Perceptions in Pixels: Analyzing Perceived Gender and Skin Tone in Real-world Image Search Results

Jeffrey Gleason
gleason.je@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Ronald E. Robertson
ronalder@stanford.edu
Stanford University
Palo Alto, California, USA

Christo Wilson
cbw@ccs.neu.edu
Northeastern University
Boston, Massachusetts, USA

ABSTRACT

The results returned by image search engines have the power to shape peoples' perceptions about social groups. Existing work on image search engines leverages hand-selected queries for occupations like "doctor" and "engineer" to quantify racial and gender bias in search results. We complement this work by analyzing peoples' real-world image search queries and measuring the distributions of perceived gender, skin tone, and age in their results. We collect 54,070 unique image search queries and analyze 1,481 open-ended people queries (i.e., not queries for named entities) from a representative sample of 643 US residents. For each query, we analyze the top 15 results returned on both Google and Bing Images.

Analysis of real-world image search queries produces multiple insights. First, less than 5% of unique queries are open-ended people queries. Second, fashion queries are, by far, the most common category of open-ended people queries, accounting for over 30% of the total. Third, the modal skin tone on the Monk Skin Tone scale is two out of ten (the second lightest) for images from both search engines. Finally, we observe a bias against older people: eleven of our top fifteen query categories have a median age that is lower than the median age in the US.

CCS CONCEPTS

• **Information systems** → **Image search**; **Query log analysis**.

KEYWORDS

image search; gender; skin tone; information retrieval; algorithm auditing

ACM Reference Format:
Jeffrey Gleason, Avijit Ghosh, Ronald E. Robertson, and Christo Wilson. 2024. Perceptions in Pixels: Analyzing Perceived Gender and Skin Tone in Real-world Image Search Results. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3589334.3645666



This work is licensed under a Creative Commons Attribution International 4.0 License.

WWW '24, May 13–17, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0171-9/24/05.
https://doi.org/10.1145/3589334.3645666

1 INTRODUCTION

Search engines are widely trusted sources of information [4], but the fact that people trust them grants them power to shape peoples' perceptions. For example, prior work has found that the political information presented in search engine result pages (SERPs) may influence voting behavior [6, 7], and that the demographics of people that appear in image search results can alter perceptions of social groups [14, 24, 43]. The fact that search engines like Google measure their audience in billions means that these systems must be carefully scrutinized to understand the potential impacts they may have on individuals and society [25].

In this work, we focus specifically on representational harms [37] in image search results. There is a robust literature on these harms that begins with Kay et al. [14] and continues through many other studies [22, 24, 28, 30, 40, 42]. This literature has primarily leveraged controlled experiments in which authors send hand-selected queries to image search engines to uncover scenarios where the images in SERPs are biased along gender and racial lines. As a result, existing scholarship focuses on narrow categories of queries, such as queries for occupations (e.g., "doctor" and "engineer") [14, 24] or queries that include gender-neutral adjectives (e.g., "intelligent") [30, 40]. This narrow focus on occupation queries was also adopted in a recent study of images produced by generative models [20].

In contrast to existing work that focuses on hand-selected queries, this study analyzes representational harms in response to peoples' real-world image search queries. Specifically, we focus on ecologically-valid *open-ended people queries* that real people send to search engines. We define a *people query* as a query that produces a SERP where a fraction of the images contain people. We define an *open-ended* people query as a people query that does not predetermine the identities of people in the results. For example, queries for named entities (e.g., "Taylor Swift") are not open-ended and we would not expect the resulting images to be demographically diverse. Further, queries that contain demographic words (e.g., "fall outfits women") are not open-ended because they circumscribe the results that a search engine might return. In contrast, open-ended people queries offer image search engines the opportunity to produce a diverse results page—whether they do or not determines whether their output causes representational harm.

To implement our study, we collect 54,070 unique image search queries and analyze a subset of 1,481 open-ended people queries from a representative sample of 643 US residents. For each of these

queries, we analyze the top 15 results returned on Google and Bing Image Search. We apply a series of filters (e.g., named entity recognition) to identify open-ended people queries in our dataset and then measure the distributions of perceived gender, skin tone, and age in the corresponding results pages. This approach enables us to understand demographic representation in image search results under real-world, ecological conditions, and compare representativeness of results between Google and Bing.

Using this dataset, we investigate four research questions:

- **RQ1**: What are the most popular categories for open-ended people queries?
- **RQ2**: How representative, in terms of perceived gender, skin tone, and age, are results for open-ended people queries?
- **RQ3**: Are there differences in representativeness across Google and Bing or across categories?
- **RQ4**: To what extent do people use demographic words (e.g., 'Black' or 'women') to refine people queries, and what kinds of people engage in this refinement?

Overall, our results show that open-ended people queries are relatively rare, accounting for less than 5% of unique queries. However, results for these queries, on both Google and Bing, demonstrate systematic skews toward lighter skin tones and away from older people. Further, our results highlight new categories of queries (e.g., fashion) that controlled studies have not explored extensively, pointing the way for future algorithm audits.

The outline of our study is as follows: in §2 we discuss related work on image search engines. We introduce our datasets in §3 and present our methods in §4. We present the results of our analysis in §5 and conclude in §6.

2 BACKGROUND

We begin by presenting an overview of work on representational issues and harms in the context of image search engines.

2.1 Representation in Image Search

There is a robust literature on representational harms [37] in image search engines. In their seminal 2015 study, Kay et al. [14] examined representation of men and women in Google Image Search results in response to queries for occupations. They found that search results for many occupations overrepresented men relative to their baseline level of employment from government statistics. Furthermore, users judged images that matched gender stereotypes (e.g., a man as a doctor) as more 'professional' and 'appropriate.' Otterbacher et al. [30] found similar representational and stereotyping issues when they queried Bing for a 'person' that had various attributes. Men were overrepresented in search results when agentic adjectives (e.g., 'competent', 'decisive') modified the query, but women were overrepresented when 'warm' adjectives modified the query. Ulloa et al. [40] observed similar overrepresentation of men when they added the adjective 'intelligent' to image queries on Google, Yandex, and Baidu. Additionally, they observed *face-ism* in search results from these search engines, a stereotype in which photos of men tend to focus on the face, while photos of women include a greater proportion of the body.

