RISE

Research Centre on **Interactive** Media
Smart Systems and **Emerging** Technologies

Towards Automatic Detection of Social Stereotypes in Algorithmic Output: The Case of Image Search

OPEN UNIVERSITY OF CYPRUS
www.ouc.ac.cy

Introduction

There is growing recognition that algorithmic processes can have consequences in the social world, such as biases in their outputs that result in discrimination against individuals or groups. **Web search algorithms** are a prime example; they are opaque to the user, but hold great power to shape his or her view of the world. Previous research has demonstrated, via *manual detection*, that search perpetuates racial and gender stereotypes, e.g., in image searches surrounding the professions [Kay et al., 2015]. We developed a theoretically grounded, **automated method to test for gender-based stereotypes in image searches** concerning *character traits*.

Theoretical background

--According to social psychologists, our perceptions of others are based on two dimensions [Fiske et al., 2002]

- Agency: whether we perceive someone as being capable of achieving goals.
- Warmth: whether we think someone has pro-social intentions or is a threat.

--Stereotypes are captured by combinations of the two dimensions. Women are expected to be low in agency and high in warmth, while men are seen as being high in agency and lower in warmth.
--The **Trait Adjective Checklist method** [Katz & Braly, 1933] has been used for decades to measure the content and strength of stereotypes. Participants are asked to describe target social groups using a set list of trait adjectives.
--We administer a similar test to the **Bing Image Search API**. Using a list of 68 character traits [Abele et al., 2008], we submit queries, e.g., "intelligent person," "emotional person." We then analyze the gender distribution of the persons depicted in the retrieved images.

Image collection & analysis

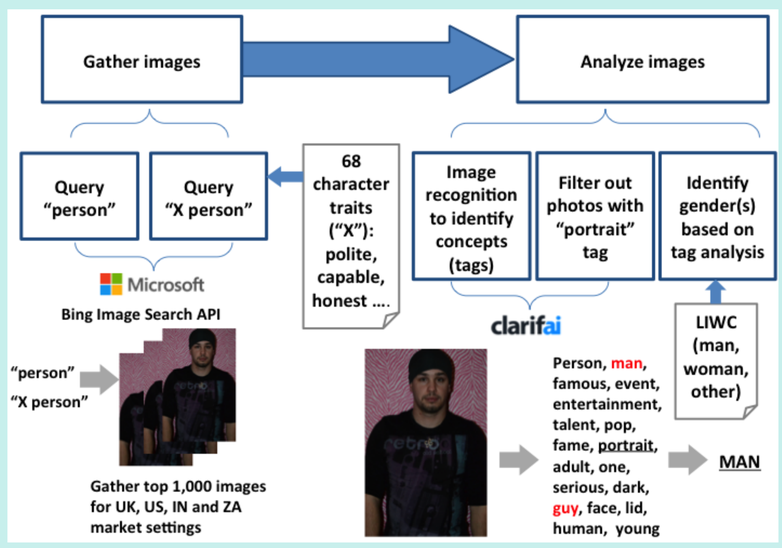

Figure 1: Image collection and analysis - gender detection.

Method and performance

--Figure 1 illustrates the process of collecting and analyzing images. To develop the process, 1.000 images from Bing were retrieved on the query "person."
--Images were manually annotated by three judges, who answered two questions: 1) Is the image a *photo*, *sketch* or *other* image? 2) Does it depict one or more *women/girls*, *men/boys*, *mixed gender group*, *no persons*, or *cannot tell*?
--The mean agreement on the 5-way classification of gender was 0.94 in photos, and 0.91 in sketches. However, as shown in Table 1, the gender of persons depicted in sketches is often ambiguous. Thus, we focus on *detecting gender in photos* only.
--We used Clarifai API, a general purpose image recognition tool. As shown in Figure 1, for an input image, Clarifai outputs 20 concept tags. The tag "portrait" is used to detect photos (recall of .75, precision of .91).
--We use the "female" and "male" dictionaries within the Linguistic Inquiry and Word Count (LIWC) tool to process remaining tags; this results in a three-way classification – photos are inferred as depicting women, men and "others," including mixed-gender groups. Performance is reported in Table 2.
--The method was used to find the gender distribution in the photos retrieved for each of 68 character traits. Figure 2 plots the proportion of photos depicting women, men and others for each trait.
--Bing's gendering of character traits is somewhat consistent with predominant stereotypes. Future work will examine the extent to which this varies as a function of localization and/or personalization of search results.

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

Table 1: Gender distribution in manually annotated images.

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

Table 2: Performance on gender detection on *photos*.

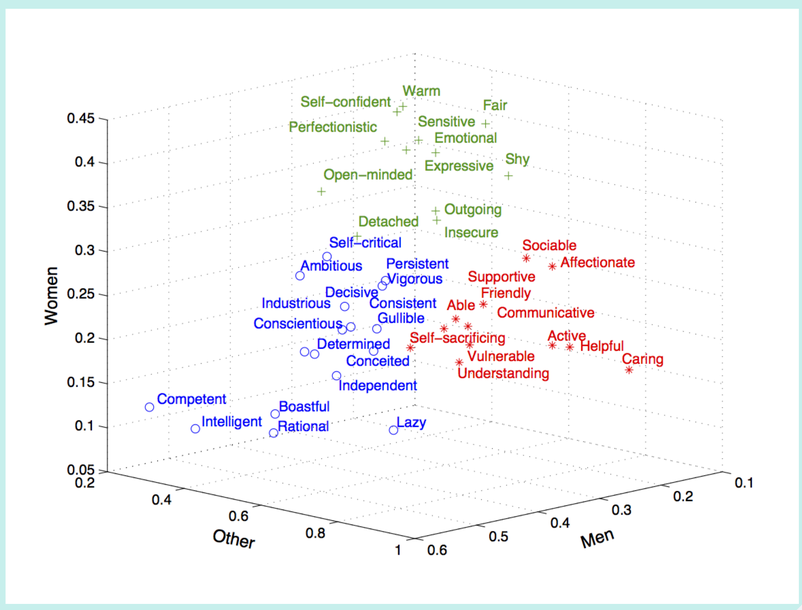

Figure 2: Gendering of 68 character traits.

Relevant publication: **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 (CHI '17). ACM, New York, NY, USA, 6620-6631.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 739578

This project has received funding from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development