Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic and occupation shifts over time. The embedding captures societal shifts—e.g., the women's movement in the 1960s and Asian immigration into the United States—and also illuminates how specific adjectives and occupations became more closely associated with certain populations over time. Our framework for temporal analysis of word embedding opens up a fruitful intersection between machine learning and quantitative social science.

word embedding | gender stereotypes | ethnic stereotypes

The study of gender and ethnic stereotypes is an important topic across many disciplines. Language analysis is a standard tool used to discover, understand, and demonstrate such stereotypes (1–5). Previous literature broadly establishes that language both reflects and perpetuates cultural stereotypes. However, such studies primarily leverage human surveys (6–16), dictionary and qualitative analysis (17), or in-depth knowledge of different languages (18). These methods often require time-consuming and expensive manual analysis and may not easily scale across types of stereotypes, time periods, and languages. In this paper, we propose using word embeddings, a commonly used tool in natural language processing (NLP) and machine learning, as a framework to measure, quantify, and compare beliefs over time. As a specific case study, we apply this tool to study the temporal dynamics of gender and ethnic stereotypes in the 20th and 21st centuries in the United States.

In word-embedding models, each word in a given language is assigned to a high-dimensional vector such that the geometry of the vectors captures semantic relations between the words—e.g., vectors being closer together has been shown to correspond to more similar words (19). These models are typically trained automatically on large corpora of text, such as collections of Google News articles or Wikipedia, and are known to capture relationships not found through simple co-occurrence analysis. For example, the vector for France is close to vectors for Austria and Italy, and the vector for XBox is close to that of PlayStation (19). Beyond nearby neighbors, embeddings can also capture more global relationships between words. The difference between London and England—obtained by simply subtracting these two vectors—is parallel to the vector difference between Paris and France. This pattern allows embeddings to capture analogy relationships, such as London to England is as Paris to France.

Recent works demonstrate that word embeddings, among other methods in machine learning, capture common stereotypes because these stereotypes are likely to be present, even if subtly, in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study historical trends—specifically trends in the gender and ethnic stereotypes in the 20th and 21st centuries in the United States. We develop a systematic framework and metrics to analyze word embeddings trained over 100 y of text corpora. We show that temporal dynamics of the word embedding capture changes in gender and ethnic stereotypes over time. In particular, we quantify how specific biases decrease over time while other stereotypes increase. Moreover, dynamics of the embedding strongly correlate with quantifiable changes in US society, such as demographic and occupation shifts. For example, major transitions in the word embedding geometry reveal changes in the descriptions of genders and ethnic groups during the women's movement in the 1960s–1970s and Asian-American population growth in the 1960s and 1980s. We validate our findings on external metrics and show that our results are robust to the different algorithms for training the word embeddings. Our framework reveals and quantifies how stereotypes toward women and ethnic groups have evolved in the United States.

Significance

Word embeddings are a popular machine-learning method that represents each English word by a vector, such that the geometry between these vectors captures semantic relations between the corresponding words. We demonstrate that word embeddings can be used as a powerful tool to quantify historical trends and social change. As specific applications, we develop metrics based on word embeddings to characterize how gender stereotypes and attitudes toward ethnic minorities in the United States evolved during the 20th and 21st centuries starting from 1910. Our framework opens up a fruitful intersection between machine learning and quantitative social science.

Author contributions: N.G., L.S., D.J., and J.Z. designed research; N.G. and J.Z. performed research; and N.G. and J.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: Data and code related to this paper are available on GitHub (https://github.com/nikhgarg/EmbeddingDynamicStereotypes).

[1]To whom correspondence may be addressed. Email: nkgarg@stanford.edu or jamesz@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720347115/-/DCSupplemental.

Our results demonstrate that word embeddings are a powerful lens through which we can systematically quantify common stereotypes and other historical trends. Embeddings thus provide an important quantitative metric which complements existing, more qualitative, linguistic and sociological analyses of biases. In *Embedding Framework Overview and Validations*, we validate that embeddings accurately capture sociological trends by comparing associations in the embeddings with census and other externally verifiable data. In *Quantifying Gender Stereotypes* and *Quantifying Ethnic Stereotypes* we apply the framework to quantify the change in stereotypes of women, men, and ethnic minorities. We further discuss our findings in *Discussion* and provide additional details in *Materials and Methods*.

Embedding Framework Overview and Validations

In this section, we briefly describe our methods and data and then validate our findings. We focus on showing that word embeddings are an effective tool to study historical biases and stereotypes by relating measurements from these embeddings to historical census and survey data. The consistent replication of such historical data, both in magnitude and in direction of biases, validates the use of embeddings in such work. This section extends the analysis of refs. 20 and 21 in showing that embeddings can also be used as a comparative tool over time as a consistent metric for various biases.

Summary of Data and Methods. We now briefly describe our datasets and methods, leaving details to *Materials and Methods* and *SI Appendix*, section A. All of our code and embeddings are available publicly*. For contemporary snapshot analysis, we use the standard Google News word2vec vectors trained on the Google News dataset (24, 25). For historical temporal analysis, we use previously trained Google Books/Corpus of Historical American English (COHA) embeddings, which are a set of nine embeddings, each trained on a decade in the 1900s, using the COHA and Google Books (26). As additional validation, we train, using the GLoVe algorithm (27), embeddings from the *New York Times* Annotated Corpus (28) for every year between 1988 and 2005. We then collate several word lists to represent each gender† (men, women) and ethnicity‡ (White, Asian, and Hispanic), as well as neutral words (adjectives and occupations). For occupations, we use historical US census data (29) to extract the percentage of workers in each occupation that belong to each gender or ethnic group and compare it to the bias in the embeddings.

Using the embeddings and word lists, one can measure the strength of association (embedding bias) between neutral words and a group. As an example, we overview the steps we use to quantify the occupational embedding bias for women. We first compute the average embedding distance between words that represent women—e.g., she, female—and words for occupations—e.g., teacher, lawyer. For comparison, we also compute the average embedding distance between words that represent men and the same occupation words. A natural metric for the embedding bias is the average distance for women minus the average distance for men. If this value is negative, then the embedding more closely associates the occupations with men. More generally, we compute the representative group vector by taking the average of the vectors for each word in the given gender/ethnicity group. Then we compute the average Euclidean distance between each representative group vector and each vector in the neutral word list of interest, which could be occupations or adjectives. The difference of the average distances is our metric for bias—we call this the relative norm difference or simply embedding bias.

We use ordinary least-squares regressions to measure associations in our analysis. In this paper, we report r^2 and the coefficient P value for each regression, along with the intercept confidence interval when relevant.

Validation of the Embedding Bias. To verify that the bias in the embedding accurately reflects sociological trends, we compare the trends in the embeddings with quantifiable demographic trends in the occupation participation, as well as historical surveys of stereotypes. First, we use women and minority ethnic participation statistics (relative to men and Whites, respectively) in different occupations as a benchmark because it is an objective metric of social changes. We show that the embedding accurately captures both gender and ethnic occupation percentages and consistently reflects historical changes.

