Diagnosing Gender Bias in Image Recognition Systems

[accepted for publication in journal Socius on September 3, 2020]

Authors

Carsten Schwemmer[1], Carly Knight[2], Emily D. Bello-Pardo[3], Stan Oklobdzija[4], Martijn Schoonvelde[5] and Jeffrey W. Lockhart[6]

Affiliations

1 Center for Information Technology Policy, Princeton University, c.schwem2er@gmail.com.

2 Department of Sociology, New York University.

3 Department of Government, American University.

4 Research Director, California YIMBY.

5 School of Politics and International Relations, University College Dublin.

6 Department of Sociology, University of Michigan, Ann Arbor.

Abstract

Images recognition systems offer the promise to learn from images at scale without requiring expert knowledge. However, past research suggests that machine learning systems often produce biased output. In this article, we evaluate potential gender biases of commercial image recognition platforms using photographs of U.S. Members of Congress and a large number of Twitter images posted by these politicians. Our crowdsourced validation shows that commercial image recognition systems can produce labels that are correct and biased at the same time as they selectively report a subset of many possible true labels. We find that images of women received three times more annotations related to physical appearance. Moreover, women in images are recognized at substantially lower rates in comparison to men. We discuss how encoded biases like these affect the visibility of women, reinforce harmful gender stereotypes, and limit the validity of the insights we can gather from such data.

Keywords

gender, image recognition, computational social science, bias, stereotypes

INTRODUCTION

Bias in the visual representation of women and men has been endemic throughout the history of media, journalism, and advertising (Becker 1974; Goffman 1979; Ferree and Hall 1990). As Goffman (1979: 11) argued, such "public pictures" are a key symbolic arena in which gendered "social structure of hierarchy or value" is manifested and reproduced. Yet despite their importance, social science research has largely neglected the analysis of images as an arena of social and political valuation. Until recently, the complexity of images rendered large-scale, systemic analysis a near impossibility.

The advent of automated image labeling and recognition systems has increased the importance of images as a form of social data, facilitating their use for commercial purposes (e.g., Greenfield 2018; HG Insights 2020) and, increasingly, for social research (e.g., Xi et al 2019; Gelman et al 2018; Garimella and Eckles 2020; Geboers and Van de Wiele 2020; Ronco and Allen-Robertson 2020; Webb Williams et al. 2020). At the same time, recent research has shown algorithmic classification systems to be a mechanism for the reproduction, and even amplification, of more general social biases (Friedman and Nissenbaum 1996; Noble 2018). Thus far, several recent studies have detailed gender biases affecting supervised image recognition systems. For example, image search algorithms, when asked to return images for occupations, generated results that reproduced gendered stereotypes, exaggerating gender disparities (Kay, Matuszeck and Munson 2015) and featuring women less prominently than men (Lam et al 2018).

While these studies have shown how image recognition systems produce bias in the representation of women and men— that is, how many appear in photos—less research has systematically explored bias in the content of these algorithms' results—that is, how images of

women and men are differently labeled, tagged, and categorized. In this article, we present an analysis of bias in both the identification of people and the content labeling of images of women and men across a set of popular commercial image recognition systems. To the best of our knowledge, this paper is the first to systematically evaluate biases across both these dimensions of person identification and content labeling. We draw upon data from a particularly salient social arena: the visual communications of American politicians. Using two datasets of images from Members of the 115th Congress, we analyze how Google Cloud Vision (GCV)—a widely utilized service in industry and scientific research—categorizes these politicians' images. We replicate our analysis across other popular off-the-shelf-alternatives, including Microsoft Azure Vision and Amazon Rekognition. Across both datasets and all three platforms, we find consistent evidence of two distinct types of algorithmic gender bias. Image search algorithms not only exhibit bias in identification—algorithms "see" men and women and different rates-- but bias in content—assigning high-powered female politicians labels related to lower social status.

Following studies of gender, classification, and status inequalities (Ridgeway and Correll 2004; Ridgeway 2011; 2014) we suggest that image recognition systems reproduce the status inequalities and gender stereotypes at play in the wider social structure. These algorithms not only lead to differences in the representation of men and women but systematically categorize women and men with labels differentiated by status. Empirically, we conclude that the systematic nature of such biases in image recognition classifiers renders these classifiers unsuitable for gender-related analyses. The pervasive and not-always-obvious nature of these biases mean they may also confound analyses that are not gender-focused. Theoretically, our findings identify these algorithms as an important case of what Ridgeway (2011: 40) terms an "amplification process"—

that is, a mechanism through which gender differentials are re-inscribed into novel social arenas and social forms.

GENDER INEQUALITY, CATEGORIZATION, AND ALGORITHMIC BIAS

Gender inequality is characterized by, and reproduced through, the persistence of gendered stereotypes that associate women with lower social status than men (Ridgeway and Correll 2004; Eagley, Wood and Diekman 2000; Ridgeway 2011; 2014; 2016). As Ridgeway (2011: 11) has argued, gender is "at root a status inequality"—one based on cultural beliefs about the differential hierarchical status between men and women. Widely-held and enduring gender beliefs characterize women as less agentic, less worthy, and less competent than men (Conway, Pizzamiglio, and Mount 1996; Spence and Buckner 2000; Lueptow, Garovich-Szabo, and Lueptow 2001; Fiske et al. 2002). While women are typically associated with "communal tasks," men are typically seen as "more competent at the things that 'count most'" and that earn the highest esteem (Ridgeway and Correll 2004). These same stereotypes have been shown to be at play in the visual representation of men and women (Goffman 1976; Ferree and. Hall 1990). For instance, Goffman's 1976 Gender Advertisements demonstrated how advertisements systematically portrayed women in an "unserious," child-like fashion. Ferree and Hall (1990) found that even in sociology textbooks, a corpus supposedly attentive to gender inequalities, images reflected women's marginality in the domains of politics and the economy.

A great deal of social science research has investigated the puzzling endurance of these gender stereotypes over time (Luptow, Garovich, Lueptow 1995; Lueptow, Garovich-Szabo, and Lueptow 2001; England 2010; Cotter, Hermsen and Vanneman 2011) — beliefs that are

continually re-inscribed "in new social forms of social and economic organization as these forms emerge in society" (Ridgeway 2011: 4). A key mechanism for this persistence is the ability of gendered status beliefs to "transfer" to novel social arenas, what Ridgeway terms an "amplification process." This amplification process allows for categorical differences associated with gender to expand in their range of application, so that preexisting gender beliefs are carried into new industries, occupations, or social forms. Status beliefs can even be transferred to "non-status elements" (Tak, Correll, and Soule 2019). For example, gendered stereotypes about men and women can transfer to evaluations of the products that they produce, with women being disadvantaged when they produce stereotypically male-typed goods (Tak, Correll and Soule 2019).

This research has typically focused on how status inequalities are perpetuated through gender beliefs: individuals bring either conscious or subconscious gendered classifications to novel social arenas (Berger et al. 1977; Correll and Ridgeway 2003; Webster and Foschi 1988). The promise of machine learning algorithms has been that they would bypass this aspect of human bias, leading to more accurate or equitable results (Gates, Perry, and Zorn 2002; Kleinberg et al 2018; Cowgill 2018). Nevertheless, a growing body of research has shown that algorithms propagate, and even amplify, existing social structures and biases (Angwin et al., 2016; Kirkpatrick, 2016; Sandvig et al., 2016; Noble, 2018; Benjamin 2019). That is, algorithms are "not cameras onto social realities but engines" (Fourcade and Healy 2017), reproducing pre-existing categorizations found in the social institutions from which the algorithm emerges. For example, natural language processing trained on biased text has been shown to strengthen the gendered associations in language, rather than avoiding them (Bolukbasi et al 2016; Noble 2018; Benjamin 2019).

While more research has been conducted on text than images, prior studies of images have shown similar patterns (Buolamwini and Gebru 2018). Some scholars, many of them computer scientists, have begun to analyze what Ferree (1990: 505) refers to as the "first level of representation" in image bias: estimating the systemic absence of images of women in particular social arenas. For example, in a study of hundreds of thousands of news articles, Jia and colleagues (2015) found that the representation of women varied by topic, with political images featuring primarily men. This bias in representation can then be encoded into biases in algorithms. For example, in a study of occupations, Kay and colleagues (2015) found that search engine algorithms returned images that over-represented men compared to their actual numbers in the population.

To date, less research has investigated how image labeling algorithms categorize—that is, how they classify, label, and annotate images of women and men. As Ferree (1990) suggested and Noble (2018) found, the lower social status of women could result in visual portrayals of women associated with "demeaning or marginalized social positions" (Ferree 1990: 506). The capacity for algorithms to amplify these pre-existing biases is the subject to which we now turn.

From Bias in the World to Bias in the Algorithm

Images are a powerful medium of communication. They are more likely to be remembered than words (Grady et al. 1998; Whitehouse, Maybery and Durkin 2006) and evoke stronger emotions (Brader 2005) and higher levels of social engagement than text (Rogers 2014). Despite the enormous social scientific potential of images as data, their analysis remains computationally demanding. Algorithms to analyze images often require a high level of technical training and knowledge to design and use, as well as large amounts of training data and data-labels. Gathering

tens of thousands or more images, all with labels describing their content, remains both costly and time consuming (Chen et al. 2015).

Commercial image labeling services, available to the public from Google, Amazon, Microsoft and other companies since 2016-2017, provide an alternative to this onerous process: reducing the cost of labeling images and identifying their content at scale and offering the potential to make image analysis readily available to users not trained in designing neural networks. These platforms allow users to quickly and easily retrieve labels for any image, as shown in Figure 1. A recent study shows just how drastic the difference in effort between human coders and algorithms like Google Cloud Vision is: "the API codified 1,818 images in less than 5 min, whereas the human coder spent nearly 35 hours to complete the same task" (Bosch, Revilla, and Paura 2019).

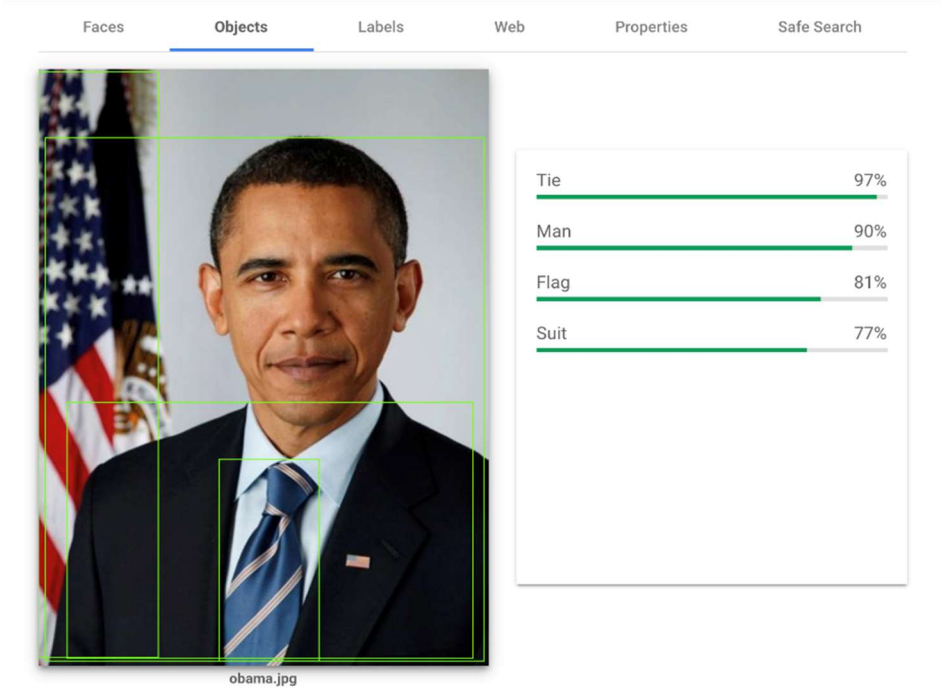

Figure 1. Example of the information that Google's Cloud Vision platform can return when asked to label a portrait of former U.S. President Barack H. Obama.