Other work on representation in image search results focuses on race and the intersection of race and gender. Noble [28] catalogued many queries that returned racist, sexist, and stereotypical results on Google. Metaxa et al. [24] replicated and expanded the Kay et al. [14] experiment, finding that White people are even more overrepresented than men in Google Image Search results for occupations. Urman et al. [42] studied the representation of migrants in response to English and German queries across six image search engines: Google, Bing, Baidu, Yandex, Yahoo, and DuckDuckGo. They found that non-White people were overrepresented, while women were underrepresented. Finally, Makhortykh et al. [22] found a predominantly White portrayal of Artificial Intelligence across the same six search engines.

2.2 Effects of Representation

Researchers have consistently found that demographic representation in image search results can impact peoples' perceptions of search result quality. Multiple studies have confirmed that participants rate image search results as higher-quality when the people in the images conform to gender stereotypes [14, 16], especially when a given participant holds strongly discriminatory views [31].

There is evidence, however, that increasing representation in image search results can lead users to correspondingly shift their views. Kay et al. [14] found that manipulating gender representation in search results for occupational queries shifted users' estimates of gender proportions in occupations by ∼7%. Metaxa et al. [24] replicated this finding, and also demonstrated that manipulating gender and racial representation affected users' level of interest in an occupation, their perception of its inclusivity, and expectations about feeling valued in that occupation. The importance of perception is echoed by Mitchell et al. [26], who present metrics to measure *diversity* and *inclusion* that go beyond traditional group fairness metrics. Finally, Vlasceanu and Amodio [43] demonstrated that manipulating gender representation affected participants' decisions in a hypothetical hiring scenario.

2.3 Situating Our Study

Existing work clearly demonstrates that demographic representational harms exist in image search engines. Work suggesting that interventions may, in the long-term, overcome peoples' initial, negative relevance judgments and meaningfully reshape their views emphasizes the importance of identifying and mitigating representational harms.

Our study is motivated by and builds on prior empirical work in two ways. First, we examine demographic representation in image search results—from Google and Bing Image Search—in response to *ecological queries* from a large, real-world panel of US residents (described in §3). This contrasts with prior studies that have utilized *controlled queries* selected by researchers themselves [14, 22, 24, 28, 30, 40, 42]. As we show in §5, access to ecological queries enables us to identify areas of concern that previous studies have not identified, as well as contextualize the prevalence of known-problematic queries (e.g., about occupations). Second, we expand the set of demographic traits from prior work by examining representation in terms of perceived gender, skin tone, and age.

Prior work on representation in image search results has framed its findings around 'bias' [14, 30]. According to Friedman and Nissenbaum [8], a computer system has a problematic normative bias

			Searches/User/Day	
Search Engine	**№ Users**	**№ Searches**	**Mean**	**Std**
Google Images	607	93510	4.89	31.69
Bing Images	127	13754	11.66	57.76

Table 1: Summary statistics from image query dataset.

		Participants		
		N	**%**	**US Census**
Gender	Female	334	51.9	50.4
	Male	310	48.1	49.6
Race/Ethnicity	White	518	80.4	58.9
	Black	49	7.6	13.6
	Hispanic	34	5.3	19.1
	Asian	14	2.2	6.3
	Native American	1	0.2	1.3
	Two or more	13	2.0	3.0
	Other	15	2.3	–
Age	< 18	0	0.0	21.7
	18-64	507	78.7	50.4
	≥ 65	137	21.3	17.3

Table 2: Demographics of participants who contributed image search queries.

if it "systematically and unfairly discriminates against certain individuals or groups." Crucially, assessing bias requires a normatively defensible baseline against which to judge a given system. In §5, we use baselines drawn from the US Census to assess bias in image search results with respect to perceived gender and age.

3 DATA COLLECTION

In this section we introduce the image query and image search result datasets that facilitate our study.

3.1 Image Search Queries

From August to December 2020, we worked with the survey company YouGov to recruit a nationally representative sample of 2,000 US residents. Participants answered survey questions about their demographics and 926 people opted to install a browser extension that we developed for Chrome and Firefox. Our extension collected multiple types of data from participants' web browsers, but we only analyze participants' browsing histories in this study. Our study was IRB approved and §6.2 describes participants' protections. This dataset has been used in other studies to study online behavior pertaining to Alphabet products [3, 9, 35].

We identified and extracted queries that participants made on Google and Bing Image Search using the URL structures of these services.[1] We ignored consecutive URLs with identical queries, which represented user interactions with the initial search, e.g., clicking on an image thumbnail. Table 1 shows the total number of users, total number of searches, and summary statistics about user daily activity on each image search engine. We define a participant as a user of a search engine if they made at least one search during our observation window on that search engine. According to this definition, 66% of our participants use Google Images, 14% use Bing Images, and 10% use both. Overall, we observe 107,264 total image searches and 54,302 unique queries from 644 participants across Google and Bing. Table 2 describes the demographics of these participants: they are substantially Whiter (80.4% vs. 58.9%) and older (by virtue of none being under 18) than the US population.

[1] google.com/search?tbm=isch&q=QUERY and bing.com/images/search?q=QUERY.

			Images/Query	
Search Engine	**№ Screenshots**	**№ Images**	**Mean**	**Std**
Google Images	54211	2510331	46.31	8.09
Bing Images	54127	2688838	49.68	2.70

Table 3: Summary statistics from image search crawls.

3.2 Image Search Results

We developed an open-source web crawler[2] that collected the top 50 image search results from both Google and Bing for each unique query that our participants issued. The crawler iterated through queries in a random order to minimize spillover and used a 1920×1080 viewport, which is the most common desktop screen resolution.[3] The crawler collected two types of image data—(1) full-page screenshots and (2) individual image files along with their metadata, e.g., position on the SERP—and saved both as Base64 encoded images. Table 3 shows the total number of screenshots and image files collected, as well as summary statistics about the number of images returned per query. Overall, we collect image data from both Google and Bing for 54,070 unique queries.

We ran the crawler in August 2022, without a user profile, from an IP address in Boston. Changes in online services over time imply that we observed different images in some cases than our participants. The effects of geolocation and user profile are unknown, though prior work finds that user profile has a limited effect on general Google Search results, while location has effects, but only for location-relevant queries [11, 15]. It's unclear whether these results hold for image search engines. All of that said, like other audits of image search engines, our analysis accurately represents results for a set of queries, from a specific location, at a specific point in time.

4 METHODS

In this section, we describe how we identified and categorized open-ended people queries and how we labeled the demographics of people in a sample of images.

4.1 Identifying Open-Ended People Queries

We applied four filters—listed in Table 4—to isolate and validate a set of open-ended people queries from our larger query corpus.