Next, we validate that the embeddings capture personality trait stereotypes. A difficulty in social science is the relative dearth of historical data to systematically quantify stereotypes, which highlights the value of our embedding framework as a quantitative tool but also makes it challenging to directly confirm our findings on adjectives. Nevertheless, we make use of the best available data from historical surveys, gender stereotypes from 1977 and 1990 (6, 7) and ethnic stereotypes from the Princeton trilogy from 1933, 1951, and 1969 (8–10).

Comparison with women's occupation participation. We investigate how the gender bias of occupations in the word embeddings relates to the empirical percentage of women in each of these occupations in the United States. Fig. 1 shows, for each occupation, the relationship between the relative percentage (of women) in the occupation in 2015 and the relative norm distance between words associated with women and men in the Google News embeddings. (Occupations whose 2015 percentage is not available, such as midwife, are omitted. We further note that the Google News embedding is trained on a corpus



Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

* All of our own data and analysis tools are available on GitHub at https://github.com/nikhgarg/EmbeddingDynamicStereotypes. Census data are available through the Integrated Public Use Microdata Series (29). We link to the sources for each embedding used in *Materials and Methods*.

† There is an increasingly recognized difference between sex and gender and thus between the words male/female and man/woman, as well as nonbinary categories. We limit our analysis to the two major binary categories due to technical limitations, and we use male and female as part of the lists of words associated with men and women, respectively, when measuring gender associations. We also use results from refs. 6 and 7 which study stereotypes associated with sex.

‡ When we refer to Whites or Asians, we specifically mean the non-Hispanic subpopulation. For each ethnicity, we generate a list of common last names among the group. Unfortunately, our present methods do not extend to Blacks due to large overlaps in common last names among Whites and Blacks in the United States.

over time, and so the 2015 occupations are not an exact comparison.) The relative distance in the embeddings significantly correlates with the occupation percentage ($P < 10^{-10}$, $r^2 = 0.499$). It is interesting to note that the regression line nearly intersects the origin [intercept in $(-0.021, -0.002)$]: Occupations that are close to 50–50 in gender participation have small embedding bias. These results suggest that the embedding bias correctly matches the magnitude of the occupation frequency, along with which gender is more common in the occupation.

We ask whether the relationship between embedding and occupation percentage holds true for specific occupations. We perform the same embedding bias vs. occupation frequency analysis on a subset of occupations that are deemed "professional" (e.g., nurse, engineer, judge; full list in *SI Appendix*, section A.3) and find nearly identical correlation [$P < 10^{-5}$, $r^2 = 0.595$, intercept in $(-0.026, 0)$]. We further validate this association using different embeddings trained on Wikipedia and Common Crawl texts instead of Google News; see *SI Appendix*, section B.1 for details.

The Google News embedding reveals one aggregate snapshot of the bias since it is trained over a pool of news articles. We next analyze the embedding of each decade of COHA from 1910 to 1990 separately to validate that for a given historical period, the embedding bias from data in that period accurately reflects occupation participation. For each decade, the embedding gender bias is significantly correlated with occupation frequency ($P \leq 0.003$, $r^2 \geq 0.123$), as in the case with the Google News embedding; however, we note that the intercepts here show a consistent additional bias against women for each decade; i.e., even occupations with the same number of men and women are closer to words associated with men.

More importantly, these correlations are very similar over the decades, suggesting that the relationship between embedding bias score and "reality," as measured by occupation participation, is consistent over time. We measure this consistency in several ways. We first train a single model for all (occupation percentage, embedding bias) pairs across time. We compare this model to a model where there is an additional term for each year and show that the models perform similarly ($r^2 = 0.236$ vs. $r^2 = 0.298$). Next, we compare the performance of the model without terms for each year to models trained separately for each year, showing that the single model both has similar parameters and performance to such separate models. Finally, for each embedding year, we compare performance of the model trained for that embedding vs. a model trained using all other data (leave-one-out validation). We repeat the entire analysis with embeddings trained using another algorithm on the same dataset [singular value decomposition (SVD)]. See *SI Appendix*, section B.3.1 for details.

This consistency makes the interpretation of embedding bias more reliable; i.e., a given bias score corresponds to approximately the same percentage of the workforce in that occupation being women, regardless of the embedding decade.

Next, we ask whether the changes in embeddings over decades capture changes in the women's occupation participation. Fig. 2 shows the average embedding bias over the occupations over time, overlaid with the average women's occupation relative percentage over time. [We include only occupations for which census data are available for every decade and which are frequent enough in all embeddings. We use the linear regression mapping inferred from all of the data across decades to align the scales for the embedding bias and occupation frequency (the two y axes in the plot).] The average bias closely tracks with the occupation percentages over time. The average bias is negative, meaning that occupations are more closely associated with men than with women. However, we see that the bias steadily moves closer to 0 from the 1950s to the 1990s, suggesting that the bias

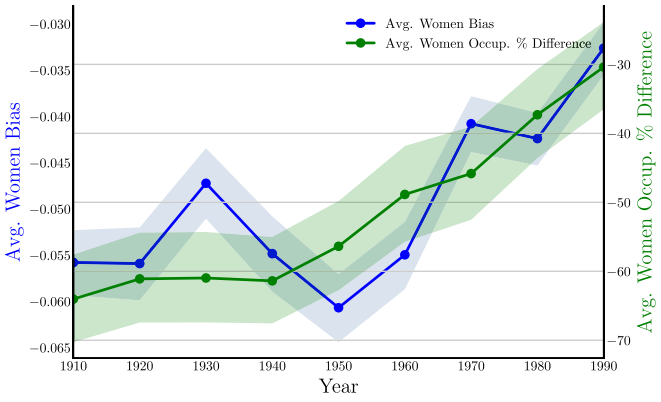


Fig. 2. Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.

is decreasing. This trend tracks with the proportional increase in women's participation in these occupations.

Comparison with ethnic occupation participation. Next, we compare ethnic bias in the embeddings to occupation participation rates and stereotypes. As in the case with gender, the embeddings capture externally validated ethnic bias. Table 1 shows the 10 occupations that are the most biased toward Hispanic, Asian, and White last names[§]. The Asian-American "model minority" (30, 31) stereotype appears predominantly; academic positions such as professor, scientist, and physicist all appear among the top Asian-biased occupations. Similarly, White and Hispanic stereotypes also appear in their respective lists. [Smith, besides being an occupation, is a common White-American last name. It is thus excluded from regressions, as are occupations such as conductor, which have multiple meanings (train conductors as well as music conductors).] As in the case with gender, the embedding bias scores are significantly correlated with the ethnic group's relative percentage of the occupation as measured by the US Census in 2010. For Hispanics, the bias score is a significant predictor of occupation percentage at $P < 10^{-5}$, $r^2 = 0.279$ and, for Asians, at $P = 0.041$, $r^2 = 0.065$. Due to the large population discrepancy between Whites and each respective minority group, the intercept values for these plots are large and are difficult to interpret and so are excluded from the main exposition (see *Discussion* for further details). The corresponding scatter plots and regression tables of embedding bias vs. occupation relative percentage are in *SI Appendix*, section C.1.