One widely known dimension of systems like GCV that rely on machine learning is that they seek out and then reproduce patterns in the data on which they are trained. Input data are typically "found data" from the "real world," containing the biases and cultural associations of human societies, which then get reproduced as "objective" and "scientific" decisions from algorithms (Benjamin 2019). For example, ImageNet is a database widely used to train image labeling algorithms that maps the categories from Princeton's WordNet to over 14 million images scraped from the internet (Crawford and Paglen 2019). WordNet is a taxonomy of English terms dating to the 1980s, based on pre-1972 Library of Congress taxonomies that contains numerous racist, ableist, and misogynistic terms (Crawford and Paglen 2019). When ImageNet's designers and human coders linked these terms to pictures of people from the internet, they encoded those biases into the database. As Crawford and Paglen (2019) showed, this profoundly shaped algorithms that were trained using the database. After their work, ImageNet removed many of the most offensive labels (Ruiz 2019).

Input data is not the only social influence on algorithmic systems. Computer engineers' design decisions and tweaking of automated systems also encode biases (Seaver 2018). For example, engineers working on music playlist algorithms use not only users' behavior to code their algorithms but also personally listen to the playlists it generates, tweaking the way the algorithm used its input data until the engineers thought the output sounded good (Seaver 2018). As Seaver states, arbitrary preferences and biases outside the code therefore become a part of the algorithm:

> "The essence of a contemporary algorithmic system [is] a steady accumulation of feedback loops, little circuits of interpretation and decision knit together into a vast textile. Every stitch is held together by a moment of human response, a potential rejection shaped by something outside the code, whether it is the arbitrariness of personal preference, the torque of structural bias, or the social force of a formal evaluation framework (2018: 377)."

Notably, the algorithmic systems trained on these input data are increasingly "black boxes." A system is a black box either if its technical design is sufficiently complex that human users cannot interpret the meaning of the inner workings, or if the details of the system's design and construction are hidden from users, for example as corporate trade secrets (Rudin, 2019). This second kind of black box describes GCV and nearly every commercially available "algorithm" or scoring system. Only some Google employees know which data sets and design decisions went into building and tuning GCV. Therefore, while researchers can audit the *results* of algorithms, they generally cannot recover the true process or logic of the black box's decisions and attempts to reverse engineer the decision process "are misleading and often wrong" (Rudin 2019: 211).

Thus far, scholars working on images have taken some initial steps to avoid the bias potentially introduced by these algorithms. For instance, in a study on social media images of legislators, Xi and colleagues (2019) removed all women and members of racial and ethnic minority groups from their data in order to sidestep gender and racial biases. While such an approach may be reasonable for specific research questions, it should be a last resort: systematically excluding large swaths of the population not only can lead to non-generalizable inferences, it can also bias social scientific research away from pivotal research questions on inequities in social, political, and economic visual communication (Rossiter 1993). We suggest that prior to resorting to such data limitations, we should develop a better understanding of the systematic nature of such biases. In what follows, we draw upon existing literature to examine gender algorithmic bias across two dimensions: bias in identification and bias in content.

Two Dimensions of Image Bias

Bias in identification is an analog to what Ferree term**s** the "first level of representation": at a very basic level, does the algorithm *see* people with equal accuracy regardless of their gender? For the most part, this has been the primary focus of the "algorithmic bias" literature, which has defined algorithmic injustice and discrimination as situations where errors disproportionately affect particular social groups (Noble 2018).

Bias in content, by contrast, is possible when algorithms output only a subset of possible labels, even if the output is correct. In this case, an algorithm might systematically return different subsets of correct labels for different groups of people. We formalize this as "conditional demographic parity" (Corbett-Davies et al. 2017). Conditional on image content, an algorithm is considered biased if it returns labels at different rates for different demographic groups. For instance, if men and women in a sample wear suits at equal rates, then an unbiased algorithm would return the label "suit" equally often for each gender. Why might the presence or absence of women in a photo affect the identification of such seemingly non-gendered classifications like clothing items? Algorithms learn by observing associations in the data they are trained on (i.e. data the models are fitted to). If we fit an algorithm to a data set where all men had suits, and no women did, it might well learn that the probability of "suit" being the right answer, given that it sees a woman or features that were associated with women in the data used to train the algorithm like long hair, is extremely low. When later presented with images of women in suits, then, it would be unlikely to label them "suit," even though that is a correct label.

Input biases do not need to be that extreme to have these effects, however. Research on word embeddings has found that algorithms can pick up far more subtle associations (Kozlowski et al. 2019). For example, one team found that word2vec trained on Google News articles produces

gendered analogies like "man is to computer programmer as woman is to homemaker." This is because gender-specific words (like "sister" or "mother") may be statistically associated with gender-neutral words (like "homemaker") in text, and thus algorithms that attempt to identify meaning through observed associations amplify these biases (Bolukbasi et al. 2016). Similarly, algorithms trained on real world images may convert associations between gender-specific labels and gender-neutral labels into biased results for image content.

DATA

To identify bias in identification and content in image recognition systems, we use two datasets containing images associated with Members of the 115th United States Congress: a dataset of official headshots and a set of images tweeted by these members. We have several reasons for focusing our analysis on political images. First, politicians' image use is substantively important. The political realm has consistently revealed gender bias in the representation of women in images (Jia et al. 2015). It is important to know whether and how human bias in the production and use of images plays out in algorithmic labeling of images. To date, politics has been an important domain of social science research on images (e.g. Anastasopoulos et al. 2016; Casas and Webb Williams 2019; Xi et al 2019; Webb Willams et al. 2020).

Second, our datasets offered a unique opportunity to study the bias in black-boxed image classification algorithms. We compiled two matched data sets: (1) a control dataset consisting of uniform portraits of the Members of Congress themselves and (2) a found dataset of images that these politicians tweeted. The control dataset limits the variation in image content and style, making it easier to detect biases in algorithmic performance, while remaining a real-world image data set. It includes social markers of gender, age, race, and politics such as clothing, hair, jewelry,

and flags that are essential to sociological understandings of identity and appearance, but which are typically cropped or abstracted away in the controlled photographs of laboratory studies. The found dataset is composed of images shared by the politicians' official Twitter accounts, which are highly variable in content, style, and purpose, but which still share a general context. These characteristics mirror those of many digital sociology and archival research projects, allowing us to evaluate algorithmic bias in a setting relevant to other researchers. Both data sets are linked to the same set of politicians, and thus the same demographics, enabling us to compare findings.

Control Dataset. We acquired the control dataset by extracting official portraits of Members of Congress from Wikipedia. These photos are produced by the United States Government Publishing Office for the official Congressional Pictorial Directory which contains photos and biographical details for all MCs during a given session. The vast majority of these images are taken in front of a neutral monochrome background. In many photos, an American flag is positioned to the MC's right and in a subset of those photos, the flag of the MC's home state is also displayed to that person's left. Many photos are taken either somewhere in the U.S. Capitol Building or an MC's office. In every photo, the vast majority of the frame is occupied by the MC. Similarly, in all photos, MC's are clothed in civilian business attire and looking at the camera. Members of Congress all have the same occupation, nationality, and motivation for taking their portrait. These photographs are as homogeneous as any real-world set of images might be, without artificially removing socially- meaningful aspects of age, gender, race, and ethnicity, such as hair and clothing, which are often removed in laboratory facial recognition data sets. All images fall under the public domain and are included in our replication material. We merged these photos with information about the MC's from government websites as well as a public Github repository (United States Project 2020).

Found Dataset. Our found dataset is composed of images posted on Twitter by MCs between January 2017 (the start of the 115th Congress) and June 2018 (n = 198,170). We obtained the set of images by using the Twitter Application Programming Interface (API) to download each MC's timeline, limited to their most recent 3,200 tweets due to data restrictions from the API. We then downloaded all of the images that these tweets contained.

From these sets of images, we selected a weighted sample in order to validate GCV's labels with human's labeling (N=9,250). An image's weight for sampling is calculated using both the labels from Google Vision and by the characteristics of the MC's posting the image. Image weights are inversely proportional to how rare their features are, such that images with uncommon labels and coming from MCs from underrepresented groups are more likely to be sampled. More details on our sampling strategy are available in our online appendix. On average, GCV returned 5.12 labels per image, and we only selected labels that GCV assigned ≥ 0.75 confidence to (confidence scores from GCV vary between 0.5 and 1.0). In that sense, our validation sample can be regarded as conservative; we only evaluate labels GCV considers as highly likely to be applicable to the specific image.

METHODS

Our main analysis is conducted on Google's Cloud Vision (GCV). As discussed above, GVC is widely used in industry, and unlike its primary competitors, Amazon Rekognition and Microsoft Azure Vision, GCV shares its underlying technology with the world's largest internet image search platform (Google Image Search) and other ubiquitous services like Google Photos (integrated with every Android phone). We also provide brief analysis of both other platforms showing that our findings generalize outside of GCV.

Validation. To validate the image labels produced by the algorithm, we hired workers through Amazon's Mechanical Turk (MTurk). This service has become popular with researchers in several disciplines over the past decade and allows for hiring a readily accessible and diverse population of research assistants. While "MTurkers" have often been a population sampled for survey research (Huff and Tingley 2015), these workers have also been employed to assist in the research process itself (Shank 2016), as was the case for our project. The use of temporary and anonymous workers who lack the labor protections of traditional research assistants employed through a higher education institution has been discussed extensively by other scholars (see Williamson 2016; Pittman and Sheehan 2016). Aspiring to maintain ethical research practices, we paid MTurkers working on our project a "living wage" of $15/hour—more than twice the U.S. federal minimum wage at the time of writing.

We presented each worker with 30 images and a set of five potential labels for each. Some labels were assigned by GCV for corresponding images (positive labels); others were chosen at random from the set of GCV labels assigned to other images but not to the one at hand (negative labels). Each image was coded by at least three people.

Workers were presented with an image and two questions. The first question presented all labels in random order and asked workers to select all labels that apply to the image they were seeing. The second question asked workers to indicate if they saw any men, women, children, or none in the image. Each person validated the labels of 30 images, and multiple people saw each combination of labels and images. Overall, respondents had an agreement rate of 0.77 with one another.

In order to identify bias in identification, we evaluate whether GCV recognizes men and women in images. With our control data, we have ground truth about the presence of and gender

for MCs depicted. With our found data we do not know the true gender of people in images. Instead, we compare whether GCV recognizes men, women, both, or neither in an image to whether human coders do. Human coders and GCV both rely on the same visual gender cues, so our research design measures whether those cues influence the algorithm's person identification.

Bias in content requires a slightly different approach. There are many things that could be labeled in any image ("an image is worth a thousand words"), but image labeling systems typically return only a handful of labels (an average of 5.3 per image in our data). Even if labels a system returns are correct, it is possible to have bias in which subset of possible correct labels gets returned for a given image. Thus, we measure bias in content as conditional demographic disparity: conditional on actual image contents, we examine whether some labels are disproportionately applied to images containing one demographic group or another.

To measure bias, we rely on two procedures. First, we use χ^2 test statistics with Yates correction on labels to identify which labels are identified relatively most often in portraits of and images tweeted by women compared to men (see our Online Appendix). Second, we use negative binomial regressions to obtain the expected counts of GCV labels in each of five coded categories for the MCs. A negative binomial distribution allows us to model counts while correcting for overdispersion.

Finally, we include several controls. Because recent research suggests GCV results may depend upon race and skin tone (Noble 2018), we control for race (coded as White or Non-White). Women are unequally distributed across parties, and to ensure results are not party-dependent, our models also control for party membership (Democrat or Republican). Finally, as studies have shown that the performance of image recognition algorithms may depend upon the age of

individuals in the images (Grother and Ngan 2014; Michalski 2017), we control for age (see Online Appendix). Results are robust to the inclusion or exclusion of these controls.

DETECTING GENDER BIAS IN GOOGLE CLOUD VISION

Bias differs across image classification systems and changes over time. Because of this, researchers using these algorithms will need to do their own evaluations, specific to the tool they are using, the time they are using it, and even the location they are accessing it from. We propose that such evaluations should measure several components. First, as a baseline, researchers should verify the correctness of the labels provided; many applied papers already evaluate this dimension (e.g., Bosch et al. 2019) but because accuracy will be context dependent, such verification is an important first step every time one uses an algorithm. Second, we suggest that researchers identify two forms of algorithmic bias: biases in identification, which is the focus of much "algorithmic bias" literature (e.g., Kleinberg et al. 2018); and biases in content. In what follows, we discuss each of these components drawing on US Members of Congress' use of images on Twitter as a case study, but the procedure we propose is generalizable to other substantive domains.