4.1.1 Detecting People Queries. We use YOLOv3 [32], an object detection model pretrained on the COCO dataset [18], to detect the number of people in each image in our corpus.[4] We summarize these inferences at the query-level by measuring the fraction of

[2] https://github.com/jlgleason/google-image-scraper

[3] https://gs.statcounter.com/screen-resolution-stats/desktop/worldwide

[4] Appendix §7.1 demonstrates that person detections from YOLOv3 strongly agree with those from a state-of-the-art object detection model.

	Filtering Step	№ Queries	Query Fraction	№ Users
	Original sample	54070	1.00	643
1.	>= 25% of images have people	21539	0.40	550
2.	Not named entity	4387	0.08	415
3.	Safe for work	3728	0.07	404
4.	Manual review	1481	0.03	296

Table 4: Summary of sample size after each filtering step.

images on the corresponding SERP that contain at least one person. Figure 1 presents a histogram of this distribution for each search engine. In 45% of Bing SERPs and 42% of Google SERPs, fewer than 10% of images have people. At the other end of the spectrum, more than 90% of images have people in only 21% of Bing SERPs and 17% of Google SERPs.

We choose a conservative threshold and only remove queries where fewer than 25% of images on either Google or Bing have people. This leaves us with 21,539 queries (40%) that are potentially people related.

4.1.2 Filtering Named Entities. When filtering named entities, our goal is to minimize the number of false negatives, i.e., queries labeled as open-ended, but which actually have a named entity. We make this decision because the demographics of images returned for a query with a named entity, e.g., "Taylor Swift", are predetermined.

We combine three labeling approaches and remove the query if any approach identifies a named entity:

(1) We use the named entity predictions from a spaCy CNN model pre-trained on the Ontonotes dataset [44].[5] This model identifies entities in 14,693 (68%) of the people queries.
(2) We search the query on general Google Search and label it a named entity if the results page contains a `knowledge-panel` or a `top-image-carousel` (see [29] and [9] for examples of these SERP components). This approach identifies entities in 7,439 (35%) of the people queries.
(3) We label the query a named entity if it contains the keyword 'meme' or 'gif'. We observed that these queries often returned a specific meme or gif with a specific person. This approach identifies entities in 728 (3%) of the people queries.

This leaves us with 4,387 open-ended people queries (20% of people queries and 8% of all queries).

4.1.3 Filtering NSFW Images. It is important to analyze open-ended people searches that return not-safe-for-work (NSFW) images to audit sexualization of racial and gender groups [28, 41]. However, we choose to remove these images from our study because we hire crowd workers to label perceived demographics (see §4.1.4) and we choose not to risk exposing them to sexual images. We use Yahoo's OpenNSFW model [21] to identify NSFW images, which is one of the best performing models for CSAM detection [17].[6] Specifically, we make NSFW predictions for each image on a SERP and filter out queries where the average NSFW probability is >= 20%. This approach flags 659 (15%) of open-ended people queries as NSFW.

[5]Appendix §7.2 evaluates the sensitivity of named entity predictions to the specific choice of model.
[6]https://github.com/bhky/opennsfw2

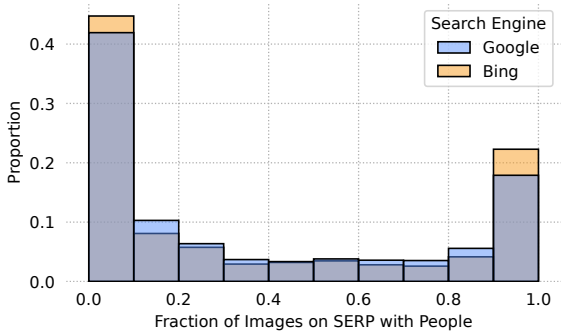


Figure 1: Histogram of fraction of images per SERP containing at least one person.

4.1.4 Expert Manual Review. Two authors manually reviewed the remaining 3,728 purportedly safe-for-work (SFW), open-ended people queries to identify any remaining false negatives. Specifically, we built an application that presented the Google and Bing full-page screenshots for each query and allowed the authors to review the automated (a) named entity and (b) NSFW labels. The two authors had a Cohen's $\kappa = 0.7$ on a random sample of 93 named entity labels, which is considered substantial [23]. Additionally, the two authors (c) recorded the presence of demographic words related to race, ethnicity, and gender (e.g., "Black" or "women") in the query, and (d) removed queries that were not sufficiently people-focused (e.g., focused on cars) or that might be triggering to crowd workers.

Overall, we identify 1,673 (45%) of the remaining queries as named entities, 49 (1%) as NSFW, 306 (8%) as having demographic words, and 209 (6%) as not sufficiently people-focused. This leaves us with 1,481 SFW, open-ended people queries (3% of all queries).

4.2 Categorizing Open-Ended People Queries

One goal of our study is to examine demographic representation in image search results stratified by query category. To implement this goal, we categorize open-ended people queries according to the second level of the WordNet Domains hierarchy (Table 8 shows the full taxonomy) [2]. We made a handful of modifications to the taxonomy after exploratory analysis of our queries. Specifically we added three categories (`food`, `gastronomy`, and `animals`), removed two (`play` and `alimentation`), and changed two (`sport` → `sports` and `earth` → `earth science`).

We assign a query to a category by computing the cosine similarities between an embedding of the query and embeddings of each category name. The embeddings are generated using a pre-trained language model that was fine-tuned to identify semantically similar sentence pairs [33].[7] We select the category with the maximum cosine similarity. Formally, let q represent the query, K represent the set of WordNet category names, and f represent the pre-trained language model. Our classification approach is:

$$\underset{k \in K}{\operatorname{argmax}} \frac{f(q) \cdot f(k)}{\|f(q)\| \|f(k)\|}.$$

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

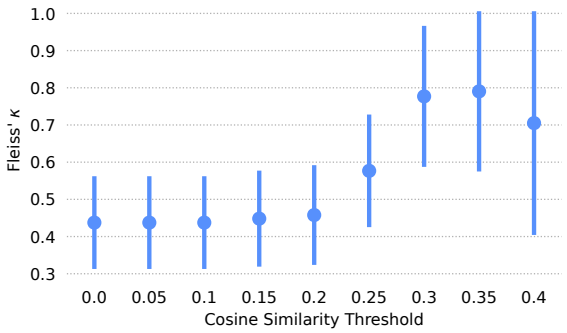


Figure 2: Category assignment agreement as cosine similarity threshold varies. Bars show 95% confidence intervals.