Similarly, as for gender, we track the occupation bias score over time and compare it to the occupation relative percentages; Fig. 3 does so for Asian Americans, relative to Whites, in the COHA embeddings. The increase in occupation relative percentage across all occupations is well tracked by the bias in the embeddings. More detail and a similar plot with Hispanic Americans are included in *SI Appendix*, section C.3.

Comparison with surveys of gender stereotypes. Now, we validate that the historical embeddings also capture gender stereotypes of personality traits. We leverage sex stereotype scores assigned to a set of 230 adjectives (300 adjectives are in the original studies; 70 adjectives are discarded due to low frequencies

[§]We adapt the relative norm distance in Eq. **3** for three groups. For each group, we compare its norm bias with the average bias of the other groups; i.e., bias(group 1) $= \sum_w \left[\frac{1}{2}(\|w - v_2\| + \|w - v_3\|) - \|w - v_1\|\right]$. This method can lead to the same occupation being highly ranked for multiple groups, such as happens for mason.

in the COHA embeddings) by human participants (6, 7). Participants scored each word for its association with men or women (example words: headstrong, quarrelsome, effeminate, fickle, talkative). This human subject study was first performed in 1977 and then repeated in 1990. We compute the correlation between the adjective embedding biases in COHA 1970s and 1990s with the respective decade human-assigned scores. In each case, the embedding bias score is significantly correlated with the human-assigned scores [$P < 0.0002, r^2 \geq 0.095$, intercepts in $(-0.017, -0.012)$ and $(-0.029, -0.024)$, respectively). *SI Appendix*, section B.3 contains details of the analysis. These analyses suggest that the embedding gender bias effectively captures both occupation frequencies as well as human stereotypes of adjectives, although noisily.

Comparison with surveys of ethnic stereotypes. We validate that the embeddings capture historical personality stereotypes toward ethnic groups. We leverage data from the well-known Princeton trilogy experiments (8–10), published in 1933, 1951, and 1969, respectively. These experiments have sparked significant discussion, follow-up work, and methodological criticism (11–16), but they remain our best method to validate our quantification of historical stereotypes.

These works surveyed stereotypes among Princeton undergraduates toward 10 ethnic groups, including Chinese people. (Other groups include Germans, Japanese, and Italians. We focus on Chinese stereotypes due to the ability to distinguish last names and a sufficient quantity of data in the embeddings.) Katz and Braly in 1933 reported the top 15 stereotypes attached to each group from a larger list of words (8) (example stereotypes: industrious, superstitious, nationalistic). (Each stereotype score is the percentage of respondents who indicated that the stereotype applies to the group. Note that these scores are not comparative across groups; i.e., a stereotype's score for one group does not directly imply its score for any other group, and so the regression intercepts are not meaningful.) In 1969, Karlins et al. reported scores for the same 15 stereotypes, among others (10). Scores for a subset of these adjectives were also reported in 1951 (9).

Using the stereotypes of Chinese people and our list of Chinese last names, we conduct two tests: First, using all reported scores for which there is sufficient text data, we correlate the stereotype scores with the given stereotype's embedding bias in the corresponding decade; second, using the stereotypes for which both 1933 and 1969 scores are available, we correlate the change in the scores with the change in the embedding bias during the period.

The results suggest, as in the case with gender, that adjective stereotypes in the embeddings reflect attitudes of the times and that the embeddings are calibrated across time. In our first test, the studies' stereotype scores are significant predictors of the corresponding embedding biases ($r^2 = 0.146,\ P = 0.023$).

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

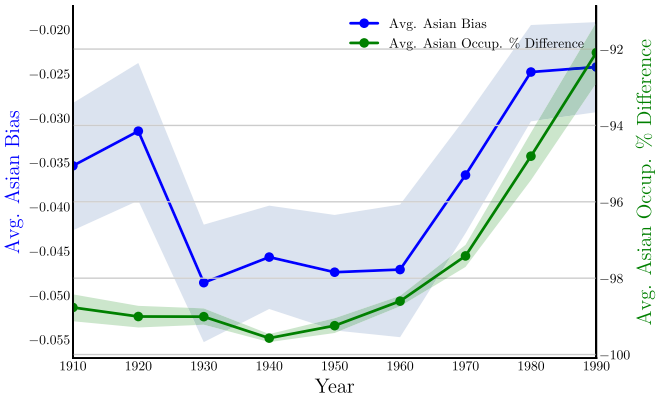


Fig. 3. Average ethnic (Asian vs. White) bias score over time for occupations in COHA (blue) vs. the average percentage of difference (green). Each shaded region is the bootstrap SE interval.

In the second test, the changes in the scores are also significant predictors of the changes in embedding biases ($r^2 = 0.472$, $P = 0.014$). See *SI Appendix*, section C.2 for regression tables and plots.

Together, the analyses in this section validate that embeddings capture historical attitudes toward both ethnic and gender groups, as well as changes in these attitudes. In the remainder of this work, we use this insight to explore such historical stereotypes to display the power of this framework.

Quantifying Gender Stereotypes

We now apply our framework to study trends in gender bias in society, both historically and in modern times. We first show that language today, such as that in the Google News corpora, is even more biased than could be accounted for by occupation data. In addition, we show that bias, as seen through adjectives associated with men and women, has decreased over time and that the women's movement in the 1960s and 1970s especially had a systemic and drastic effect in women's portrayals in literature and culture.

Due to the relative lack of systematic quantification of stereotypes in the literature, a gap that this work seeks to address, we cannot directly validate the results in this section or the next. We reference sociological literature and use statistical tests as appropriate to support the analyses.

Occupational Stereotypes Beyond Census Data. While women's occupation percentages are highly correlated with embedding gender bias, we hypothesize that the embedding could reflect additional social stereotypes beyond what can be explained by occupation participation. To test this hypothesis, we leverage the gender stereotype scores of occupations, as labeled by people on Amazon Mechanical Turk and provided to us by the authors of ref. 20[¶]. These crowdsource scores reflect aggregate human judgment as to whether an occupation is stereotypically associated with men or women. (A caveat here is that the US-based participants on Amazon Mechanical Turk may not represent the US population.) In separate regressions, both the crowdsourced stereotype scores [$r^2 = 0.655, P < 10^{-10}$, intercept confidence interval $(-0.281, 0.027)$] and the occupation relative percentage [$r^2 = 0.452, P < 10^{-6}$, intercept confidence

[¶] List of occupations available is in *SI Appendix*, section A.3. Note that the crowdsourcing experiment collected data for a larger list of occupations; we select the occupations for which both census data and embedding orientation are also available. For this reason, the regressions with just the occupation percentage score are slightly different from those in Fig. 1.

interval $(-0.027, -0.001)$] are significantly correlated with the embedding bias.