Evaluating Google Cloud Vision

The first, most general dimension for evaluating any algorithm is determining the correctness of its results. There are many different measures for evaluating labeling or classification algorithms (Nelson et al. 2018). In general, commercial labeling systems present users with only predicted positive labels (e.g. "there are cats in this photograph") and not predicted negative labels (e.g. "there are no children"). This can make calculating many measures of correctness difficult. Additionally, calculating measures of correctness requires "ground truth"

data about what is "correct." But users typically turn to labeling algorithms precisely because they do not already have ground truth information about their images.

We address both challenges using our sample of 9,250 human-coded images. Overall, we find that human crowd workers have high agreement with the labels the GCV algorithm generated, as shown in Figure 2. When presented with an image and a set of potential labels, humans typically select the positive GCV labels, but not the negative labels. Moreover, the proportion of humans who select a label is strongly correlated with the confidence score returned by GCV. That is, GCV's confidence score is a good measure of whether a human would agree that the label applied to a given image. In this sense, GCV is a high-precision image labeling system: when GCV says a label applies to an image, it is generally correct.

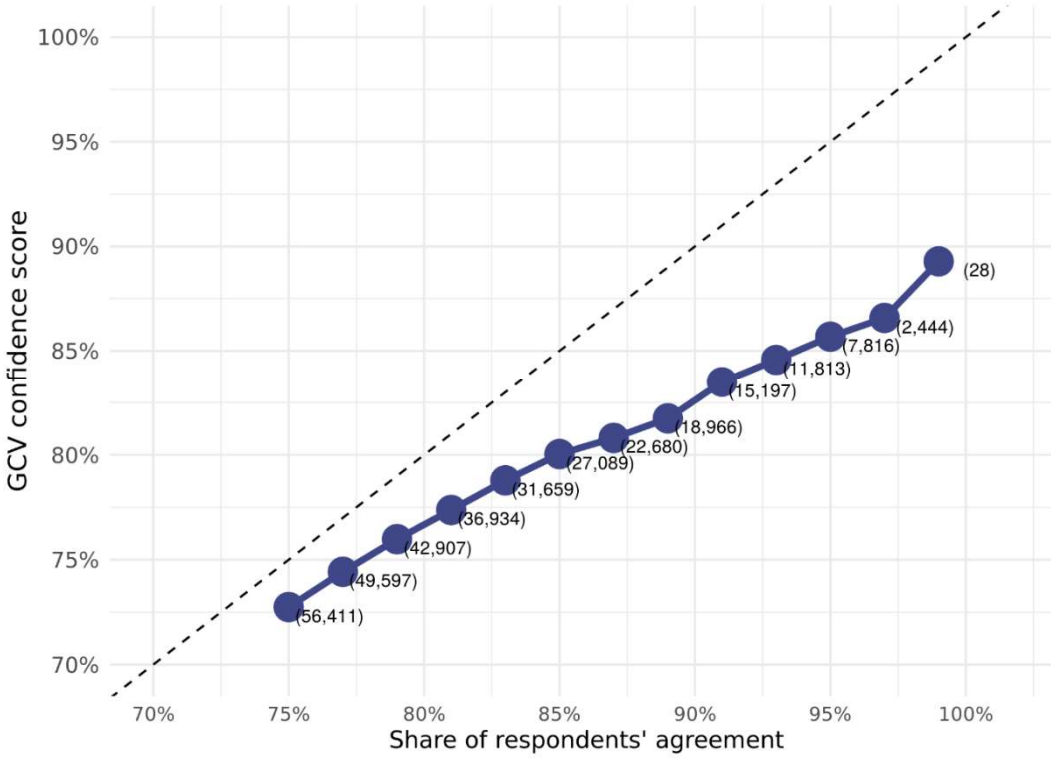

Figure 2. Relationship between Google Cloud Vision confidence and human agreement. Numbers in parentheses denote observations for corresponding confidence score thresholds.

Bias in Identification

Overall accuracy of an algorithm like GCV is not the only important measure, however. As Nelson and colleagues (2018) show, sometimes the measures of correctness for individual categories/labels are more important for sociological analysis and can lead to further insights about the data. We test this with gender. While observer-ascribed gender is a poor measure of gender identity (Hamidi, Scheuerman and Branham 2018; Lagos 2019), it can be a good measure of the stereotypically-gendered appearances that may influence GCV.

We use the object recognition module of GCV which, at the time of data collection, detects people and differentiates between men or women. We conduct this validation using all images from the controlled dataset (results shown in Figure 3) and all 9,250 images from Twitter that human workers coded (results shown in Figure 4). As the right panels of Figures 3 and 4 demonstrate, GCV has low false-positive rates for detecting people that our human coders did not identify in the images, regardless of gender. The false-positive rate is low for both women (near 0% in the Wikipedia image data and around 1% in the Twitter image data) and men (1.8% in Wikipedia images, and 2.3% in the Twitter data). In short, GCV rarely detects people in images where humans do not.

However, the algorithm's false negative rates vary substantially by ascribed gender. In our control dataset of Congressional professional portraits, women in Congress are only recognized in 75.5% of images of women MCs in comparison to 85.8% for men in Congress, a difference of nearly 10 percentage points (see the left panels of Figure 3). Thus, in high quality photos in which only one individual is presented, women are still "seen" by the algorithm significantly less than men.

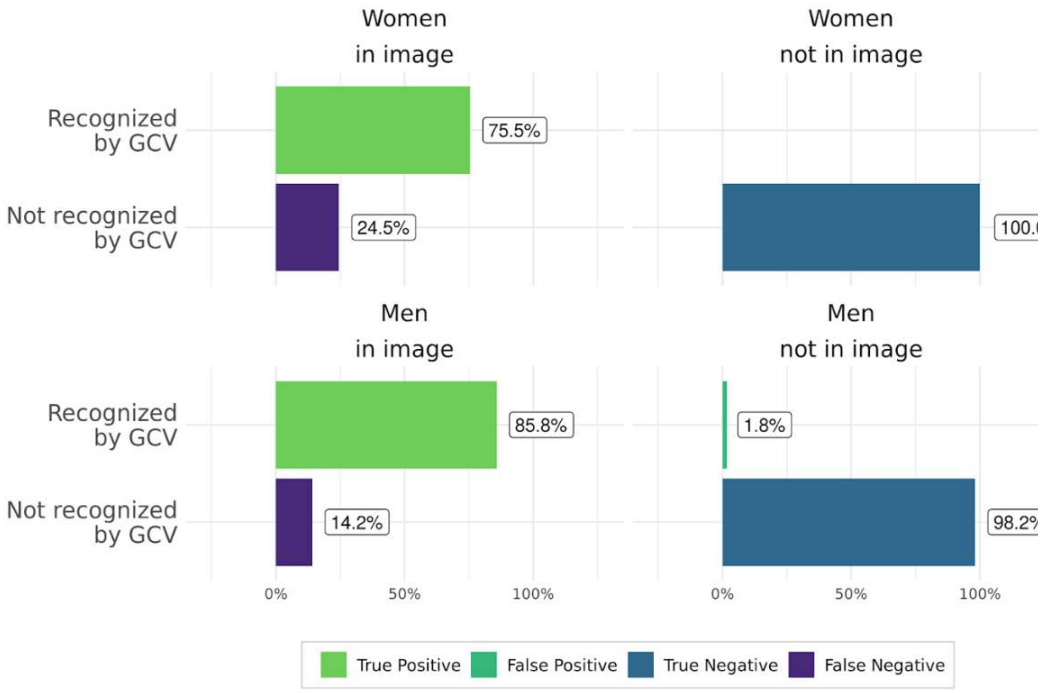

Figure 3. Accuracy of person-detection of Google Cloud Vision. Percentages shown were determined by comparing gender of MCs depicted in uniform data (professional photographs) to annotations from the object recognition software.

This difference was even more striking in our primary dataset of Twitter photos. Here, GCV identified 45.3% of the men that our human coders saw in the pictures, but only 25.8% of the women—a striking 20 percentage point gap (see the left panels of Figure 4). As with the label annotation results, GCV object-labels for people are high-precision: if GCV detects a person, it is very likely humans will agree that there is a person. However, these results indicate that GCV has poor recall: if GCV does not tag something, it may still nevertheless be in the image (ergo, the high false negative rates in recognizing individuals). High precision with low recall is likely an unavoidable feature of labeling images: for any given image, the set of possible correct labels that the algorithm could return is theoretically enormous. Our findings show, however, that there is

substantial gender bias in errors of omission: false negative rates are substantially higher for women than men.

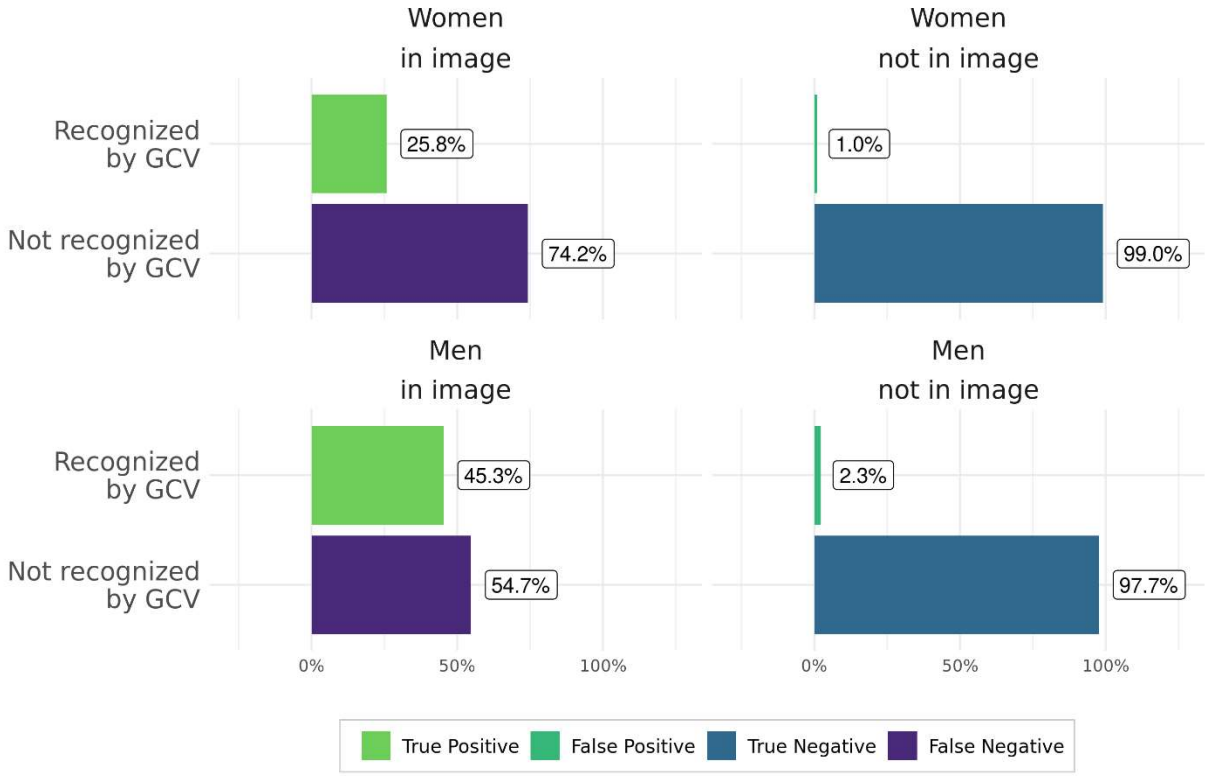

Figure 4. Accuracy of person-detection of Google Cloud Vision. Percentages shown were determined by comparing human agreement about the presence of men or women in Twitter images to annotations from the object recognition software.

Biases in Content

The second component of evaluating GCV labels concerns bias in content. Our finding that positive labels are recognized as correct by humans does not rule out bias in their distribution. Positive labels could be both correct and biased, in the sense that they might not always meet conditional demographic parity.

In order to examine this possibility, we used GCV labels from our uniform data set of Members' of Congress professional portraits. If GCV returns gender-biased labels on this set of images, those biases could affect any inferences we draw from the algorithm with other datasets,

including our analysis of whether Members of Congress engage in gendered patterns of communication on Twitter. Example images and labels from this set can be seen in Figure 5. Here, Google Cloud Vision labeled Congresswoman Lucille Roybal-Allard as a "smiling" "television presenter" with "black hair", whereas Senator Steve Daines was labeled as an "official", "businessperson", and "spokesperson."

