0.074±0.048	0.087±0.064	**0.007**	0.02

- **When the original bias is larger than 0.1 (e.g., biologist United States), gender bias normally decreases along with the increase of ϵ in the epsilon-greedy algorithm and ρ in the relevance-aware swapping algorithm.**
- **If the original bias is small (e.g., engineer United States), epsilon-greedy algorithm and relevance-aware swapping algorithm cannot mitigate gender bias.**
- **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.**

Limitation and Future Work

- **Treat gender as a binary feature, which is not True in our real world.**
- **Study Culture Factors in Image Search.**
- **Adversarial Auditing Commercial Facial Recognition Systems from the Perspective of Fairness and Trustworthiness.**

UNIVERSITY of WASHINGTON

Thank you!
Q&A