Figure 4a plots the distribution of queries over categories, which we discuss in §5.2. Table 5 shows example queries (where ≥ 2 participants searched the query, to protect individual privacy). Finally, to evaluate our categorization approach, Figure 2 plots Fleiss' κ scores between three labelers (two authors and the cosine similarity method) on a random sample of 54 queries, as we vary the cosine similarity threshold. We see that the point estimate for agreement is ≥ 0.7 when the cosine similarity is ≥ 0.3.

4.3 Labeling Open-Ended People Queries

The final step in our methodology is to obtain demographic labels for a sample of images in our corpus. Similar to prior work, we hire crowd workers to do this task [14, 24]. Our labeling task was IRB approved and §6.2 describes crowd workers' protections.

4.3.1 Query Sample. We allocate a $4,500 labeling budget by sampling up to 20 queries (where the cosine similarity between the query and category name is ≥ 0.3) from each of the top 15 categories (see Figure 4a). This produces a sample of 220 total queries. Table 5 shows the number of queries sampled from each category.

We label all images that appear in the top 15 ranks of Google or Bing search results and contain visible face(s). We focus on the top 15 ranks because image search eye-tracking studies that use five-column layouts find that attention is concentrated on the first three rows [19, 45, 46]. We require visible faces to maximize annotator agreement, especially of perceived skin tone. Specifically, we detect faces using a multi-task CNN model [48] that is pre-trained on the FDDB [13] and WIDER FACE [47] datasets.[8] We label up to three people in each image and at least two workers label each person. Before labeling, we de-duplicate images that have an embedding similarity ≥ 0.95 according to a CNN model pre-trained on ImageNet [12, 36]. Table 5 shows the average number of faces per image in each category.

4.3.2 Label Weights. Because images contain multiple people, receive multiple annotations, and attract varying attention, we apply the following weights when computing means in §5:

(1) Each annotator of a person gets equal weight.
(2) Each person in an image gets equal weight.
(3) Each image gets a weight corresponding to its rank on the SERP using the click rate distribution from Lu and Jia [19].

[8] Appendix §7.3 demonstrates that face detections from this CNN model strongly agree with those from a state-of-the-art face detection model.

Category	Example Queries	№ Queries Sampled	Average № Faces/ Image
fashion	tuxedo, face mask, masks	20	1.2
military		20	1.6
politics		20	3.5
art	photo caption, tattoo	18	1.6
medicine	covid patients	18	1.5
sports	exercise, stretching	18	2.7
children		17	1.5
food	sitting	17	1.7
body care		15	1.9
psychology	conversation, optimistic	11	1.6
sexuality		11	1.9
telecomms		10	2.1
veterinary		10	1.4
pedagogy		8	3.1
tourism		7	3.0

Table 5: Example queries, number of sampled queries, and average number of faces per image in top 15 query categories. We only provide example queries where ≥ 2 participants searched the query, to protect individual privacy.

Label Type	Fleiss' κ 95% CI	Weights
Gender Presentation	(0.81, 0.85)	Identity
Skin Tone	(0.44, 0.52)	Quadratic
Age	(0.79, 0.83)	Quadratic

Table 6: Fleiss' κ agreement statistics between labelers.

4.3.3 Mechanical Turk Task Specification. We define a Mechanical Turk task where workers (1) report their gender, skin tone, and age, and (2) label the perceived gender, skin tone, and age of ten people. Figure 7 shows our labeling interface. For gender, we provide four labels: feminine presenting, masculine presenting, non-binary presenting, and unsure. We collect skin tone labels using the Monk Skin Tone Scale, which has ten levels and better represents darker skin tones [27]. Google Research introduced this scale in 2022 and uses it for machine learning labeling and fairness testing [34]. We ask for age as a positive integer.

Each batch of ten images contains one attention check, for which we compare workers' labels against labels from two authors. 96% of workers provided the same perceived gender label as the two authors. 95% of workers were within three skin tones of the authors' skin tone range. 99% of workers were within ten years of the authors' age range.

We require that workers were located in the US, had an approval rate > 98%, and had completed 1,000 HITs. We paid workers $3.75 per HIT and the median time to complete the task was 16 minutes, which translates to $14 per hour. In total, 683 unique workers provided labels.

Finally, Table 6 presents 95% confidence intervals for Fleiss' κ scores between labelers. Skin tone and age are ordinal scales and therefore we use quadratic weights to evaluate agreement, which penalize large disagreements more than small ones. That said, agreement for perceived skin tone is much lower than agreement for perceived gender and age. This disagreement is also non-random:

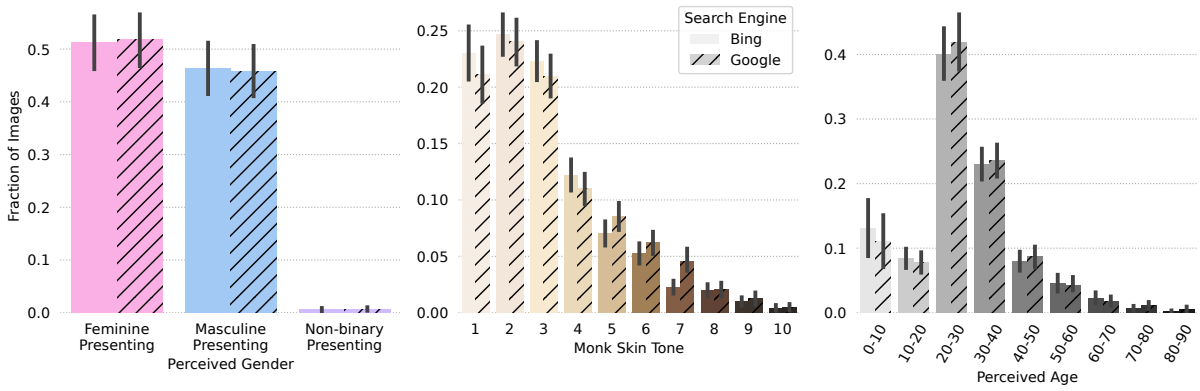


Figure 3: Perceived gender, Monk Skin Tone, and age distributions in image search results across search engines. We compute 95% confidence intervals using the percentile bootstrap with 1000 replications over queries.

workers' skin tones are correlated with their choice of skin tone labels (Pearson's $r = 0.21$).[9] This is important because the workers we recruit skew toward lighter skin tones (see Figure 6). We evaluate the sensitivity of our results to this skew in Appendix §8.

5 RESULTS

This section describes the topical distribution of open-ended people queries and analyzes the distributions of perceived gender, skin tone, and age across search engines and categories. Results in §5.1 and §5.2 are weighted using the approach we describe in §4.3.2.[10]

5.1 Representation Across Search Engines

Figure 3 compares the distributions of perceived gender, skin tone, and age across Google and Bing.[11] Continuous age labels are binned into ten-year age brackets. We compute 95% confidence intervals using the percentile bootstrap with 1000 replications over queries, which is our sampling unit [5].