Figure 5. Two images of U.S. Members of Congress with their corresponding labels as assigned by Google Cloud Vision. On the left, Steve Daines, Republican Senator for Montana. On the right, Lucille Roybal-Allard, Democratic Representative for California's 40th congressional district. Percentages next to labels denote confidence scores of Google Cloud Vision.

We then use χ^2 tests to identify the key labels by gender for this uniform dataset (see our online appendix for additional information). Figure 6 shows the top 25 key labels for both men and women, sorted by absolute frequencies. Some labels, for instance "long hair" for women, are a clear result of the underlying data we chose: there are no Congressmen with long hair in the data set, and no Congresswomen who wore neckties, so it is unsurprising that some of these labels have

strong gendered associations. Note, however, that "bald" or "short hair" do not appear among the labels that GCV returned, indicating a bias in which hairstyles the algorithm mentioned. The seemingly neutral label "hairstyle" is given to more than half of women but only a minute percentage of men. Similar patterns exist for labels like "black hair" and "brown hair." By the authors' manual count, 2% of women's portraits have no visible hair (due to hats), 3% of men's portraits have no visible hair (completely bald heads), and a further 7% of men's portraits have partial hair (hair visible on the sides but not the top of the head). Conservatively, then, women are 1.1 times as likely as men in this data to have visible hair, nowhere near the disparity in labels returned by GCV. Thus, we conclude that, conditional on hair being in the image, GCV was much more likely to comment on it if the hair belonged to a woman.

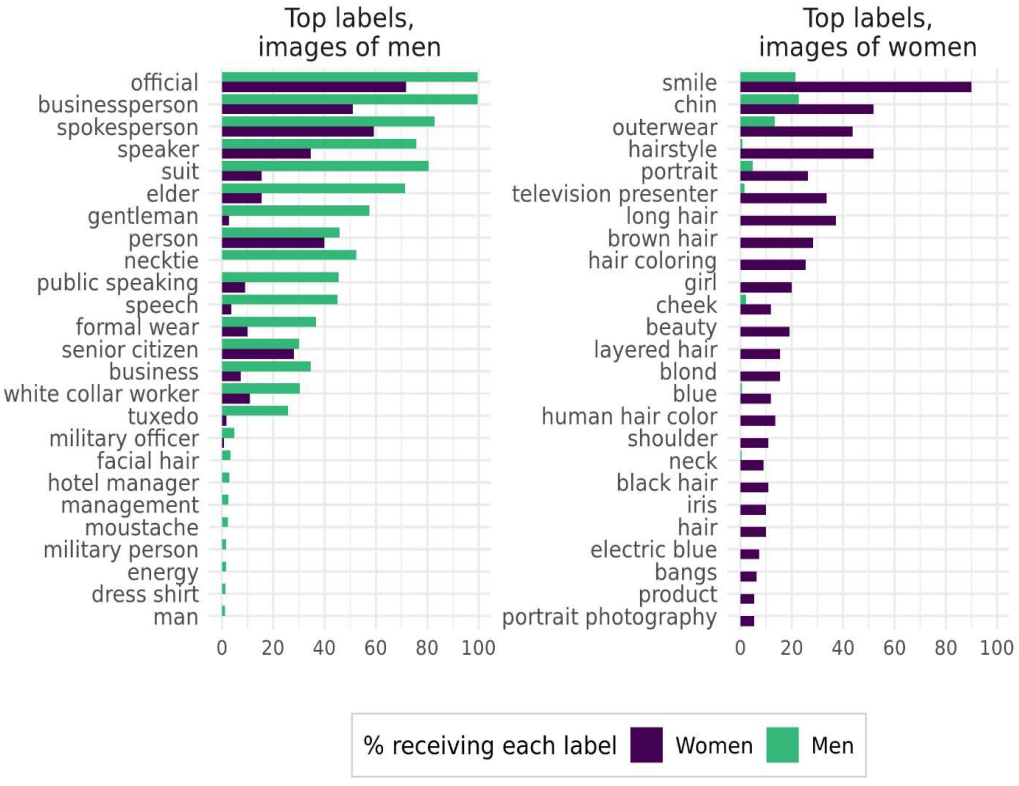

Figure 6. Google Cloud Vision labels applied to uniform dataset (professional photos). The 25 most gendered labels for men and women were identified with χ^2 tests ($p \leq .01$). Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain label by gender.

Labels like "girl" and "gentleman" encode gender directly, so their correspondence with Members of Congress's gender is unsurprising. However, labeling adult women "girls" while men are labeled with more prestigious and age-appropriate titles such as "gentlemen" is an old, sexist trope (Durepos et al. 2017) that resurfaces in image recognition algorithms.

Furthermore, we see evidence that confirms gender and occupational bias. That is, although all individuals in the dataset have the same occupation (Member of Congress), GCV labels them with a variety of occupations. Notably, the only occupation that GCV labels women with more often than men is "television presenter," while men get labeled with more authoritative variants such as "white collar worker", "spokesperson", and "military officer." That is, although these labels are ostensibly gender-neutral, their highly gendered cultural histories emerge clearly in GCV's differential application of the labels. For instance, Perryman and Theiss (2013) show that the age-diminutive "weather girl" stereotype has developed since the 1950s, when television stations began to hire non-expert women as presenters to attract viewers through theatrics and sex appeal. Today, GCV labels women as "television presenter" instead of "weather girl," but the historical gender bias remains evident.

Overall, appearance labels such as "beauty" and "hairstyle" are disproportionately applied to women. Labels most biased toward men revolve around professional and class status like "gentleman" and "white collar worker." None of these individual labels is necessarily wrong. Many men in Congress are businesspeople, and many women have brown hair. But the reverse is true as well: women are in business and men have brown hair. From the set of all possible correct

labels, GCV systematically selects appearance labels more often for women and high-status occupation labels more for men. Naive analysis using these labels may erroneously conclude that images with men or women in them are more focused on, respectively, business or fashion—even if they are all professional portraits of people with the same occupation.

We conducted further analysis to quantify the different types of labels assigned by GCV dependent on gender, race, and party of MCs by manually coding all GCV labels for the photographs of MCs into the following categories: "occupation," "physical traits & body," "clothing & apparel," "color & adjectives," "other". Three authors of this article coded the labels independently, with an intercoder reliability score of 0.88 (see our Online Appendix). For each of these labels we computed regressions to estimate the effects of gender on label counts for the MC photographs. We opted for negative binomial regressions as dispersion tests for our count-based variables suggested partial overdispersion. We control for race, age and political party of MCs. Figure 7 shows predictions by gender while holding party, race and age at observed values.

Figure 7. Predicted labels counts for images of men and women. Results are based on the Wikipedia photographs of U.S. Members of Congress and negative binomial regressions, controlling for party and ethnicity. Circles describe point estimates, bars describe 95% confidence intervals.

Images of women receive about three times more labels categorized as "physical traits & body" (5.3 for women, 1.8 for men). Images of men receive about 1.5 times more labels categorized as "occupation" (3 for women, 4.7 for men). Images of men also receive more labels related to clothing/apparel than women. We found no substantial differences in labels related to color/adjective or other types of traits.

These results provide further evidence that images of women contain more labels related to physical traits in comparison to images of men. At the same time, labels related to occupation, and to a lesser extent clothing & apparel, are more often included in images of men. Results of the same analysis for ethnicity as well as for political party do not suggest substantial effects (see our

Online Appendix). In short, our results indicate that GCV suffers from substantial biases related to gender.

To examine how these biases in uniform data manifest in "real world" data, we turn now to our "found" dataset of Members' of Congress twitter images. Again, we use χ^2 tests to identify the labels most strongly associated with images tweeted by men vs. women MCs. Figure 8 shows the top 25 key labels for both men and women, sorted by absolute frequencies.

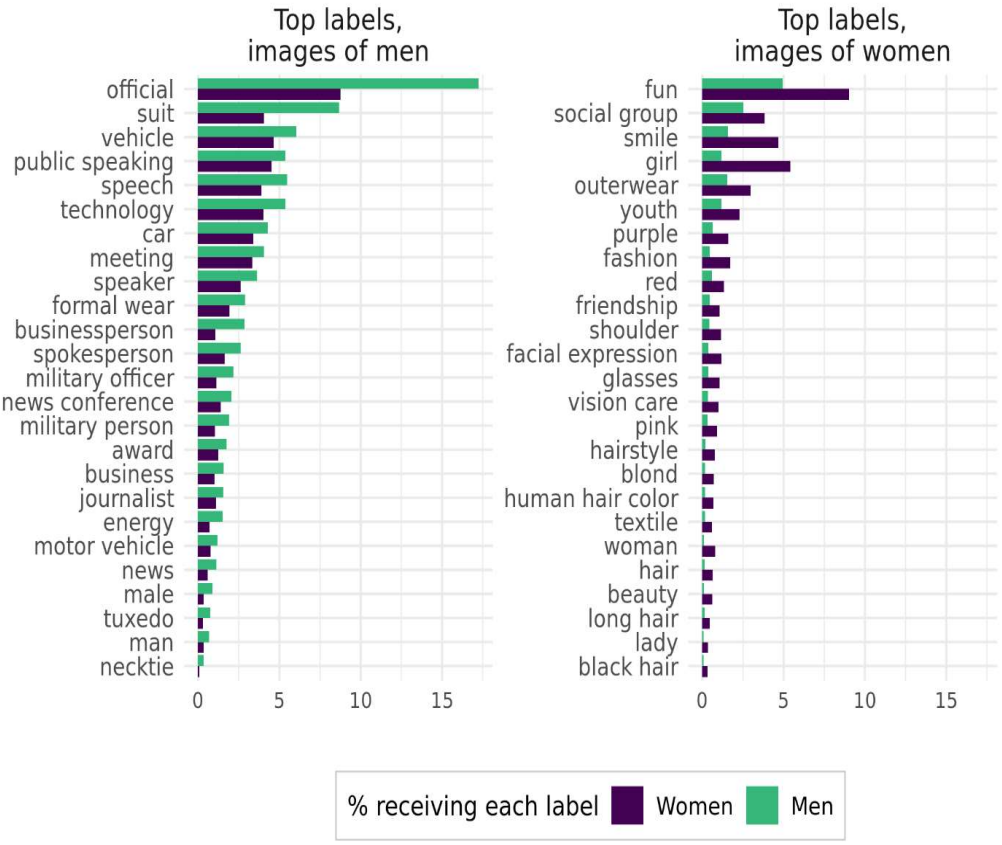

Figure 8. Google Cloud Vision labels applied to found dataset (Twitter images). The 25 most gendered labels for men and women were identified with χ^2 tests ($p \leq .01$). Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain label by gender.

Results indicate a sharp divide in content of images tweeted by men and women, such that female Members of Congress appear to be much more likely to tweet pictures of women and girls,

fashion, and other appearance-focused themes (about 5% of all images tweeted by women received the label "girl," while only 1.5% for images by men received that label). Meanwhile, men in Congress appear much more likely to tweet images of officials, vehicles, public speaking/speeches, technology, military personnel, and business. These themes conform to common gender stereotypes, and a reasonable but naive interpretation of these results might have been that MCs' gender substantially influences the content of the images they share on Twitter. The results broken out by MC's party show similarly gendered distinctions (see Online Appendix).

However, our evaluation procedure highlighted that many of those specific labels are applied with substantial gender bias, which confounds these observed differences. Indeed, when considering that women were much more likely to be given labels associated with physical traits or the body or were much more likely to be labeled as "girls," many of the most "gendered" findings about images tweeted by MCs are revealed to be artifacts of algorithmic bias. The label "girl," for instance, does not necessarily indicate the presence of a child, as we identified the biased application of the label "girl" to our control dataset of images of adult women. Thus, rather than women tweeting more images of girls than men in congress, all MCs might simply be tweeting images of themselves and getting labeled differently by GCV.

Our analysis reveals that GCV's biases severely limit the kind of inferences that scholars interested in gendered political communication could accurately draw from visual evidence if they were to use this black-box algorithm. Indeed, among the top labels associated with "gendered" images tweeted by MCs, it is clear that very few point towards reliable, unbiased differences. We therefore conclude that labels produced by GCV are too biased to yield meaningful insights into gender differences in visual political communication patterns.

