Gender Distributions in Google Images Search Results: Beliefs and Statistics

Juan López Martín[1,2], Jack Cao[3], and Mahzarin R. Banaji[3]

[1] Oxford Internet Institute, University of Oxford; [2] Department of Psychology, Autonomous University of Madrid;
[3] Department of Psychology, Harvard University

Abstract

Discovering what people really believe is a complex task, especially in contentious domains like gender. In psychology, beliefs are typically assessed using explicit questions or more indirect, implicit measures. However, advances in machine learning coupled with the availability of big data have surfaced new methods for evaluating beliefs. For instance, word embeddings that capture semantic relationships between words have been shown to reflect human-like biases [1,2].

We capitalized on these advances by scraping the results for more than 200 search terms on Google Images, over 100,000 images in total, and comparing the gender distribution for each search term with people's beliefs and actual statistics. With this approach, we show that internet images reflect beliefs and statistics about gender.

Methods

Search Term Selection

First, 104 **professions** and 62 **college majors** were selected from census data [3,4]. To measure people's beliefs, 23 and 25 participants, respectively, estimated the gender distribution for each profession (e.g. "Percentage of accountants who are men") and college major (e.g. "Percentage of psychology college students who are men").

Second, 12 **alcoholic drinks**, 17 **music genres and instruments**, 15 **personality traits**, and 20 **sports** were selected. Note that these domains are strongly gendered, but contrary to the previous ones there is no easily ascertainable ground truth about gender distributions.

Processing Google Images Search Results

We calculate the proportion of women in the Google Images search results for each of these search terms using the algorithm described below. Additionally, the faces in these results were averaged to form a composite face.

Web Scraping

Download the first 500 Google Images search results

Face Detection

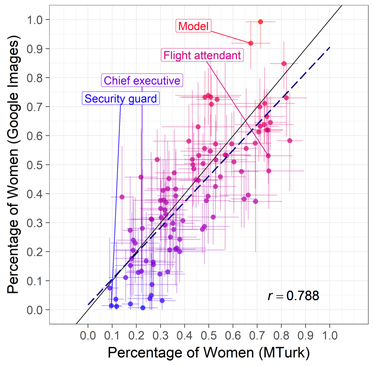
Support Vector Machine trained on Histogram of Oriented Gradient features

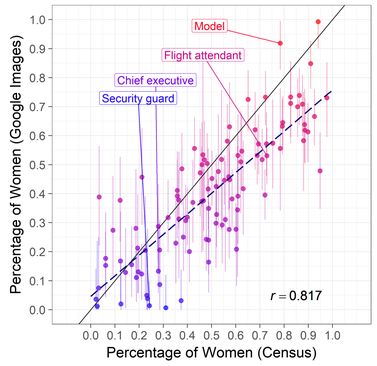
Gender Recognition

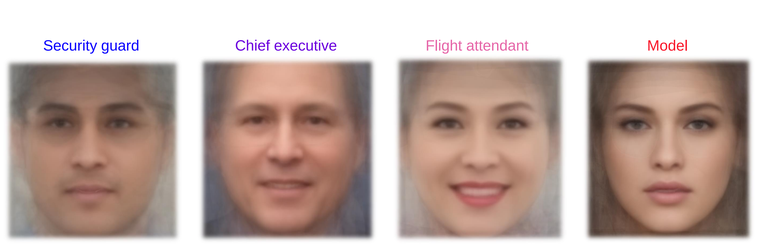
Convolutional Neural Network

Results

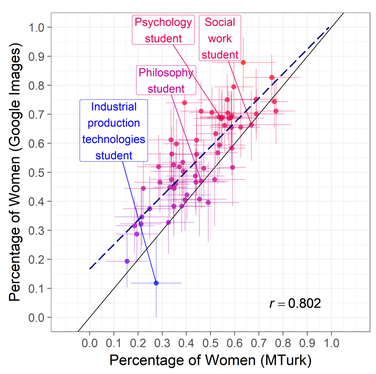
1. Professions

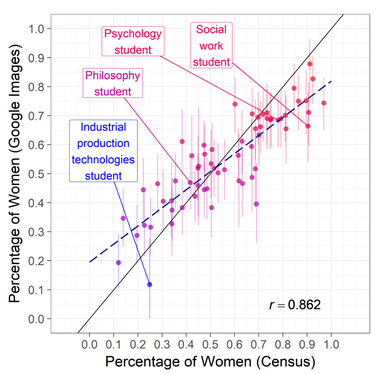


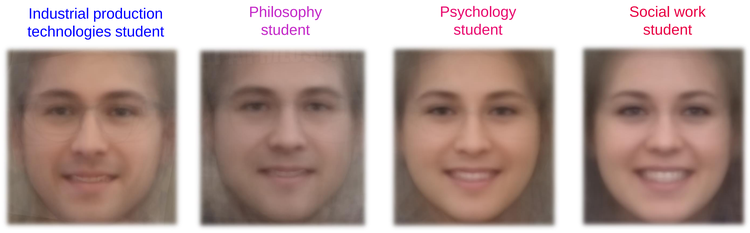




2. College Majors

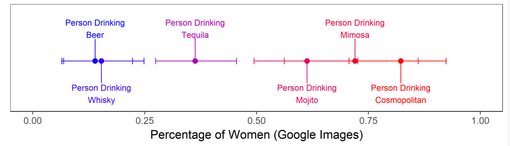






3. Alcoholic Drinks

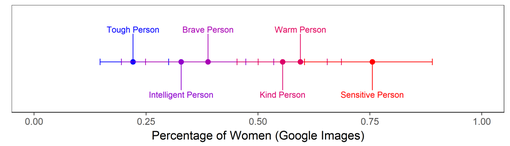
Person Drinking Beer



Person Drinking Cosmopolitan

4. Personality Traits

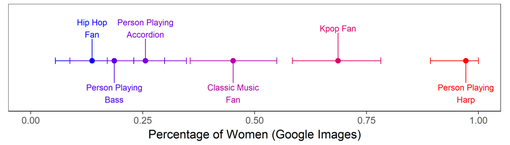
Intelligent Person



Sensitive Person

5. Music Genres and Instruments

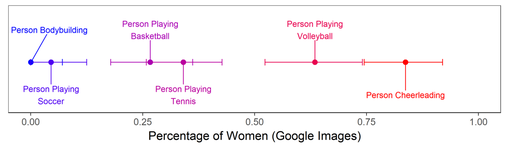
Hip Hop Fan



K-pop Fan

6. Sports

Person Bodybuilding



Person Cheerleading

Discussion

These results indicate that the percentage of women in different professions and college majors corresponds both with census data and participants' beliefs. Similarly, the images in the domains for which gender distributions are unknown reproduced people's commonly held beliefs. Additionally, face averaging further supported these findings.

In summary, our findings suggest that search results from Google Images – a free, vastly rich, and ubiquitous source of information for millions of people – reflect both beliefs about the world and actual statistics. However, these gender distributions may undercut people's desires for equal representation between men and women, in which case a commonly used source of information may perpetuate gender inequalities.

[1] Caliskan et al. (2017). *Science;* [2] Garg et al. (2017). *PNAS;*
[3] Bureau of Labor Statistics. (2018). *CPS;* [4] Casselman. (2014). *FiveThirtyEight*