5.1.1 Gender. Google and Bing have similar perceived gender distributions. Both search engines have slightly higher fractions of feminine than masculine presenting people, but these differences are not distinguishable from zero.

5.1.2 Skin Tone. Search results for both Google and Bing are heavily skewed toward lighter skin tones. The modal perceived skin tone for both search engines is two out of ten. 63–69% of Google images and 67–72% of Bing images have a perceived skin tone ≤ 3. The mean skin tone on Google (3.19) is slightly higher than the mean skin tone on Bing (2.99) (95% CI 0.07–0.35).

5.1.3 Age. The modal perceived age bracket for both search engines is 20–30. Perceived age is ≤ 40 in 81–87% of both Google and Bing images. The 0–10 age bracket represents 7–15% of Google images and 9–17% of Bing images.

[9]The association between workers' genders and their choice of gender labels (Cramer's $V = 0.03$) and between workers' age and their choice of age labels (Pearson's $r = 0.02$) are both small.

[10]Replication material for this section: https://doi.org/10.7910/DVN/ARDEDK

[11]Appendix §9 also examines intersectional distributions for (1) perceived gender and skin tone, and (2) perceived gender and age .

5.2 Representation Across Categories

Figure 4a plots the categorical distribution of all open-ended people queries (i.e., not only queries from our labeled subset). We observe that fashion is by far the most popular category, comprising more than 30% of queries. Art, children, and sports are the next three largest categories, each comprising 5–10% of queries. All other categories account for less than 5% of queries.

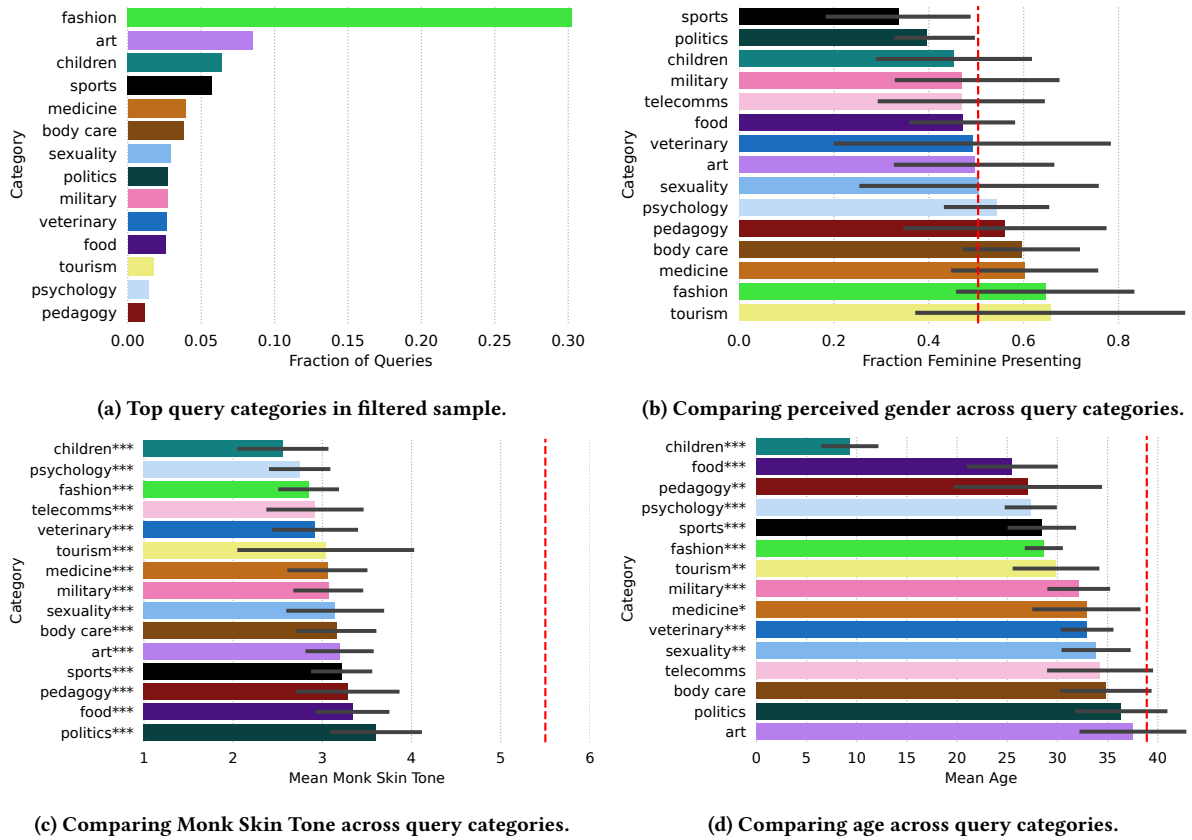
The rest of Figure 4 compares the distributions of perceived gender, skin tone, and age across categories to reference baselines. For each category, we compute 95% confidence t-intervals, where each data point is the average for one query (our sampling unit). We compute p-values using t-tests and correct for multiple comparisons using the Benjamini-Hochberg procedure [1]. We compare the fraction of feminine presenting images to 50.4% (see Table 2) and average age to 38.9 years.[12] In lieu of an existing baseline for the Monk Skin Tone Scale, we use the midpoint of the scale: 5.5.

5.2.1 Gender. Sports has the lowest fraction of feminine presenting images (19–48%), while medicine (45-75%), fashion (46-83%), and tourism (38–94%) have the highest. However, after adjusting for multiple comparisons, none are distinguishable from the US Census baseline.

5.2.2 Skin Tone. All categories have significantly lighter average perceived skin tones ($p < 0.001$) than the midpoint of the Monk Skin Tone Scale. Additionally, confidence intervals for all categories overlap, so we cannot distinguish them from one another.

5.2.3 Age. All but four categories have lower average ages than the US median age. This is expected for the children category, but perhaps surprising for other categories, such as fashion and psychology. Overall, this demonstrates the bias that Google and Bing have away from images of older people.

[12]https://www.census.gov/newsroom/press-releases/2023/population-estimates-characteristics.html



(a) Top query categories in filtered sample.

(b) Comparing perceived gender across query categories.

(c) Comparing Monk Skin Tone across query categories.

(d) Comparing age across query categories.