Detecting Bias in Other Image Recognition Tools

While our results so far focused on examining gender biases of one particular system, Google Cloud Vision, we also replicated our analysis of our uniform dataset of professional photos on two other popular image recognition tools: Amazon Rekognition and Microsoft Azure Computer Vision. We found that labels assigned by these tools produce gender biases similar to GCV (see Figure 9 and Figure 10 below). For example, Amazon Rekognition assigns the prestigious occupation labels "attorney" and "executive" for photographs of men. Photographs of women are labeled with "teen", "girl", and "kid", although the lowest age for both men and women in our dataset is 34 years. In addition, images of women are also labeled with "home decor" even when they are from the uniform portrait dataset. Unlike GCV and Amazon Rekognition, labels from Microsoft Azure Computer Vision do not seem to be of high precision in general. The system produces biased labels such as "girl", "cake," and "kitchen" for portraits of adult women, where no kitchens or food are present. This demonstrates the need for users to evaluate the specific biases of the system they are using at the time they are doing so.

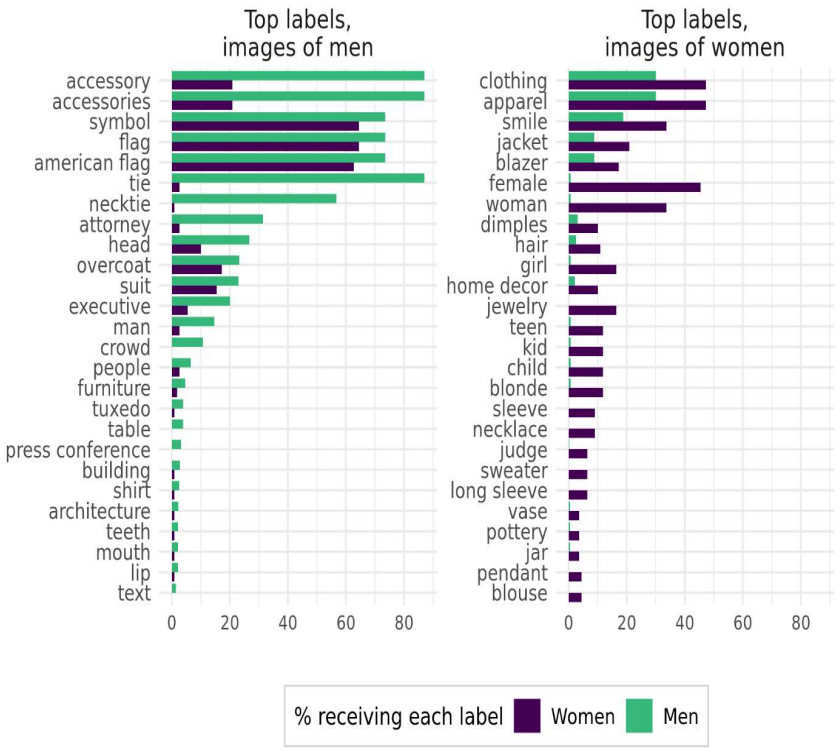

Figure 9. Amazon Rekognition labels applied to professional photographs of Members of Congress. The 25 most gendered labels for men and women were identified with χ^2 tests ($p \leq .01$). Bars denote the percentage of images for a certain label by gender.

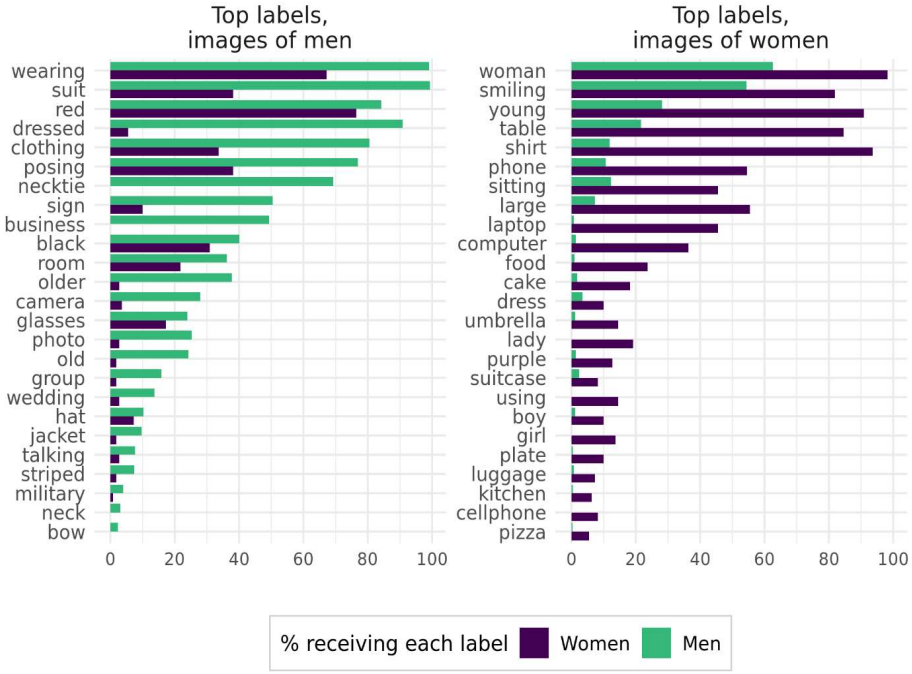

Figure 10. **Microsoft Azure Computer Vision labels** applied to professional photographs of Members of Congress. The 25 most gendered labels for men and women were identified with χ^2 tests ($p \leq .01$). Bars denote the percentage of images for a certain label by gender.

DISCUSSION

In this article, we have identified systemic and pervasive bias in how images including men and women are processed, such that image recognition systems mimic and even amplify real world bias. Specifically, we have shown how bias in identification and bias in content skew the results for even uniform political images, labeling photos of women according to their appearance and photos of men according to their occupation. In other words, image labeling algorithms "see" American Congresswomen through the classic gendered stereotypes that have historically beset the visual representation of women, if they see women at all (Goffman 1974; Ferree and Hall 1990). For any project seeking to draw conclusions from labels that image recognition systems apply with a gender bias, gender may further operate as a confounding variable.

While prior work has sought to either use algorithms (Anastasopoulos et al 2016; Casas and Williams 2018; Xi et al 2019) or identify biases in them (Buolamwini and Gebru 2018; Crawford and Paglen 2019; Eubanks 2018), we argue that it is critical for scholars to do both at the same time. Further, we demonstrated that this is different from simply evaluating the correctness of an algorithm's output, as many applied studies already do. An algorithm like GCV might be both correct and biased at the same time if it selectively reports a subset of many possible true labels. There is an active field of research focused on constructing algorithms to avoid specific biases (e.g., Kleinberg et al 2018). But unless algorithms are consciously constructed and tested

for that specific purpose, biases are likely to taint applications that rely on their output in unforeseen ways.

Though we have addressed algorithm's classification of men and women here, it is important to note that a smaller body of work has begun to examine the systematic exclusion of trans and nonbinary people in algorithmic image recognition systems, which relies on conceptions of sex/gender as binary, immutable, and visually legible (Keyes 2018). That is, such algorithms assume a person or computer can look at someone and know that they are either a man or woman from visual cues such as hair style. To be sure, perception by others is a critical dimension of gender and a part of the interactional process of "doing gender" (West and Zimmerman 1987). But because gender is an accomplishment, rather than a pre-social fact, observer perception and other dimensions of gender such as individual identity may differ in consequential ways (Lagos 2019). The genders we measure in this paper are mostly binary and observer-ascribed, either by algorithms or by humans tasked with validating the algorithms. Here, we demonstrate gender biases and stereotypes even within the constrained, binary terms that the algorithms operate in. This complements work on who can be represented in these algorithms by critically evaluating how those who can be represented by a system's logic are represented by it in practice.

Our findings are necessarily time- and context-dependent. New training data and model changes will alter these results and may alleviate some of the biases **we** identified or generate new, unmeasured ones. Nevertheless, research using image labeling algorithms must be attentive to such biases when drawing conclusions about image content. Our particular results are also specific to the image recognition systems we tested. Among the three systems we evaluated — GCV, Microsoft Azure Vision, and Amazon Rekognition — there was substantial gender bias in every system, but also variation in the specific content and magnitude of biases. Further, the algorithms

deployed by Google and other technology companies change frequently. To give one example, Google Cloud Vision has recently removed its gender-identification feature from all of their public-facing services (Ghosh 2020).

Furthermore, some kinds of labels we analyzed are not amenable to our bias measurement approach and, we argue, pose substantial measurement reliability challenges. A prominent example of this in our data was the label "smile," which was applied to women much more often than men in all three commercial image labeling systems we examined. GCV applied the label to Congresswomen more than 90% of time, while applying it to Congressmen less than 25% of the time. It would be tempting to do analysis of gender bias here: smiling is a highly-gendered behavior, particularly in images of women (Goffman 1974). But smiling is far more ambiguous to classify than labels like "hair," "outdoors," "child," or "military officer." Researchers who try to create metrics for what counts as a smile invariably find that age, race, gender, nationality, dental health, and more influence not only how people smile, but also whether observers see a particular facial expression as a smile (Jensen, Joss, and Lang 1999; Liébart et al. 2004). When one of the authors attempted to tally the presence of smiles in the congressional portraits data, this ambiguity rapidly became apparent: many facial expressions seemed borderline. Was that really a smile? Do smirks count? What if teeth are showing, but they do not seem happy? This is why flight attendants and other emotional laborers are formally trained not just that they are expected to smile, but specifically how they should be smiling (Hochschild, [1983] 2012). By our count, 91% of women and 86% of men were smiling—very far from the ratio of smiles in GCV labels and suggestive of substantial gender bias. But our recommendation is that researchers and users should avoid labels with this level of measurement ambiguity altogether.

Beyond simply calling attention to specific, significant gender biases in GCV, this article also serves as a template for future researchers seeking to use commercial algorithms. By comparing biases identified in uniform datasets as well as "found data," researchers will be better able to evaluate the tools they use before drawing firm conclusions from the data. Although our examples are primarily concerned with gender bias in image labeling, depending on the data set and research question, researchers may use the same procedures to test for bias along any trait and automated labeling system. As our crowdsourced validation suggests that humans predominantly agree with high-confidence labels by GCV, image recognition systems may still be useful for a variety of applications unaffected by gender biases. In any case, we recommend thorough validation efforts before using a commercial image recognition system. To simplify the process of annotating, validating and analyzing images with GCV, one of the authors of this paper has developed auxiliary open source software in form of an R package (https://cschwem2er.github.io/imgrec/).

The increased accessibility of computational tools generally, and computer vision specifically, presents a novel opportunity for social science researchers to expand the study of social life. However, researchers—and practitioners writ large— cannot treat such black-box tools as infallible. With tasks such as image labeling, there are nearly infinite potential labels to describe an image. If "a picture is worth a thousand words," but an algorithm provides only a handful, the words it chooses are of immense consequence. As some academic disciplines find themselves undergoing a "replication crisis," reliance on black-box tools that often change without notice can further exacerbate patterns of incorrect inference while even obscuring the methodology used to arrive at these results. As past trends in research methodology in the social sciences have illustrated (Shank 2016), research tools often grow in popularity before their biases and limitations are widely

understood. Therefore, our research serves as an injunction to future researchers seeking to break from, rather than reinforce, the biased gendered associations of the past.

References

Anastasopoulos, L J., Dhruvil Badani, Crystal Lee, Shiry Ginosar, and Jake Williams. 2016. "Photographic Home Styles in Congress: A Computer Vision Approach." *arXiv preprint arXiv:1611.09942*.

Becker, Howard S. 1974. "Photography and Sociology." *Studies in Visual Communication* 1(1):3–26.

Benjamin, Ruha. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. 1 edition. Medford, MA: Polity.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems*: 4349–57

Bosch, Oriol J., Melanie Revilla, and Ezequiel Paura. 2019. "Answering Mobile Surveys with Images: An Exploration Using a Computer Vision API." *Social Science Computer Review* 37(5):669–83.

Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49(2):388–405.

Buolamwini, Joy and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Conference on Fairness, Accountability and Transparency* 77–91.

Burri, Regula V. 2012. "Visual Rationalities: Towards a Sociology of Images." *Current Sociology* 60(1):45–60.

Casas, Andreu and Nora W. Williams. 2019. "Images That Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72(2):360–75.

Chen, Qiang et al. 2015. "Deep Domain Adaptation for Describing People Based on Fine-Grained Clothing Attributes." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition:* 5315–24.

Chen, Tao, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. "Understanding and Classifying Image Tweets." *Proceedings of the 21st ACM International Conference on Multimedia:* 781–84.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).