Next, we conduct two additional separate regressions to test that the embedding bias captures the same extra stereotype information as do the crowdsource scores, information that is missing in the census data. In each regression, the occupation percentage difference is the independent covariate. In one, the embedding bias is the dependent variable; in the other, stereotype score is. In these regressions, a negative (positive) residual indicates that the embedding bias or stereotype score is closer to words associated with women (men) than is to be expected given the gender percentages in the occupation. We find that the residuals between the two regressions correlate significantly (Pearson coefficient 0.811, $P < 10^{-10}$). This correlation suggests that the embedding bias captures the crowdsource human stereotypes beyond that which can be explained by empirical differences in occupation proportions.

Where such crowdsourcing is not possible, such as in studying historical biases, word embeddings can thus further serve as an effective measurement tool. Further, although the analysis in the previous section shows a strong relationship between census data and embedding bias, it is important to note that biases beyond census data also appear in the embedding.

Quantifying Changing Attitudes with Adjective Embeddings. We now apply the insight that embeddings can be used to make comparative statements over time to study how the description of women—through adjectives—in literature and the broader culture has changed over time. Using word embeddings to analyze biases in adjectives could be an especially useful approach because the literature is lacking systematic and quantitative metrics for adjective biases. We find that—as a whole—portrayals have changed dramatically over time, including for the better in some measurable ways. Furthermore, we find evidence for how the women's movement in the 1960s and 1970s led to a systemic change in such portrayals.

How overall portrayals change over time. We first establish that comparing the embeddings over time could reveal global shifts in society in regard to gender portrayals. Fig. 4 shows the Pearson correlation in embedding bias scores for adjectives over time between COHA embeddings for each pair of decades. As expected, the highest correlation values are near the diagonals; embeddings (and attitudes) are most similar to those from adjacent decades. More strikingly, the matrix exhibits two clear blocks. There is a sharp divide between the 1960s and 1970s, the height of the women's movement in the United States, during which there was a large push to end legal and social barriers for women in education and the workplace (32, 33). The transition in the gender embeddings from 1960 to 1970 is statistically significant ($P < 10^{-4}$, Kolmogorov–Smirnov two-sample test) and is larger than the change between any two other adjacent decades. See *SI Appendix*, section B.3.3 for a more detailed description of the test and all statistics.

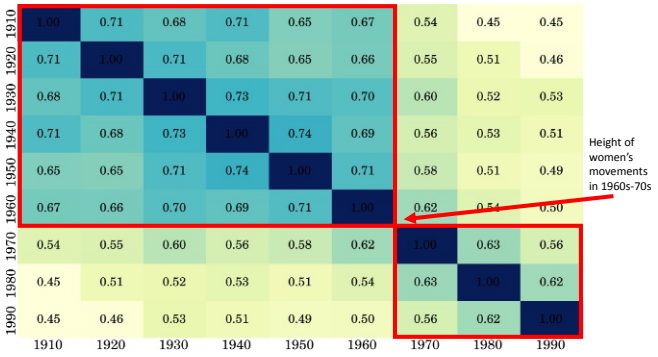


Fig. 4. Pearson correlation in embedding bias scores for adjectives over time between embeddings for each decade. The phase shift in the 1960s–1970s corresponds to the US women's movement.

We note that the effects of the women's movement, including on inclusive language, are well documented (18, 33–36); this work provides a quantitative way to measure the rate and extent of the change. A potential extension and application of this work would be to study how various narratives and descriptions of women developed and competed over time.

Individual words whose biases changed over time. As an example of such work, we consider a subset of the adjectives describing competence, such as intelligent, logical, and thoughtful (see *SI Appendix*, section A.3 for a full list of words; these words were curated from various online sources). Since the 1960s, this group of words on average has increased in association with women over time (from strongly biased toward men to less so): In a regression with embedding bias from each word as the dependent variable and years from 1960 to 1990 as the covariate, the coefficient is positive; i.e., there is a (small) positive trend (0.005 increase in women association per decade, $P = 0.0036$). At this rate, such adjectives would be equally associated with women as with men a little after the year 2020.

As a comparison, we also analyze a subset of adjectives describing physical appearance—e.g., attractive, ugly, and fashionable—and the bias of these words did not change significantly since the 1960s (null hypothesis of no trend not rejected with $P > 0.2$). Although the trend regarding intelligence is encouraging, the top adjectives are still potentially problematic, as displayed in Table 2.

We note that this analysis is an exploration; perceived competence and physical appearance are just two components of gender stereotypes. Models in the literature suggest that stereotypes form along several dimensions, e.g., warmth and competence (16). A more complete analysis would first collect externally validated lists of words that describe each such dimension and then measure the embedding association with respect to these lists over time.

The embedding also reveals interesting patterns in how individual words evolve over time in their gender association. For example, the word hysterical used to be, until the mid-1900s, a catchall term for diagnosing mental illness in women but has since become a more general word (37); such changes are clearly reflected in the embeddings, as hysterical fell from a top 5 woman-biased word in 1920 to not in the top 100 in 1990 in the COHA embeddings[#]. On the other hand, the word emotional becomes much more strongly associated with women over time in the embeddings, reflecting its current status as a word that is largely associated with women in a pejorative sense (38).

These results together demonstrate the value and potential of leveraging embeddings to study biases over time. The embeddings capture subtle individual changes in association, as well as larger historical changes. Overall, they paint a picture of a society with decreasing but still significant gender biases.

Quantifying Ethnic Stereotypes

We now turn our attention to studying ethnic biases over time. In particular we show how immigration and other 20th-century trends broadly influenced how Asians were viewed in the United States. We also show that embeddings can serve as effective tools to analyze finer-grained trends by analyzing the portrayal of Islam in the *New York Times* from 1988 to 2005 in the context of terrorism.

[#]We caution that due to the noisy nature of word embeddings, dwelling on individual word rankings in isolation is potentially problematic. For example, hysterical is more highly associated with women in the Google News vectors than emotional. For this reason we focus on large shifts between embeddings.

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Trends in Asian Stereotypes. To study Asian stereotypes in the embeddings, we use common and distinctly Asian last names, identified through a process described in *SI Appendix*, section A.2. This process results in a list of 20 last names that are primarily but not exclusively Chinese last names.

The embeddings illustrate a dramatic story of how Asian-American stereotypes developed and changed in the 20th century. Fig. 5 shows the Pearson correlation coefficient of adjective biases for each pair of embeddings over time. As with gender, the analysis shows how external events changed attitudes. There are two phase shifts in the correlation: in the 1960s, which coincide with a sharp increase in Asian immigration into the United States due to the passage of the 1965 Immigration and Nationality Act, and in the 1980s, when immigration continued and the second-generation Asian-American population emerged (39). Using the same Kolmogorov–Smirnov test on the correlation differences described in the previous section, the phase shifts between the 1950s–1960s ($P = 0.011$) and 1970s–1980s ($P < 10^{-3}$) are significant, while the rest are not ($P > 0.070$).

We extract the most biased adjectives toward Asians (when compared with Whites) to gain more insights into factors driving these global changes in the embedding. Table 3 shows the most Asian-biased adjectives in 1910, 1950, and 1990. Before 1950, strongly negative words, especially those often used to describe outsiders, are among the words most associated with Asians: barbaric, hateful, monstrous, bizarre, and cruel. However, starting around 1950 and especially by 1980, with a rising Asian population in the United States, these words are largely replaced by words often considered stereotypic (40–42) of Asian Americans today: sensitive, passive, complacent, active, and hearty, for example. See *SI Appendix*, Table C.8 for the complete list of the top 10 most Asian-associated words in each decade.