Figure 4: Perceived gender, Monk Skin Tone, and age across categories. The red lines in (b), (c), and (d) compare each category to a reference baseline: fraction of women from the US Census, the midpoint of the Monk Skin Tone Scale, and the median age from the US Census, respectively. Bars show 95% confidence t-intervals and asterisks represent Benjamini-Hochberg adjusted p-values ($^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

5.3 Use of Demographic Words in Queries

In this section, we analyze participants' use of demographic words related to race, ethnicity, and gender when searching for and refining people queries. Specifically, we analyze 1,481 open-ended people queries without demographic words and 306 people queries with demographic words (identified in §4.1.4).

We construct query refinement sequences by sorting participants' queries in time and comparing the semantic similarity of consecutive queries. For instance, an example sequence is: 'fall outfits' → 'fall outfits for women' → 'fall outfits for black women.' We represent queries using the language model described in §4.2 and measure semantic similarity using the cosine similarity.

Figure 5 compares the probability that a refined query contains a demographic word to the probability that an initial query contains one, as we vary the similarity threshold used to identify refinement sequences. The point estimate for the difference in proportions is positive for all values of the similarity threshold. This indicates that refined people queries are more likely to contain demographic words than initial people queries.

Finally, Table 7 presents results from linear mixed-effects models that regress the use of specific demographic words on participants' self-reported gender and race. We include random effects for participants to control for individual differences in behavior. Specifically, we analyze the use of 'female' words ('woman', 'women', 'female', and 'girl'), 'male' words ('man', 'men', 'male', and 'boy'), and 'Black' words ('Black'). Other demographic words are used infrequently by our participants. We observe that Black participants are substantially more likely to use the word 'Black' in their people queries and male participants are slightly more likely to use 'male' words.

6 DISCUSSION

This study generates new insights about representation on image search engines by focusing on real-world, *open-ended people queries*. First, we find that less than 5% of unique queries are open-ended people searches (i.e., not searches for named entities). This suggests that fairness interventions, which can be computationally expensive [38], are not required for all of an image search engine's traffic. Second, we categorize open-ended people queries and find that fashion is by far the most popular category, accounting for over 30% of queries. Another prominent category is children: 7% of open-ended people queries are children-related and 7–17% of images across Google and Bing fall into the 0–10 age bracket. Fashion

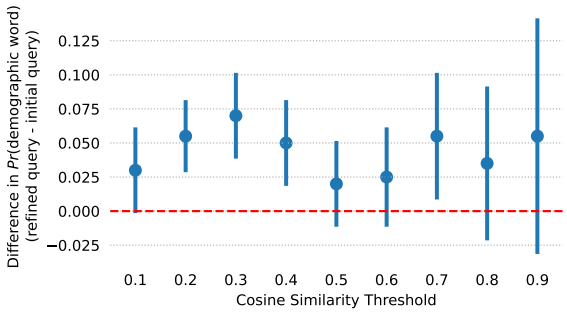


Figure 5: Refined queries are more likely to contain demographic words than the initial query in a sequence. Bars show 95% confidence intervals.

	Demographic Word Use		
	'Female'	'Male'	'Black'
intercept	0.068***	0.018	0.010
	(0.014)	(0.009)	(0.006)
Male	0.005	0.030*	−0.002
	(0.022)	(0.014)	(0.010)
Black	0.009	0.037	0.093***
	(0.057)	(0.036)	(0.026)
Male:Black	−0.065	0.040	0.013
	(0.081)	(0.051)	(0.036)
Observations	3,327	3,327	3,327
Groups	310	310	310
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 7: Regressions of demographic word use on self-reported participant demographics.

and children are two categories that seem ripe for future controlled audits and user perception experiments. Pinterest's focus on fashion and beauty content in their end-to-end diversification of search and recommendation is further evidence for the importance of fashion-related image search.

Our labeled sample of open-ended people queries across categories on Google and Bing also generates findings about perceived skin tone, age, and gender. First, Google and Bing are heavily skewed toward lighter skin tones. Across both search engines, the modal skin tone on the Monk Skin Tone Scale [34] is two out of ten, and around 2/3 of images have a skin tone ≤ 3. Our use of the ten-level Monk Skin Tone Scale, which Google introduced to better represent darker skin tones, emphasizes the concentration of image results at the light-end of the scale. Second, both search engines also demonstrate a bias away from older people. Over 80% of images are of people ≤ 40 and eleven out of fifteen categories have an average age that is significantly lower than the US median age. Age bias is a representational harm that has not yet been studied in controlled settings—e.g., occupational queries—but which could have important effects. We also observe that two popular query categories, sports and fashion, conform to gender stereotypes.

Finally, we explore participants' use of demographic words related to race, ethnicity, and gender in the query refinement process. We find that refined queries are slightly more likely to contain demographic words and that Black participants are significantly more likely to include the word 'Black' in their queries. This suggests that some users might need to use demographic words to arrive at results that better represent them. Thus, image search engines have an opportunity to improve the user experience—a motivation reflected in Pinterest's system overhaul [38].

6.1 Limitations

Our study has several limitations. Our approach to identifying open-ended people queries relies on pre-trained models for person detection, named entity recognition, and NSFW detection, as well as manual review. We didn't incorporate uncertainty from these specific choices into our analyses further downstream. The same is true of our taxonomy for open-ended people queries and the corresponding classification approach. Furthermore, although we leverage real-world image search queries, we acknowledge that 80% of the participants who generated these queries were White and all participants were U. S. residents. Biases in image search results may vary with the demographics and geography of searchers. Similar limitations arise from the sample of crowd workers who annotated our images, as they skewed male and White. Lastly, we operationalize skin tone using a light-to-dark scale, but this fails to incorporate variable skin tone hues. Assessment utilizing a multidimensional scale [39] may uncover more representational problems in image search results.

6.2 Ethics

Our query data collection was approved by the Northeastern IRB under protocol #20-03-04. We informed potential participants about the data our browser extension would collect and asked for their consent to collect this data. Participants were compensated, could revoke consent at any time (none did), and our extension uninstalled itself at the end of the study period. Participant data was encrypted in transit and only project personnel may access it. Due to privacy concerns we cannot release participant data.

Our image labeling protocol was approved by the Northeastern IRB under protocol #22-12-11. We took extensive measures to remove NSFW images from our corpus before seeking labels. That said, out of an abundance of caution, we informed workers about the potential risks of our task (e.g., viewing disturbing images) before they could complete our task. We did not collect identifiable information from workers. We accepted all submissions from workers and compensated them.