Cowgill, Bo. 2018. "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening." *Working Paper; Columbia Business School.*

Crawford, K and T Paglen. 2019. "Excavating AI: The Politics of Images in Machine Learning Training Sets." The AI Now Institute, NYU (available at https://www.excavating.ai)

De Vries, T., Misra, I., Wang, C., & van der Maaten, L. 2019. Does Object Recognition Work for Everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 52-59).

Di Ronco, Anna, and James Allen-Robertson. 2020. *Crime, Media, and Culture*. (available at: http://repository.essex.ac.uk/28271/3/Main%20Document.pdf)

Durepos, Gabrielle, Alan McKinlay, and Scott Taylor. 2017. "Narrating Histories of Women at Work: Archives, Stories, and the Promise of Feminism." *Business History* 59(8):1261–79.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Ferree, Myra M. and Elaine J. Hall. 1990. "Visual Images of American Society: Gender and Race in Introductory Sociology Textbooks." *Gender & Society* 4(4):500–533.

Forsyth, David A., and Jean Ponce. 2002. *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference.

Garimella, Kiran and Dean Eckles. 2020. "Images and Misinformation in Political Groups: Evidence from WhatsApp in India". (available at https://arxiv.org/pdf/2005.09784.pdf)

Gates, Susan W., Vanessa G. Perry, and Peter M. Zorn. 2002. "Automated Underwriting in Mortgage Lending: Good News for the Underserved?" *Housing Policy Debate* 13(2):369–91.

Geboers & Van de Wiele (2020): Geboers, Marloes Annette, and Chad Thomas Van De Wiele. "Machine Vision and Social Media Images: Why Hashtags Matter." Social Media and Society 6.2: 2056305120928485.

Gelman, Andrew, Greggor Mattson, and Daniel Simpson. 2018. "Gaydar and the Fallacy of Decontextualized Measurement." *Sociological Science* 5: 270–80.

Goffman, Erving. 1979. *Gender Advertisements*. Macmillan International Higher Education.

Grady, Cheryl L., Anthony R. McIntosh, M N. Rajah, and Fergus I. Craik. 1998. "Neural Correlates of the Episodic Encoding of Pictures and Words." *Proceedings of the National Academy of Sciences* 95(5):2703–8.

Greenfield, Sam. 2018. "Picture what the cloud can do: How the New York Times is using Google Cloud to find untold stories in millions of archived photos" *Google Cloud Blog*. (available at: https://cloud.google.com/blog/products/ai-machine-learning/how-the-new-york-times-is-using-google-cloud-to-find-untold-stories-in-millions-of-archived-photos)

Grother, Patrick, Mei Ngan, and Kayee Hanaoka. 2019. *Face Recognition Vendor Test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology.

Ghosh, Shona. 2020. "Google AI Will No Longer Use Gender Labels like 'Woman' or 'Man' on Images of People to Avoid Bias." Business Insider. February 20, 2020.

Hamidi, Foad, Morgan K. Scheuerman, and Stacy M. Branham. 2018. "Gender Recognition or Gender Reductionism." *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems:* 1–13.

HG Insights. 2020. "Companies Using Google Cloud Vision API, Market Share, Customers and Competitors." HG Insights. Retrieved August 3, 2020 (available athttps://discovery.hgdata.com/product/google-cloud-vision-api).

Huff, C., & Tingley, D. 2015) "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Research & Politics.

Jia, Sen, Thomas Lansdall-Welfare, and Nello Cristianini. 2015. "Measuring Gender Bias in News Images." *Proceedings of the 24th International Conference on World Wide Web:* 893–98.

Joo, Jungseock and Zachary C. Steinert-Threlkeld. 2018. "Image as Data: Automated Visual Content Analysis for Political Science." *arXiv preprint arXiv:1810.01544*.

Kay, Matthew, Cynthia Matuszek, and Sean A. Munson. 2015. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems:* 3819–28.

Keyes, Os. 2018. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133(1):237–93.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." *AEA papers and proceedings* 108:22–27.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review* 84 (5): 905–49.

Lagos, Danya. 2019. "Hearing Gender: Voice-Based Gender Classification Processes and Transgender Health Inequality." *American Sociological Review* 84(5):801–27.

Lam, Onyi, Brian Broderick, Stefan Wojcik, and Adam Hughes. n.d. "Gender and Jobs in Online Image Searches." *Pew Social Trends*. Retrieved March 14, 2020 (https://www.pewsocialtrends.org/2018/12/17/gender-and-jobs-in-online-image-searches/).

Long, J. Scott, and Jeremy Freese. 2006. "*Regression Models for Categorical Dependent Variables using Stata*. Stata press.

Michalski, D. 2017. "The impact of age-related variables on facial comparisons with images of children: Algorithm and practitioner performance," Doctoral Dissertation, University of Adelaide, Australia.

Mullainathan, Sendhil and Ziad Obermeyer. 2017. "Does Machine Learning Automate Moral Hazard and Error?" *The American Economic Review* 107(5):476–80.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2018. "The Future of Coding." *Sociological Methods & Research* 18(4):1-36.

Pauwels, Luc. 2010. "Visual Sociology Reframed: An Analytical Synthesis and Discussion of Visual Methods in Social and Cultural Research." *Sociological Methods & Research* 38(4):545–81.

Perryman, Nyssa, and Sandra Theiss. 2013. "'Weather Girls' on the Big Screen: Stereotypes, Sex Appeal, and Science." *Bulletin of the American Meteorological Society* 95 (3): 347–56.

Pittman, M. and Sheehan, K., 2016. Amazon's Mechanical Turk a digital sweatshop? Transparency and accountability in crowdsourced online research. Journal of media ethics, 31(4), pp.260-262.

Rogers, S. 2014. "What Fuels a Tweet's Engagement. Twitter Media Blog." *Twitter Media Blog* (available at https://blog.twitter.com/en_us/a/2014/what-fuels-a-tweets-engagement.html)

Rossiter, Margaret W. 1993. "The ~~Matthew~~ Matilda Effect in Science." *Social Studies of Science* 23 (2): 325–41.

Schwartz, Dona. 1989. “Visual Ethnography: Using Photography in Qualitative Research.” *Qualitative Sociology* 12(2):119–54.

Seamster, Louise, and Victor Ray. 2020. “Racism Without Race: Proxies and Algorithmic Inequality.” in American Sociological Association. San Francisco.

Sen, Shilad, Margaret E. Giesel, Rebecca Gold, Benjamin Hillmann, Matt Lesicko, Samuel Naden, Jesse Russell, Zixiao (Ken) Wang, and Brent Hecht. 2015. “Turkers, Scholars, ‘Arafat’ and ‘Peace’: Cultural Communities and Algorithmic Gold Standards.” In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW ’15*, 826–38. Vancouver, BC, Canada: ACM Press.

Shank, Daniel B. 2016. “Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk.” *The American Sociologist* 47(1):47–55.

Shor, Eran, Arnout van de Rijt, and Babak Fotouhi. 2019. “A Large-Scale Test of Gender Bias in the Media.” *Sociological Science* 6:526–50.

Torres Pacheco, Silvia Michelle. 2019. “A Visual Political World: Determinants and Effects of Visual Content.” Dissertation, Seattle: Washington University. (available at https://openscholarship.wustl.edu/art_sci_etds/1767).

United States Project. 2020. Github repository for data on Congress legislators. (available at https://github.com/unitedstates/congress-legislators)

Webb Williams, Nora, Andreu Casas, and John D. Wilkerson. 2020. Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification. 1st ed. Cambridge University Press.

West, Candace, and Don H. Zimmerman. 1987. “Doing Gender.” *Gender and Society* 1 (2): 125–51.

Whitehouse, Andrew J., Murray T. Maybery, and Kevin Durkin. 2006. “The Development of the Picture-Superiority Effect.” *British Journal of Developmental Psychology* 24(4):767–73.

Williamson, V., 2016. On the ethics of crowdsourced research. PS: Political Science & Politics, 49(1), pp.77-81.

Xi, Nan et al. 2019. “Understanding the Political Ideology of Legislators from Social Media Images.” *arXiv preprint arXiv:1907.09594*.

Author contributions

All authors jointly designed the study. J.L. devised the sampling strategy. C.S. and J.L. collected the data. C.K. managed human validation, with E.B.P and J.L. designing the validation survey. C.S. performed the statistical analysis and produced the replication materials, with input from all authors. C.K. and J.L. took the lead in writing the manuscript, with contributions from all authors. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Acknowledgments

The authors thank Laura Adler, Chris Bail, David Barker, Nel Escher, Parker Koch, Ryan T. Moore, Stefan Müller, Arvind Narayanan and Matthew Salganik for feedback on this project. We also thank participants at the 2018 European Symposium Series on Societal Challenges in Computational Social Science in Cologne; the 2019 IC2S2 in Amsterdam; the 2019 PSAI in Maynooth, Ireland, and American University's School of Public Affairs PhD student speaker series for feedback on this project.

Funding

We thank the Summer Institute in Computational Social Science (SICSS) and its funders, the Alfred P. Sloan Foundation and the Russell Sage Foundation, for funding and incubating this project. Lockhart received funding through an NICHD training grant to the Population Studies Center at the University of Michigan (T32HD007339). Schwemmer received funding through an NSF Award at the Center for Information Technology Policy of Princeton University (IIS-1763642).

Online Appendix: Diagnosing Gender Bias in Image Recognition Systems

This online appendix includes additional information for our crowdsourced validation, additional notes on methodology as well as results of additional analysis. We relied on a variety of open source tools to conduct our analyses. Replication materials, including the R code used to generate all results shown in the main paper as well as for those included in this supplementary file, will be available at Harvard Dataverse.

Crowdvalidation Survey

For human evaluation of image labels, we hired crowd workers on Amazon Mechanical Turk at a rate of $15/hour. The validation survey was reviewed and exempted by Institutional Review Boards at American University (Protocol#: IRB-2020-9), the University of Michigan (HUM00156287), and New York University (IRB-FY2019-3266). From the images tweeted by Members of Congress, we selected a stratified sample of N = 9,250 to conduct the external human validation of GCV labels. For stratification, we used a weighted randomization strategy. An image's weight is calculated using both the labels from Google Vision and by the characteristics of the Member of Congress (MC's) posting the image. Image weights are inversely proportional to how rare their features are, such that images with uncommon labels and coming from MCs from underrepresented groups are more likely to be sampled. Specifically, MCs' images were weighted by the inverse frequency of images from their state, party, house, age, gender, race/ethnicity, and unique ID in the set of all tweets, as well as by the inverse frequency of the labels GCV applied to them. We excluded some labels before sampling. Because the labels "font" and "text" were empirically found to be redundant, we dropped the "font" label. We also dropped 15,631 images that were only labeled "font" and "text," with no additional labels, as these were the most common images and least informative labels.

We further excluded images that were thumbnail previews of videos. In our validation data, we only selected labels that GCV assigned ≥ 0.75 confidence to. We presented each worker with 30 images and a set of potential labels for each image. Some labels were assigned by GCV (positive labels); others were chosen at random from the set of GCV labels assigned to other images but not to the one at hand (negative labels). We asked workers to select all labels that applied to each

image individually. Each image was coded by at least three workers. An example for one image as included in our validation survey is shown in the following Figure.