Using our methods regarding trends, we can quantify this change more precisely: Fig. 6 shows the relative strength of the Asian association for words used to describe outsiders over time. As opposed to the adjectives overall, which see two distinct phase shifts in Asian association, the words related to outsiders steadily decrease in Asian association over time—except around World War II—indicating that broader globalization trends led to changing attitudes with regard to such negative portrayals. Overall, the word embeddings exhibit a remarkable change in adjectives and attitudes toward Asian Americans during the 20th century.

Trends in Other Ethnic and Cultural Stereotypes. Similar trends appear in other datasets as well. Fig. 7 shows, in the *New York Times* over two decades, how words related to Islam (vs. those related to Christianity) associate with terrorism-related words. Similar to how we measure occupation-related bias, we create a list of words associated with terrorism, such as terror, bomb, and violence. We then measure how associated these words appear to be in the text to words representing each religion, such as mosque and church, for Islam and Christianity, respectively. (Full word lists are available in *SI Appendix*, section A.) Throughout the time period in the *New York Times*, Islam is more associated with terrorism than is Christianity. Furthermore, an increase in the association can be seen both after the 1993 World Trade Center bombings and after September 11, 2001. With a more recent dataset and using more news outlets, it would be useful to study how such attitudes have evolved since 2005.

We illustrate how word embeddings capture stereotypes toward other ethnic groups. For example, *SI Appendix*, Fig. C.4, with Russian names, shows a dramatic shift in the 1950s, the start of the Cold War, and a minor shift during the initial years of the Russian Revolution in the 1910s–1920s. Furthermore, *SI Appendix*, Fig. C.5, the correlation over time plot with Hispanic names, serves as an effective control group. It shows more steady changes in the embeddings rather than the sharp transitions found in Asian and Russian associations. This pattern is consistent with the fact that numerous events throughout the 20th century influenced the story of Hispanic immigration into the United States, with no single event playing too large a role (43).

These patterns demonstrate the usefulness of our methods to study ethnic as well as gender bias over time; similar analyses can be performed to examine shifts in the attitudes toward other ethnic groups, especially around significant global events. In particular, it would be interesting to more closely measure dehumanization and "othering" of immigrants and other groups using a suite of linguistic techniques, validating and extending the patterns discovered in this work.

Discussion

In this work, we investigate how the geometry of word embeddings, with respect to gender and ethnic stereotypes, evolves over time and tracks with empirical demographic changes in the United States. We apply our methods to analyze word embeddings trained over 100 y of text data. In particular, we quantify the embedding biases for occupations and adjectives. Using occupations allows us to validate the method when the embedding associations are compared with empirical participation rates for each occupation. We show that both gender and ethnic occupation biases in the embeddings significantly track with the actual occupation participation rates. We also show that adjective associations in the embeddings provide insight into how different groups of people are viewed over time.

As in any empirical work, the robustness of our results depends on the data sources and the metrics we choose to represent bias or association. We choose the relative norm difference metric for its simplicity, although many such metrics are reasonable. Refs. 20 and 21 leverage alternate metrics, for example. Our metric agrees with other possible metrics—both

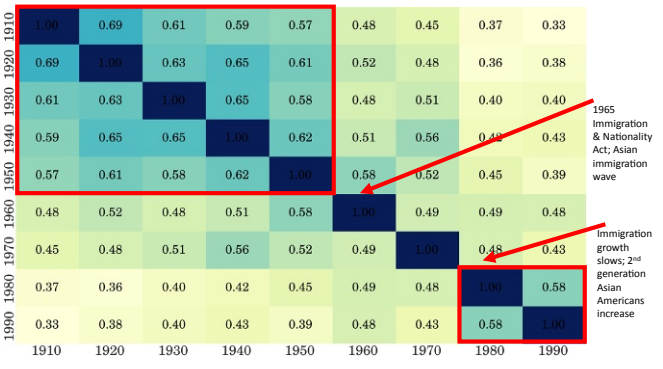


Fig. 5. Pearson correlation in embedding Asian bias scores for adjectives over time between embeddings for each decade.

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

qualitatively through the results in the snapshot analysis for gender, which replicates prior work, and quantitatively as the metrics correlate highly with one another, as shown in *SI Appendix*, section A.5.

Furthermore, we primarily use linear models to fit the relationship between embedding bias and various external metrics; however, the true relationships may be nonlinear and warrant further study. This concern is especially salient when studying ethnic stereotypes over time in the United States, as immigration drastically shifts the size of each group as a percentage of the population, which may interact with stereotypes and occupation percentages. However, the models are sufficient to show consistency in the relationships between embedding bias and external metrics across datasets over time. Further, the results do not qualitatively change when, for example, population logit proportion instead of raw percentage difference is used, as in ref. 44; we reproduce our primary figures with such a transformation in *SI Appendix*, section A.6.

Another potential concern may be the dependency of our results on the specific word lists used and that the recall of our methods in capturing human biases may not be adequate. We take extensive care to reproduce similar results with other word lists and types of measurements to demonstrate recall. For example, in *SI Appendix*, section B.1, we repeat the static occupation analysis using only professional occupations and reproduce an identical figure to Fig. 1 in *SI Appendix*, section B.1. Furthermore, the plots themselves contain bootstrapped confidence intervals; i.e., the coefficients for random subsets of the occupations/adjectives and the intervals are tight. Similarly, for adjectives, we use two different lists: one list from refs. 6 and 7 for which we have labeled stereotype scores and then a larger one for the rest of the analysis where such scores are not needed. We note that we do not tune either the embeddings or the word lists, instead opting for the largest/most general publicly available data. For reproducibility, we share our code and all word lists in a repository. That our methods replicate across many different embeddings and types of biases measured suggests their generalizability.

A common challenge in historical analysis is that the written text in, say 1910, may not completely reflect the popular social attitude of that time. This is an important caveat to consider in interpreting the results of the embeddings trained on these earlier text corpora. The fact that the embedding bias for gender and ethnic groups does track with census proportion is a positive control that the embedding is still capturing meaningful patterns despite possible limitations in the training text. Even this control may be limited in that the census proportion does not fully capture gender or ethnic associations, even in the present day. However, the written text does serve as a window into the attitudes of the day as expressed in popular culture, and this work allows for a more systematic study of such text.

Another limitation of our current approach is that all of the embeddings used are fully "black box," where the dimensions have no inherent meaning. To provide a more causal explanation of how the stereotypes appear in language, and to understand how they function, future work can leverage more recent embedding models in which certain dimensions are designed to capture various aspects of language, such as the polarity of a word or its parts of speech (45). Similarly, structural properties of words—beyond their census information or human-rated stereotypes—can be studied in the context of these dimensions. One can also leverage recent Bayesian embeddings models and train more fine-grained embeddings over time, rather than a separate embedding per decade as done in this work (46, 47). These approaches can be used in future work.