ACKNOWLEDGMENTS

The collection of data used in this study was funded in part by the Anti-Defamation League, the Russell Sage Foundation, and the Democracy Fund. This research was supported in part by NSF grant IIS-1910064. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

REFERENCES

[1] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[2] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the workshop on multilingual linguistic resources*. 94–101.

[3] Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. 2023. Subscriptions and external links help drive resentful users to alternative and extremist YouTube videos. *Science Advances* 9, 35 (August 2023).

[4] Edelman. 2022. Edelman Trust Barometer 2022. Daniel J. Edelman Holdings, Inc..

[5] Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

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

[7] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, 2 (November 2017).

[8] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347.

[9] Jeffrey Gleason, Desheng Hu, Ronald E Robertson, and Christo Wilson. 2023. Google the Gatekeeper: How Search Components Affect Clicks and Attention. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 245–256.

[10] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and Computation Redistribution for Efficient Face Detection. In *International Conference on Learning Representations*.

[11] 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*. 527–538.

[12] Tanuj Jain, Christopher Lennan, Zubin John, and Dat Tran. 2019. Imagededup. https://github.com/idealo/imagededup.

[13] Vidit Jain and Erik Learned-Miller. 2010. *Fddb: A benchmark for face detection in unconstrained settings*. Technical Report. UMass Amherst technical report.

[14] 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*.

[15] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 internet measurement conference*. 121–127.

[16] Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022. Do Perceived Gender Biases in Retrieval Results Affect Relevance Judgements?. In *Advances in Bias and Fairness in Information Retrieval*.

[17] Camila Laranjeira da Silva, João Macedo, Sandra Avila, and Jefersson dos Santos. 2022. Seeing without looking: Analysis pipeline for child sexual abuse datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2189–2205.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[19] Wanxuan Lu and Yunde Jia. 2014. An eye-tracking study of user behavior in web image search. In *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*. Springer, 170–182.

[20] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Evaluating societal representations in diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[21] Jay Mahadeokar and Gerry Pesavento. 2016. Open sourcing a deep learning solution for detecting NSFW images. *Retrieved August* 24 (2016), 2018.

[22] Mykola Makhortykh, Aleksandra Urman, and Roberto Ulloa. 2021. Detecting Race and Gender Bias in Visual Representation of AI on Web Search Engines. In *Advances in Bias and Fairness in Information Retrieval*.

[23] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[24] Danaë Metaxa, Michelle A. Gan, Su Goh, Jeff Hancock, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (apr 2021).

[25] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends in Human-Computer Interactions* 14, 4 (Nov. 2021), 272–344.

[26] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 117–123.

[27] Ellis Monk. 2023. The Monk Skin Tone Scale. (2023).

[28] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York university press.

[29] Bruno Oliveira and Carla Teixeira Lopes. 2023. The Evolution of Web Search User Interfaces-An Archaeological Analysis of Google Search Engine Result Pages. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 55–68.

[30] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.

[31] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

[32] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[33] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

[34] Google Research. 2022. Developing the Monk Skin Tone Scale. https://skintone.google/the-scale.

[35] Ronald E. Robertson, Jon Green, Damian J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* 618 (May 2023).

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.

[37] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. [n. d.]. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

[38] Pedro Silva, Bhawna Juneja, Shloka Desai, Ashudeep Singh, and Nadia Fawaz. 2023. Representation Online Matters: Practical End-to-End Diversification in Search and Recommender Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

[39] William Thong, Przemyslaw Joniak, and Alice Xiang. 2023. Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color. arXiv:2309.05148 [cs.CV]

[40] Roberto Ulloa, Ana Carolina Richter, Mykola Makhortykh, Aleksandra Urman, and Celina Sylwia Kacperski. 2022. Representativeness and face-ism: Gender bias in image search. *New Media & Society* (2022).

[41] Aleksandra Urman and Mykola Makhortykh. 2022. "Foreign beauties want to meet you": The sexualization of women in Google's organic and sponsored text search results. *new media & society* (2022), 14614448221099536.

[42] Aleksandra Urman, Mykola Makhortykh, and Roberto Ulloa. 2022. Auditing the representation of migrants in image web search results. *Humanities and Social Sciences Communications* 9, 130 (2022).

[43] Madalina Vlasceanu and David M. Amodio. 2022. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences* 119, 29 (2022).

[44] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA* 23 (2013), 170.

[45] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 275–284.

[46] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based evaluation metrics for web image search. In *The world wide web conference*. 2103–2114.

[47] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5525–5533.

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[49] Zhuofan Zong, Guanglu Song, and Yu Liu. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6748–6758.

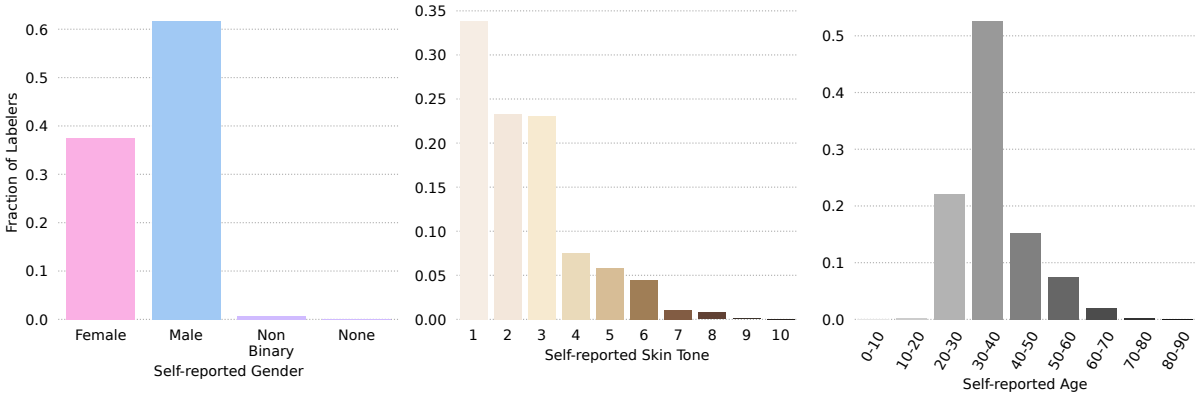


Figure 6: Labeler gender, skin tone, and age distributions.