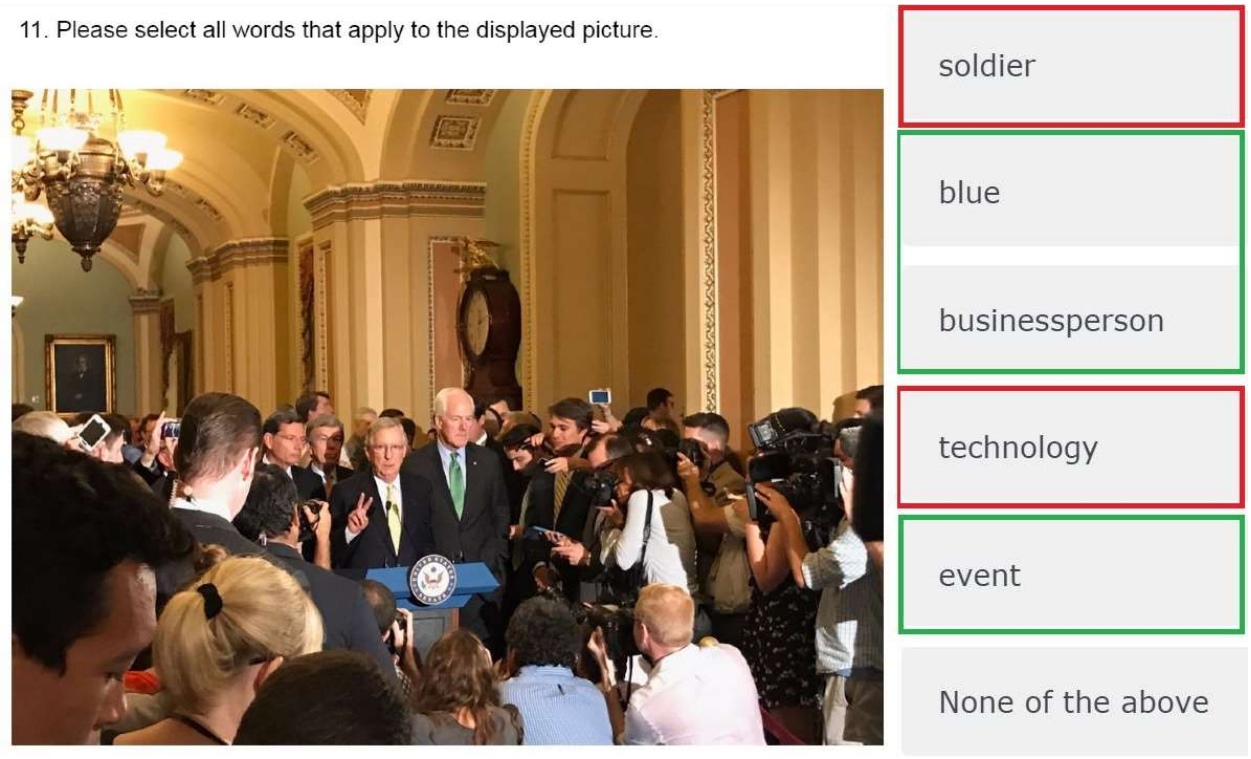

Figure A1. Example image as shown in our validation survey to crowd workers (without colored margins). Labels with green margins are positive (GCV annotations) and those with red margins are negative (chosen at random from all other labels).

Workers were required to live in the United States. In addition, we required workers to be at least 18 years old. The median age of workers in our sample is 33. Asked what ethnic group best describes themselves, 74% of workers selected “White". Regarding gender, about 60% identified themselves as men, 38% as women and 2% as non-binary, genderqueer, or something else. Workers were also asked with which party they identify: 47% identified themselves as Democrats, 26% as Republicans and 27% either identified themselves with another party or with no party at all. Workers generally agreed with one another. Each person validated the labels of 30 images, and multiple people saw each combination of labels and images. We compute each person’s agreement with the other workers as the fraction of their answers that match other workers’ answers to the same questions. If a worker is guessing at random, we can expect that they will have a label-wise agreement around 0.5, because the selection for each label is a binary choice. Values below 0.5

indicate that a worker disagrees with other workers about labels more often than they agree. Overall, the average agreement is at 0.77 (median=0.79).

Before proceeding with analysis, we excluded responses from 22 people (2.6% of workers) whose agreement with others is below 0.6. We suspect these workers either misunderstood the task or rushed to complete it by randomly guessing (in accordance with the financial incentives of MTurk). Perfect agreement is not expected in this task: many of the labels provided by Google Vision are ambiguous, such as "academic conference" and "tuxedo." A room full of people with a speaker at the front may look like an academic conference to some workers and a political fundraiser to others. Similarly, some may label any dark suit a "tuxedo," while others may draw sartorial distinctions between business wear and evening wear.

Additional notes on methodology

For identifying the most gendered labels assigned by image recognition software, we rely on two procedures in particular. First, we use χ^2 test statistics with Yates correction on labels returned by GCV, Amazon Rekognition and Microsoft Azure Vision. Our test statistics are specified as

$$\chi^2_{\text{corrected}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where O_i denotes the observed frequency of a label and E_i denotes the expected frequency of a label for men or women. Second, we use negative binomial regressions to obtain the expected counts of GCV labels in each coded category for men and women MCs while controlling for ethnicity and party. Regression models are specified as

$$\frac{E(y|\boldsymbol{x}, x_k + \delta)}{E(y|\boldsymbol{x}, x_k)} = \exp \beta_k \delta$$

such that for a change δ in variable x_k, the expected count of labels in a category, increases by a factor of $\exp \beta_k \delta$, holding all other variables constant (Long and Freese 2006).

Label categories for Google Cloud Vision

As explained in the main paper, we further categorized the image labels obtained by GCV for the professional photographs. Three coders, all of whom are authors of this paper, categorized labels into 5 larger categories (see table A1): physical trait / body, occupation, clothing and apparel, color / adjective, and other.

Table A1. Image labels as assigned by Google Cloud Vision as well as manually assigned categories.

Label	***Category***	***Label***	***Category***
actor	occupation	*long_hair*	physical trait / body
afro	physical trait / body	*magenta*	color / adjective
bangs	physical trait / body	*male*	physical trait / body
beard	physical trait / body	*man*	physical trait / body
beauty	physical trait / body	*management*	occupation
black_hair	physical trait / body	*military_officer*	occupation
blazer	clothing / apparel	*military_person*	occupation
blond	physical trait / body	*moustache*	physical trait / body
blue	color / adjective	*mouth*	physical trait / body
bob_cut	physical trait / body	*music_artist*	occupation
brown_hair	physical trait / body	*neck*	physical trait / body
business	occupation	*necktie*	clothing / apparel
businessperson	occupation	*newscaster*	occupation
carpet	other	*newsreader*	occupation
cheek	physical trait / body	*nose*	physical trait / body
chin	physical trait / body	*official*	occupation
dress_shirt	clothing / apparel	*outerwear*	clothing / apparel
ear	physical trait / body	*pattern*	other
elder	physical trait / body	*person*	other
electric_blue	color / adjective	*photograph*	other
energy	other	*photography*	other
eyebrow	physical trait / body	*pink*	color / adjective
eyewear	clothing / apparel	*pixie_cut*	physical trait / body
face	physical trait / body	*portrait*	other
facial_expression	physical trait / body	*portrait_photography*	other
facial_hair	physical trait / body	*product*	other
fashion	clothing / apparel	*public_speaking*	occupation
fashion_accessory	clothing / apparel	*purple*	color / adjective

flag	other	*red*	color / adjective
flag_of_the_united_states	other	*scarf*	clothing / apparel
forehead	physical trait / body	*scholar*	occupation
formal_wear	clothing / apparel	*sedan*	other
fun	color / adjective	*senior_citizen*	physical trait / body
gentleman	color / adjective	*shoulder*	physical trait / body
girl	physical trait / body	*skin*	physical trait / body
glasses	clothing / apparel	*sleeve*	clothing / apparel
grass	other	*smile*	physical trait / body
hair	physical trait / body	*speaker*	occupation
hair_coloring	physical trait / body	*speech*	other
hairstyle	physical trait / body	*spokesperson*	occupation
hat	clothing / apparel	*standing*	color / adjective
head	physical trait / body	*student*	occupation
headgear	clothing / apparel	*suit*	clothing / apparel
hotel_manager	occupation	*sunglasses*	clothing / apparel
human	color / adjective	*surfer_hair*	physical trait / body
human_hair_color	physical trait / body	*sweater*	clothing / apparel
iris	physical trait / body	*television_presenter*	occupation
jacket	clothing / apparel	*textile*	clothing / apparel
jaw	physical trait / body	*tooth*	physical trait / body
jheri_curl	physical trait / body	*turquoise*	color / adjective
job	occupation	*tuxedo*	clothing / apparel
lady	physical trait / body	*uniform*	clothing / apparel
laughter	physical trait / body	*vision_care*	clothing / apparel
layered_hair	physical trait / body	*white_collar_worker*	occupation
lip	physical trait / body		

Coding reached a satisfactory inter-rater agreement (kappa = 0.878). Our main interest concerns the extent to which men and women are tagged with different label categories as evidence for label bias or its absence. Using the modal category for labels, we then estimated negative binomial count models for each category, predicting the number of labels within that category as a function of the gender, ethnicity, *age,* and party of the portrayed MC. The results are presented in the main text and show that, when estimated on a sample of highly comparable professional photographs, the image labels produced by GCV are strongly gendered. Physical traits / body labels are assigned

much more often to images of women MCs, whereas occupation (and to a lesser extent clothing & apparel) are more prevalent among images of men MCs.

We find no such differences for ethnicity, *where we distinguish between "White" and "Non-White" MCs (see the Figure A2).* This may in part stem from the lack of ethnic diversity in the US Congress, as we were only able to compare categories "White" versus "Non-White" due to small number MCs who identify themselves with ethnic groups other than "White".

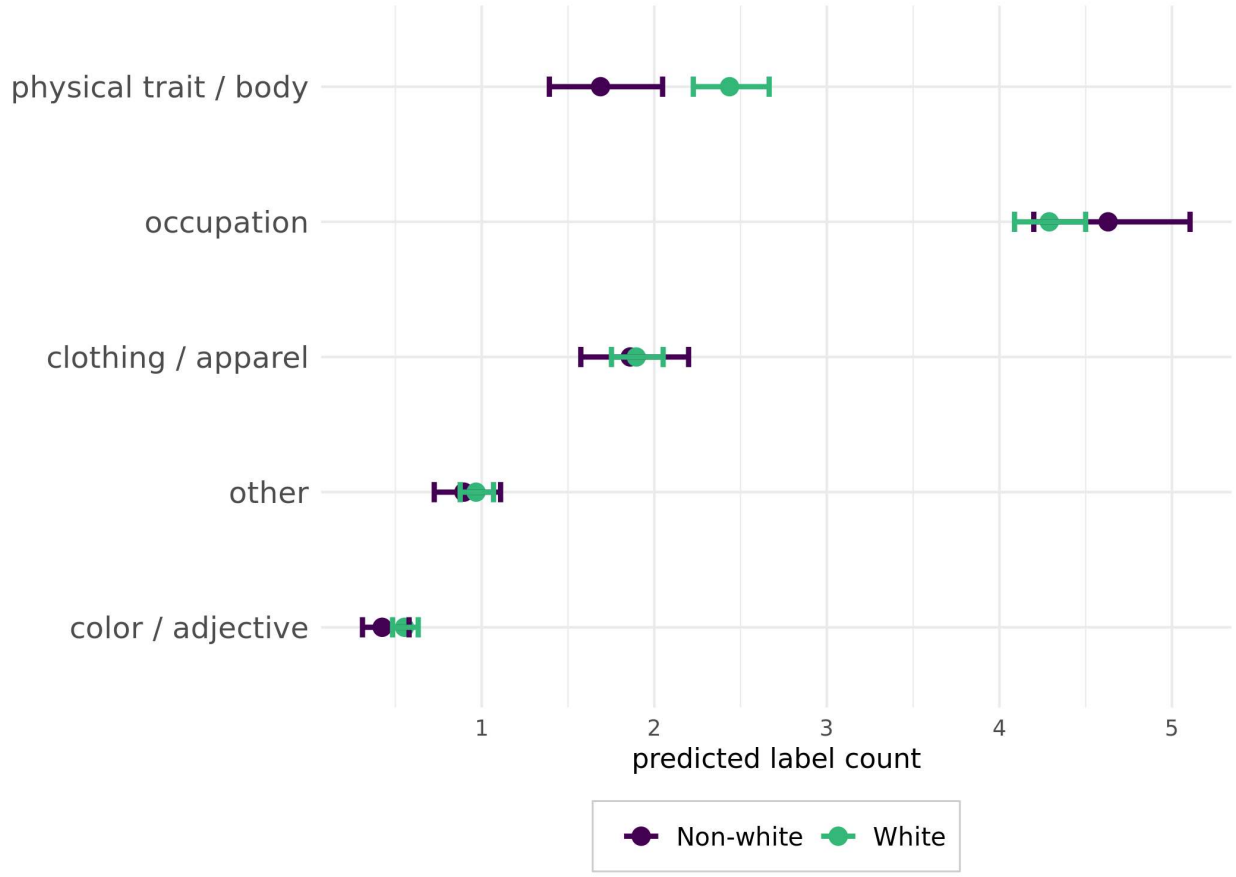

Figure A2. Predicted counts of Google Cloud Vision labels in comparison between images of White and Non-white persons. Results are based on photographs of U.S. Members of Congress and negative binomial regressions, controlling for gender, age and party. Circles describe point estimates, bars describe 95% confidence intervals.

Furthermore, we find no substantial differences between Democrats and Republicans and Representatives or Senators for these label categories (see Figure A3).