We view the main contribution of our work as introducing and validating a framework for exploring the temporal dynamics of stereotypes through the lens of word embeddings. Our framework enables the computation of simple but quantitative measures of bias as well as easy visualizations. It is important to note that our goal in *Quantifying Gender Stereotypes* and *Quantifying Ethnic Stereotypes* is quantitative exploratory analysis rather than pinning down specific causal models of how certain stereotypes arise or develop, although the analysis in *Occupational Stereotypes Beyond Census Data* suggests that common language is more biased than one would expect based on external, objective metrics. We believe our approach sharpens the analysis of large cultural shifts in US history; e.g., the women's movement of the 1960s correlates with a sharp shift in the encoding matrix (Fig. 4) as well as changes in the biases associated with specific occupations and gender-biased adjectives (e.g., hysterical vs. emotional).

In standard quantitative social science, machine learning is used as a tool to analyze data. Our work shows how the artifacts of machine learning (word embeddings here) can themselves be interesting objects of sociological analysis. We believe this paradigm shift can lead to many fruitful studies.

Materials and Methods

In this section we describe the datasets, embeddings, and word lists used, as well as how bias is quantified. More detail, including descriptions of additional embeddings and the full word lists, are in *SI Appendix*, section A. All of our data and code are available on GitHub (https://github.com/nikhgarg/EmbeddingDynamicStereotypes), and we link to external data sources as appropriate.

Embeddings. This work uses several pretrained word embeddings publicly available online; refer to the respective sources for in-depth discussion of their training parameters. These embeddings are among the most commonly used English embeddings, vary in the datasets on which they were

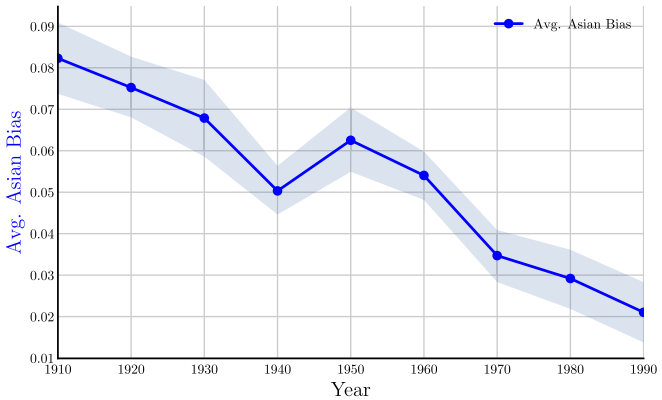


Fig. 6. Asian bias score over time for words related to outsiders in COHA data. The shaded region is the bootstrap SE interval.

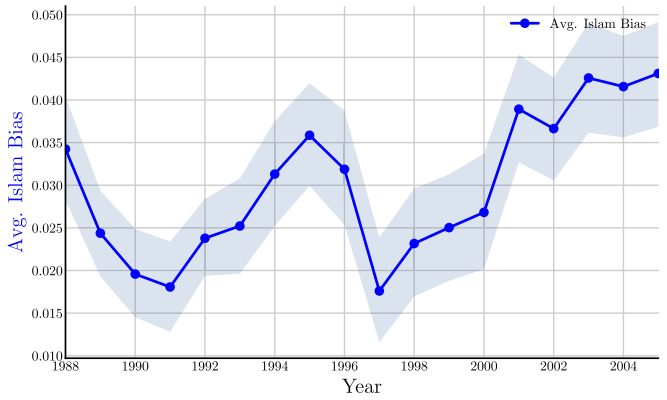


Fig. 7. Religious (Islam vs. Christianity) bias score over time for words related to terrorism in *New York Times* data. Note that embeddings are trained in 3-y windows, so, for example, 2000 contains data from 1999–2001. The shaded region is the bootstrap SE interval.

trained, and between them cover the best-known algorithms to construct embeddings. One finding in this work is that, although there is some heterogeneity, gender and ethnic bias is generally consistent across embeddings. Here we restrict descriptions to embeddings used in the main exposition. For consistency, only single words are used, all vectors are normalized by their l_2 norm, and words are converted to lowercase.

Google News word2vec vectors. Vectors trained on about 100 billion words in the Google News dataset (24, 25). Vectors are available at https://code.google.com/archive/p/word2vec/.

Google Books/COHA. Vectors trained on a combined corpus of genre-balanced Google Books and the COHA (48) by the authors of ref. 26. For each decade, a separate embedding is trained from the corpus data corresponding to that decade. The dataset is specifically designed to enable comparisons across decades, and the creators take special care to avoid selection bias issues. The vectors are available at https://nlp.stanford.edu/projects/histwords/, and we limit our analysis to the SVD and skip-gram with negative sampling (SGNS) (also known as word2vec) embeddings in the 1900s. Note that the Google Books data may include some non-American sources and the external metrics we use are American. However, this does not appreciably affect results. In the main text, we exclusively use SGNS embeddings; results with SVD embeddings are in *SI Appendix* and are qualitatively similar to the SGNS results. Unless otherwise specified, COHA indicates these embeddings trained using the SGNS algorithm.

New York Times. We train embeddings over time from *The New York Times* Annotated Corpus (28), using 1.8 million articles from the *New York Times* between 1988 and 2005. We use the GLoVe algorithm (27) and train embeddings over 3-y windows (so the 2000 embeddings, for example, contain articles from 1999 to 2001).

In *SI Appendix* we also use other embeddings available at https://nlp.stanford.edu/projects/glove/.

Related Work. Word embedding was developed as a framework to represent words as a part of the artificial intelligence and natural language processing pipeline (25). Ref. 20 demonstrated that word embeddings capture gender stereotypes, and ref. 21 additionally verified that the embedding accurately reflects human biases by comparing the embedding results with that of the implicit association test. While these two papers analyzed the bias of the static Google News embedding, our paper investigates the temporal changes in word embeddings and studies how embeddings over time capture historical trends. Our paper also studies attitudes toward women and ethnic minorities by quantifying the embedding of adjectives. The focus of ref. 20 is to develop algorithms to reduce the gender stereotype in the embedding, which is important for sensitive applications of embeddings. In contrast, our aim is not to debias, but to leverage the embedding bias to study historical changes that are otherwise challenging to quantify. Ref. 21 shows that embeddings contain each of the associations commonly found in the implicit association test. For example, European-American names are more similar to pleasant (vs. unpleasant) words than are African-American names, and male names are more similar to career (vs. family) words than are female names. Similarly, they show that, in the Google News embeddings, census data correspond to bias in the embeddings for gender.

The study of gender and ethnic stereotypes is a large focus of linguistics and sociology and is too extensive to be surveyed here (1–5). Our main innovation is the use of word embeddings, which provides a unique lens to measure and quantify biases. Another related field in linguistics studies how language changes over time and has also recently used word embeddings as a tool (49–51). However, this literature primarily studies semantic changes, such as how the word gay used to primarily mean cheerful and now means predominantly means homosexual (26, 52), and does not investigate bias.