Top-Level Category	2nd-Level Category
doctrines	archaeology
	astrology
	history
	linguistics
	literature
	philosophy
	psychology
	art
	religion
free time	play
	sport
applied science	agriculture
	alimentation
	architecture
	computer science
	engineering
	medicine
	veterinary
pure science	astronomy
	biology
	chemistry
	earth
	mathematics
	physics
social science	administration
	anthropology
	artisanship
	body care
	commerce
	economy
	fashion
	industry
	law
	military
	pedagogy
	politics
	publishing
	sexuality
	sociology
	telecommunication
	tourism
	transport

Table 8: WordNet Domains Hierarchy

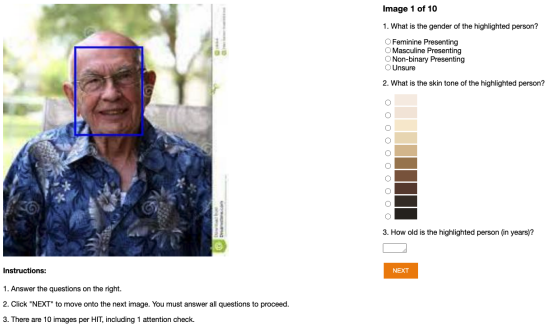


Figure 7: Mechanical Turk labeling interface.

7 SENSITIVITY OF AUTOMATED FILTERING

In this section, we evaluate the sensitivity of automated filtering steps (person detection, named entity recognition, and face detection) to specific choice of model. We evaluate sensitivity by measuring agreement (Cohen's κ) between predictions from the models we use in the main text and predictions from state-of-the-art models as of early 2024.

7.1 Person Detection

We compare predictions from YOLOv3 to predictions from Co-DETR [49] on a sample of ~10,000 images. We select a confidence threshold for Co-DETR such that the total number of detections is approximately the same as the YOLOv3 model. The binary label for the purpose of Cohen's κ is whether the model detects at least one person in the image. Cohen's κ between the two models is 0.86.

7.2 Named Entity Recognition

We compare named entity predictions from a spaCy CNN model[13] to predictions from a spaCy transformer model on the full sample of queries.[14] Specifically, we label a query a named entity if the classifier labels any subsequence of the query string a named entity.

[13]https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-2.3.1
[14]https://github.com/explosion/spacy-models/releases/tag/en_core_web_trf-3.7.2

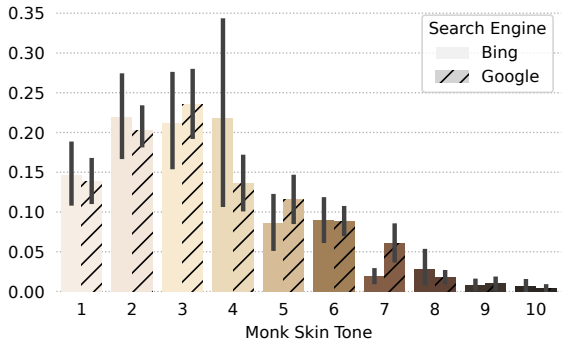


Figure 8: Perceived Monk Skin Tone under simulated set of workers with a uniform skin tone distribution.

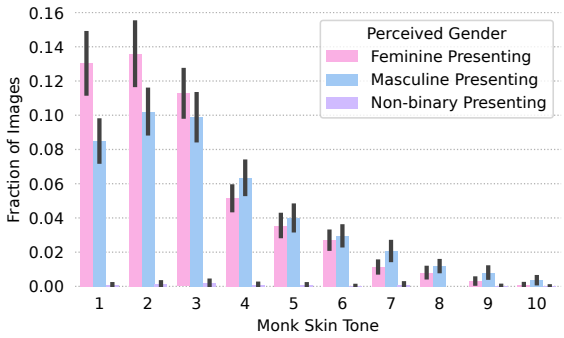


Figure 9: Intersectional distribution of perceived gender and Monk Skin Tone in image search results.

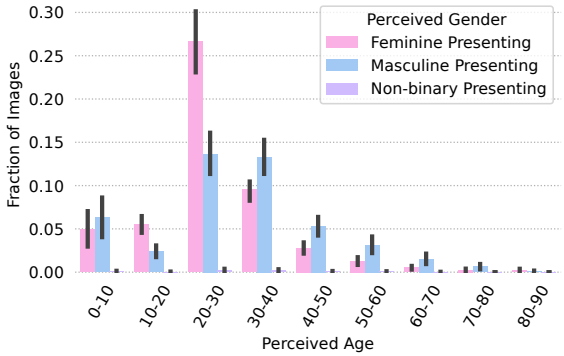


Figure 10: Intersectional distribution of perceived gender and age in image search results.

Cohen's κ between the two models is only 0.46. The relatively low agreement motivates the additional steps we take to minimize named entity false negatives, which we describe in §4.1.2 and §4.1.4.

7.3 Face Detection

We compare predictions from the CNN model to predictions from SCRFD [10] on the full sample of images in results for open-ended people queries. We select a confidence threshold for SCRFD such that the total number of detections is approximately the same as the CNN model. The binary label for the purpose of Cohen's κ is whether the model detects at least one face in the image. Cohen's κ between the two models is 0.70.

8 SENSITIVITY TO LABELER DEMOGRAPHICS

To understand how the correlation between workers' skin tones and their choice of skin tone labels affects our results, we re-weight our labeled data to simulate a set of workers with a uniform skin tone distribution. Specifically, each worker with skin tone s receives the weight:

$$weight_{skin_tone=s} = \frac{\mathbb{P}_{uniform}(skin_tone = s)}{\mathbb{P}_{observed}(skin_tone = s)}$$

where $\mathbb{P}_{uniform}(skin_tone = s) = \frac{1}{9}$ is the probability a worker has skin tone s under a uniform distribution of skin tones (we collected data from workers with skin tones 1–9) and $\mathbb{P}_{observed}(skin_tone = s)$ is the probability a worker has skin tone s under the observed distribution of skin tones.

Figure 8 shows the distribution of perceived Monk Skin Tone across Google and Bing after incorporating this weight. The mean perceived skin tones on Google and Bing increase from 3.19 to 3.55 and 2.99 to 3.41, respectively. However, this does not change our conclusions: images remain heavily skewed toward lighter skin tones, even after adjusting for worker skin tone.

9 INTERSECTIONAL REPRESENTATION

Figure 9 and Figure 10 show the intersectional distributions of (1) perceived gender and skin tone, and (2) perceived gender and age, collapsed over Google and Bing. As before, we compute 95% confidence intervals using the percentile bootstrap with 1000 replications over queries. People with the lightest perceived skin tones (i.e., 1 and 2) are significantly more likely to be feminine presenting. Furthermore, people in the age range 10–30 are significantly more likely to be feminine presenting, while people in the age range 30–70 are significantly more likely to be masculine presenting. Thus, images of people perceived as feminine presenting skew younger and have lighter skin.