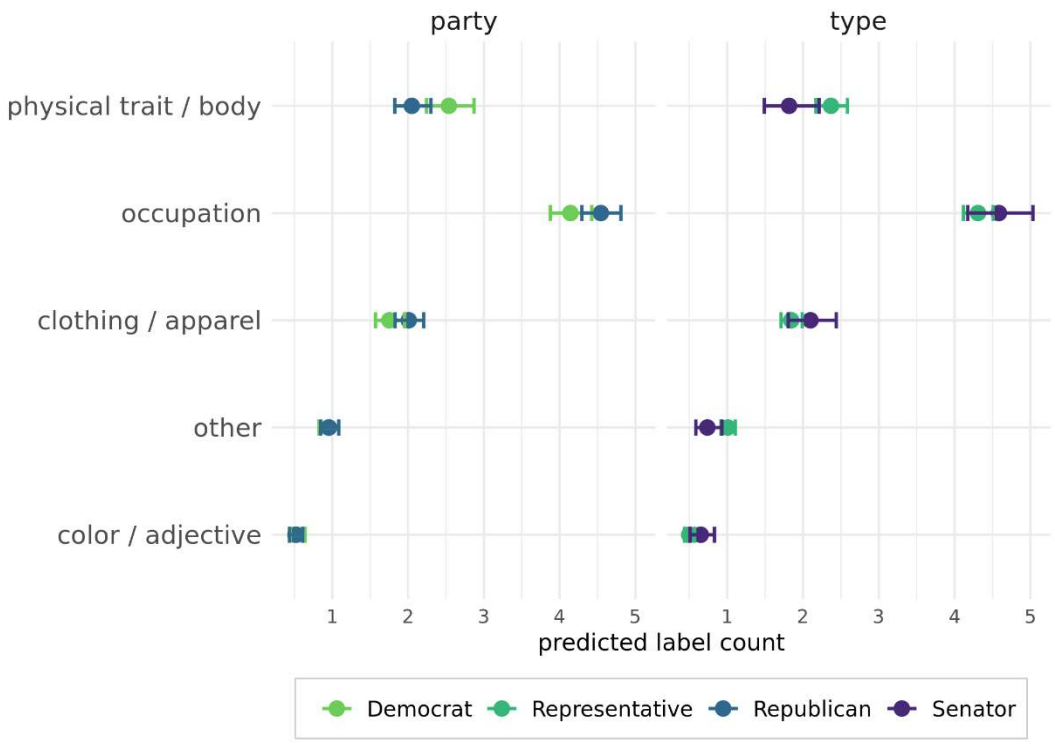

Figure A3. Predicted counts of Google Cloud Vision labels in comparison between images of Democrats and Republicans, and images of Representatives and Senators. Results are based on photographs of U.S. Members of Congress and negative binomial regressions, controlling for gender, age and ethnicity. Circles describe point estimates, bars describe 95% confidence intervals.

With regards to our control variable age, older Members of Congress were less likely to receive a label related to occupation or clothing & apparel, but more likely to receive labels related to physical traits & body. Women in are dataset are on average older than men (62 vs 60), and the youngest woman and man in our data are both 34 years.

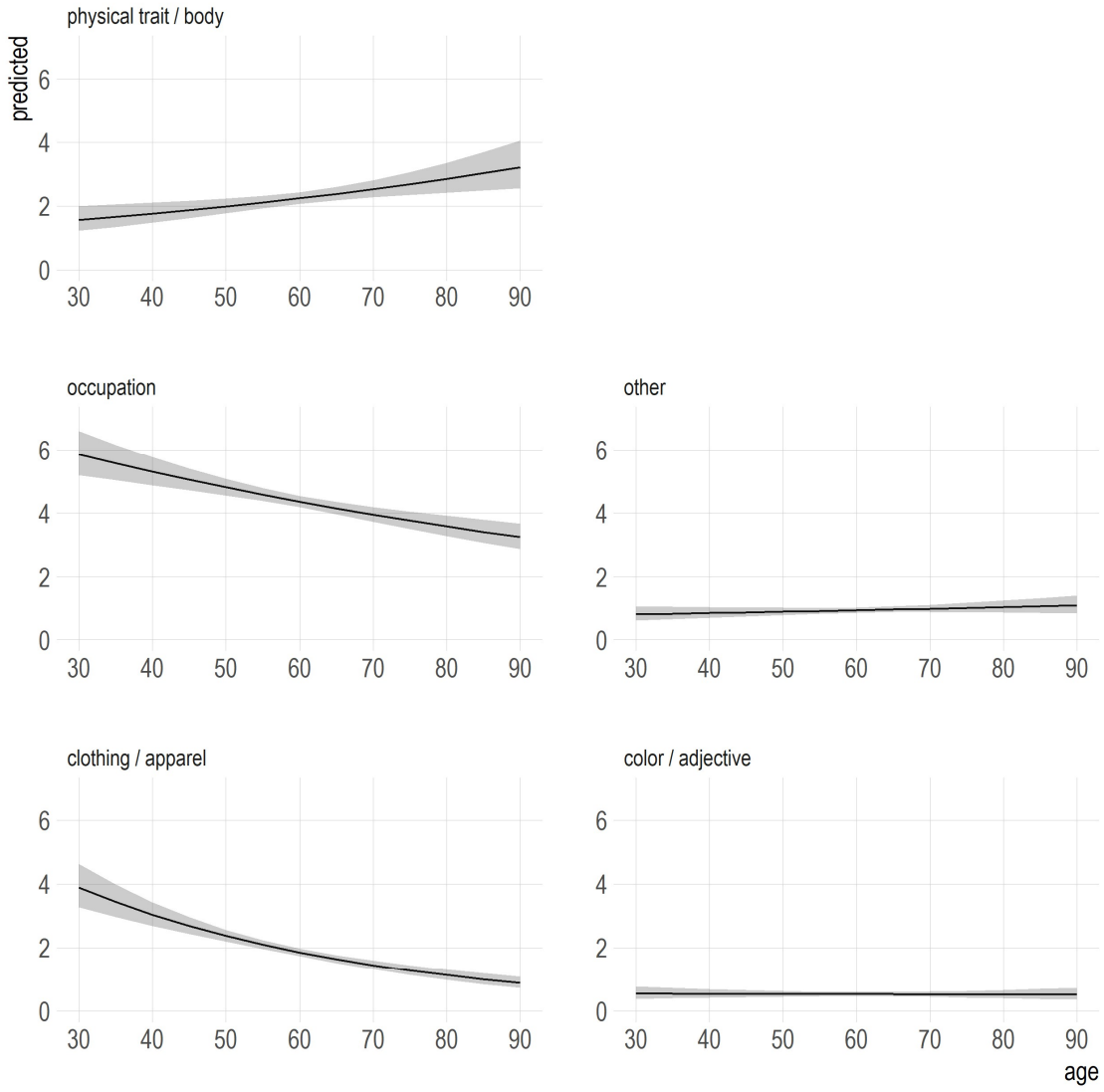

Figure A4. Predicted counts of Google Cloud Vision labels for different age values. Results are based on photographs of U.S. Members of Congress and negative binomial regressions, controlling for gender, party and ethnicity.

Label for Twitter images by Party

In our paper, we show that analyzing the key labels by gender for our Twitter dataset might lead to the wrong conclusion that MCs' gender substantially influences the content of the images they share on Twitter. Since we detect bias in how the Google Cloud Vision platform assigned the labels, we cannot be sure that the differences observed in the image labels are not at least partially

confounded by this bias. To examine whether gender differences in labels are similar in our Twitter dataset regardless of MC's partisanship, we again compute χ^2 statistics, but this time breaking out the results by both gender and party. For visualization purposes, Figure A5 only displays 20 labels per group without relative frequencies and sorted by χ^2 values. These results are similar to those broken out by gender alone. For example, the top label assigned to images tweeted by women in both parties is "girl", while the top labels for men in both parties are either text-related (e.g. Democratic men) or "official", "suit", and "businessperson" (e.g. Republican men). Without detecting biases in the GCV platform, these differences would suggest that MCs' online behavior is different in terms of gender.

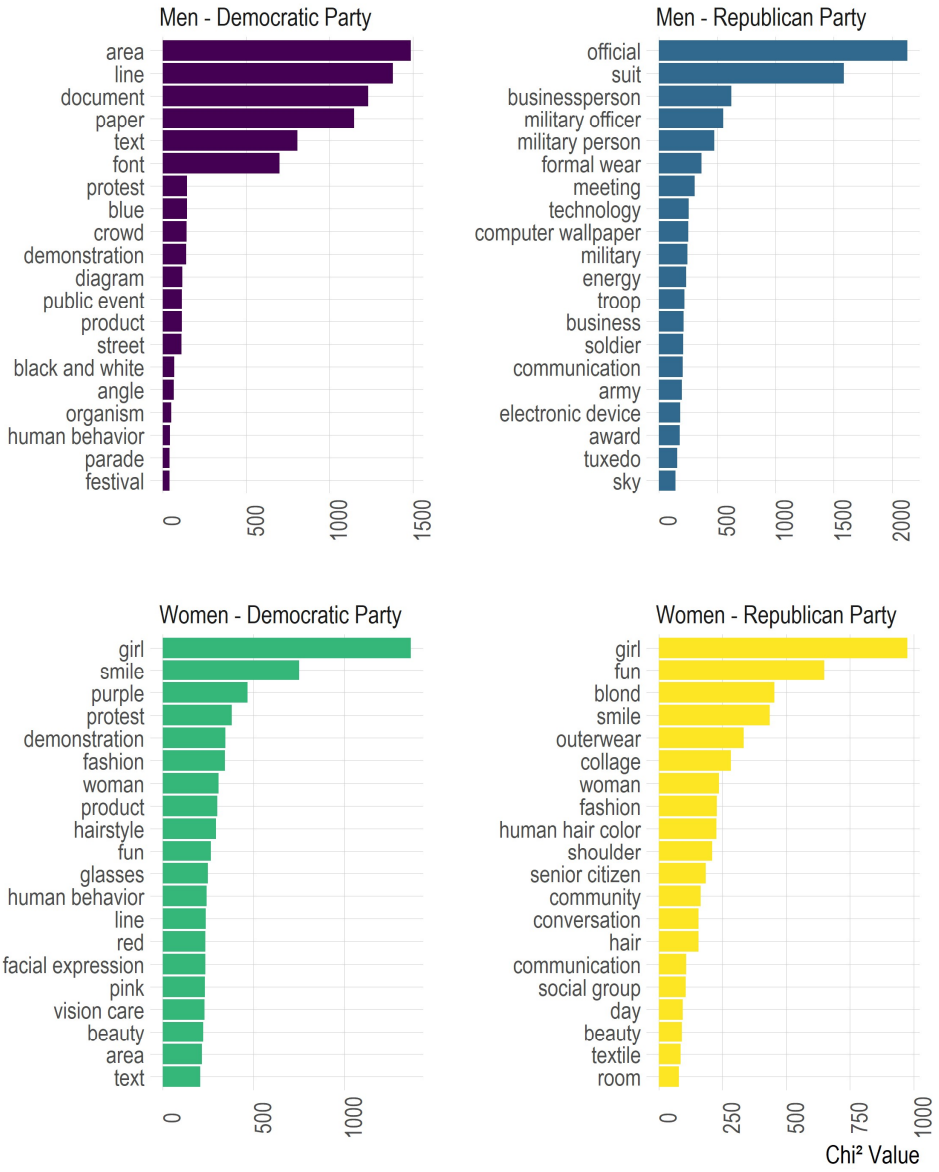

Figure A5. Google Cloud Vision labels applied to images tweeted by Members of Congress. The 25 most gendered labels by party and gender were identified with χ^2 tests. Bars denote χ^2 tests results.

Image examples for women recognized / not recognized by GCV

In our paper, we show that the object recognition module of GCV correctly identifies women in images at substantially lower rates in comparison to men. We also qualitatively examined whether any particular features of images could be an indicator for whether men and women are recognized.

We did not find any of these features. The following figure shows example of men and women recognized (left-hand column) and not recognized (right-hand column) by GCV.

Figure A6. Examples for images of men and women recognized / not recognized by Google Cloud Vision's object recognition. Top-left: Maria Cantwell, recognized. Top-right: Kathy Castor, not recognized. Bottom-left: Ben Cardin, recognized. Bottom-right: Tom Carper, not recognized.