Word Lists and External Metrics. Two types of word lists are used in this work: group words and neutral words. Group words represent groups of people, such as each gender and ethnicity. Neutral words are those that are not intrinsically gendered or ethnic (for example, fireman or mailman would be gendered occupation titles and so are excluded); relative similarities between neutral words and a pair of groups (such as men vs. women) are used to measure the strength of the association in the embeddings. In this work, we use occupations and various adjective lists as neutral words.

Gender. For gender, we use noun and pronoun pairs (such as he/she, him/her, etc.).

Race/ethnicity. To distinguish various ethnicities, we leverage the fact that the distribution of last names in the United States differs significantly by ethnicity, with the notable exception of White and Black last names. Starting with a breakdown of ethnicity by last name compiled by ref. 53, we identify 20 last names for each ethnicity as detailed in *SI Appendix*, section A.2. Our procedure, however, produces almost identical lists for White and Black Americans (with the names being mostly White by percentage), and so the analysis does not include Black Americans.

Occupation census data. We use occupation words for which we have gender and ethnic subgroup information over time. Group occupation percentages are obtained from the Integrated Public Use Microdata Series (IPUMS), part of the University of Minnesota Historical Census Project (29). Data coding and preprocessing are done as described in ref. 44, which studies wage dynamics as women enter certain occupations over time. The IPUMS dataset includes a column, OCC1950, coding occupation census data as it would have been coded in 1950, allowing accurate interyear analysis. We then hand map the occupations from this column to single-word occupations (e.g., chemical engineer and electrical engineer both become engineer, and chemist is counted as both chemist and scientist) and hand code a subset of the occupations as professional. In all plots containing occupation percentages for gender, we use the percentage difference between women and men in the occupation:

$$p_{\text{women}} - p_{\text{men}}$$
$$\text{where } p_{\text{women}} = \%\text{ of occupation that is women}$$
$$p_{\text{men}} = \%\text{ of occupation that is men.}$$

For ethnicity, we similarly report the percentage difference, except we first condition on the workers being in one of the groups in question:

$$\frac{p_{\text{min}} - p_{\text{white}}}{p_{\text{min}} + p_{\text{white}}}$$
$$\text{where } p_{\text{min}} = \%\text{ of occupation that is minority group in question}$$
$$p_{\text{white}} = \%\text{ of occupation that is White.}$$

In each case, a value of 0 indicates an equal number of each group in the occupation. We note that the results do not qualitatively change if instead the logit proportion (or conditional logit proportion) of the minority group is used, as in ref. 44 (*SI Appendix*, section A.6).

Occupation gender stereotypes. For a limited set of occupations, we use gender stereotype scores collected from users on Amazon Mechanical Turk by ref. 20. These scores are compared with embedding gender association.

Adjectives. To study associations with adjectives over time, several separate lists are used. To compare gender adjective embedding bias to external metrics, we leverage a list of adjectives labeled by how stereotypically associated with men or women they are, as determined by a group of subjects in 1977 and 1990 (6, 7). For Chinese adjective embedding bias, we use a list of stereotypes from the Princeton trilogy (8–10). For all other analyses using adjectives, a larger list of adjectives is used, primarily from ref. 54. Except when otherwise specified, adjectives are used to refer to this larger list.

Metrics. Given two vectors, their similarity can be measured either by their negative difference norm, as in Eq. **1**, or by their cosine similarity, as in Eq. **2**. The denominators are omitted because all vectors have norm 1:

$$\text{neg-norm-dif}(u, v) = -\|u - v\|_2 \quad [1]$$

$$\text{cos-sim}(u, v) = u \cdot v. \quad [2]$$

The association between the group words and neutral words is calculated as follows: Construct a group vector by averaging the vectors for each word in the group; then calculate the similarity between this average vector and each word in the neutral list as above.

The relative norm distance, which captures the relative strength of association of a set of neutral words with respect to two groups, is as described in Eq. **3**, where M is the set of neutral word vectors, v_1 is the average vector for group one, and v_2 is the average vector for group two. The more positive (negative) that the relative norm distance is, the more associated the neutral words are toward group two (one). In this work, when we say that a word is biased toward a group with respect to another group, we specifically mean in the context of the relative norm distance. Bias score also refers to this metric:

$$\text{relative norm distance} = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2. \quad [3]$$

We can also use cosine similarity rather than the Euclidean 2-norm. *SI Appendix*, section A.5 shows that the choice of similarity measure is not important; the respective metrics using each similarity measure correlate highly with one another (Pearson coefficient > 0.95 in most cases). In the main text, we exclusively use the relative norm.

ACKNOWLEDGMENTS. J.Z. is supported by a Chan–Zuckerberg Biohub Investigator grant and National Science Foundation (NSF) Grant CRII 1657155. N.G. is supported by the NSF Graduate Research Fellowship under Grant DGE-114747.

1. Hamilton DL, Trolier TK (1986) *Stereotypes and Stereotyping: An Overview of the Cognitive Approach in Prejudice, Discrimination, and Racism* (Academic, San Diego), pp 127–163.
2. Basow SA (1992) *Gender: Stereotypes and Roles* (Thomson Brooks/Cole Publishing Co, Belmont, CA), 3rd Ed.
3. Wetherell M, Potter J (1992) *Mapping the Language of Racism: Discourse and the Legitimation of Exploitation* (Columbia Univ Press, New York).
4. Holmes J, Meyerhoff M, eds (2004) *The Handbook of Language and Gender* (Blackwell Publishing Ltd, Oxford).
5. Coates J (2016) *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language* (Routledge, London).
6. Williams JE, Best DL (1977) Sex stereotypes and trait favorability on the adjective check list. *Educ Psychol Meas* 37:101–110.
7. Williams JE, Best DL (1990) *Measuring Sex Stereotypes: A Multination Study* (Sage Publications, Thousand Oaks, CA), Rev Ed.
8. Katz D, Braly K (1933) Racial stereotypes of one hundred college students. *J Abnorm Soc Psychol* 28:280–290.
9. Gilbert GM (1951) Stereotype persistence and change among college students. *J Abnorm Soc Psychol* 46:245–254.
10. Karlins M, Coffman TL, Walters G (1969) On the fading of social stereotypes: Studies in three generations of college students. *J Pers Soc Psychol* 13:1–16.
11. Devine PG, Elliot AJ (1995) Are racial stereotypes really fading? The Princeton trilogy revisited. *Pers Soc Psychol Bull* 21:1139–1150.
12. Diekman AB, Eagly AH (2000) Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Pers Soc Psychol Bull* 26:1171–1188.
13. Bergsieker HB, Leslie LM, Constantine VS, Fiske ST (2012) Stereotyping by omission: Eliminate the negative, accentuate the positive. *J Pers Soc Psychol* 102:1214–1238.
14. Madon S, et al. (2001) Ethnic and national stereotypes: The Princeton trilogy revisited and revised. *Pers Soc Psychol Bull* 27:996–1010.
15. Gaertner SL, Dovidio JF (1986) *The Aversive Form of Racism* (Academic, San Diego).
16. Fiske ST, Cuddy AJC, Glick P, Xu J (2002) A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol* 82:878–902.
17. Henley NM (1989) Molehill or mountain? What we know and don't know about sex bias in language. *Gender and Thought: Psychological Perspectives*, eds Crawford M, Gentry M (Springer, New York), pp 59–78.
18. Hellinger M, Bußmann H eds (2001) *Gender Across Languages: The Linguistic Representation of Women and Men,* IMPACT: Studies in Language and Society (John Benjamins Publishing Company, Amsterdam), Vol 9.
19. Collobert R, et al. (2011) Natural language processing (almost) from scratch. *J Machine Learn Res* 12:2493–2537.
20. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems 29,* eds Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (Curran Associates, Inc, Barcelona), pp 4349–4357.
21. Caliskan A, Bryson JJ, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186.
22. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2017) Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, eds Palmer M, Hwa R (Association for Computational Linguistics, Copenhagen), pp 2979–2989.
23. van Miltenburg E (2016) Stereotyping and bias in the Flickr30k dataset. arXiv: 1605.06083.
24. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781.
25. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, eds Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Curran Associates, Inc, Lake Tahoe, NV), pp 3111–3119.
26. Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds Erk K, Smith NA (Association for Computational Linguistics, Berlin), Vol 1, pp 1489–1501.
27. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), eds Moschitti A, Pang B (Association for Computational Linguistics, Doha, Qatar), pp 1532–1543.
28. Sandhaus E (2008) *The New York Times Annotated Corpus* (Linguistic Data Consortium, Philadelphia).
29. Ruggles S, Genadek K, Goeken R, Grover J, Sobek M (2015) Integrated Public Use Microdata Series: Version 6.0. Available at doi.org/10.18128/D010.V6.0. Accessed August 16, 2017.
30. Osajima K (2005) Asian Americans as the model minority: An analysis of the popular press image in the 1960s and 1980s. *A Companion to Asian American Studies*, ed Ono KA (Blackwell Publishing Ltd, Malden, MA), pp 215–225.
31. Fong TP (2002) *The Contemporary Asian American Experience: Beyond the Model Minority* (Prentice Hall, Upper Saddle River, NJ).
32. Bryson V (2016) *Feminist Political Theory* (Palgrave Macmillan, New York).
33. Rosen R (2013) *The World Split Open: How the Modern Women's Movement Changed America* (Tantor eBooks, Old Saybrook, CT).
34. Thorne B, Henley N, Kramarae C, eds (1983) *Language, Gender, and Society* (Newbury House, Rowley, MA).
35. Eckert P, McConnell-Ginet S (2003) *Language and Gender* (Cambridge Univ Press, Cambridge, UK).
36. Evans S (2010) *Tidal Wave: How Women Changed America at Century's End* (Simon and Schuster, New York).
37. Tasca C, Rapetti M, Carta MG, Fadda B (2012) Women and hysteria in the history of mental health. *Clin Pract Epidemiol Ment Health* 8:110–119.
38. Sanghani R (2016) Feisty, frigid and frumpy: 25 Words we only use to describe women. The Telegraph. Available at https://www.telegraph.co.uk/women/life/ambitious-frigid-and-frumpy-25-words-we-only-use-to-describe-wom/. Accessed August 21, 2017.
39. Zong J, Batalova J (2016) *Asian Immigrants in the United States* (Migration Policy Institute, Washington, DC).
40. Lee SJ (1994) Behind the model-minority stereotype: Voices of high- and low-achieving Asian American students. *Anthropol Educ Q* 25:413–429.
41. Kim A, Yeh CJ (2002) *Stereotypes of Asian American students* (ERIC Digest New York, NY).
42. Lee SJ (2015) *Unraveling the "Model Minority" Stereotype: Listening to Asian American Youth* (Teachers College Press, New York), 2nd Ed.
43. Gutiérrez DG (2016) A historic overview of Latino immigration and the demographic transformation of the United States. *The New Latino Studies Reader: A Twenty-First-Century Perspective*, eds Gutierrez RA, Almaguer T (Univ of California Press, Oakland, CA), pp 108–125.
44. Levanon A, England P, Allison P (2009) Occupational feminization and pay: Assessing causal dynamics using 1950–2000 U.S. Census data. *Soc Forces* 88:865–891.
45. Rothe S, Schütze H (2016) Word embedding calculus in meaningful ultradense subspaces. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, eds Erk K, Smith NA (Association for Computational Linguistics, Berlin), Vol 2, pp 512–517.
46. Rudolph M, Ruiz F, Athey S, Blei D (2017) *Structured Embedding Models for Grouped Data in Advances in Neural Information Processing Systems 30*, eds Guyon I, et al. (Curran Associates, Inc, Long Beach, CA), pp 250–260.
47. Rudolph M, Blei D (2017) Dynamic Bernoulli embeddings for language evolution. arXiv:1703.08052.
48. Davies M (2010) The 400 million word corpus of historical American English (1810–2009). *Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23–27 August 2010*, eds Hegedűs I, Fodor A (John Benjamins Publishing, Amsterdam), Vol 325.
49. Ullmann S (1962) *Semantics: An Introduction to the Science of Meaning* (Barnes & Noble, New York).
50. Blank A (1999) *Why Do New Meanings Occur? A Cognitive Typology of the Motivations for Lexical Semantic Change in Historical Semantics and Cognition*, ed Koch P (Walter de Gruyter, New York).

51. Kulkarni V, Al-Rfou R, Perozzi B, Skiena S (2015) Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*, eds Gangemi A, Leonardi S, Panconesi A (International World Wide Web Conferences Steering Committee, New York), pp 625–635.
52. Hamilton WL, Leskovec J, Jurafsky D (2016) Cultural shift or linguistic drift? comparing two computational measures of semantic change. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, eds Su J, Duh K, Carreras X (Association for Computational Linguistics, Austin, TX), pp 2116–2121.
53. Chalabi M, Flowers A (2018) Dear Mona, what's the most common name in America? Available at https://fivethirtyeight.com/features/whats-the-most-common-name-in-america/. Accessed September 3, 2017.
54. Gunkel P (2013) 638 Primary personality traits. Available at ideonomy.mit.edu/essays/traits.html. Accessed August 21, 2017.

Appendix A Data

This section contains information on the additional embeddings used in this supplement and the various word lists used. These lists will also be provided in a public repository alongside code to reproduce our results.

A.1 Additional Embeddings

A.1.1 Wikipedia 2014+Gigaword 5 GloVe

Trained on Wikipedia data from 2014 and newswire data from the mid 1990s through 2011 [1] using GloVe [2]. Available online at `https://nlp.stanford.edu/projects/glove/`. We refer to these vectors as the Wikipedia vectors.

A.1.2 Common Crawl GloVe

Trained on Common Crawl using GloVe [2]. Also available online at `https://nlp.stanford.edu/projects/glove/`. We refer to these vectors as the Commmon Crawl vectors.

A.2 Group words

Man words *he, son, his, him, father, man, boy, himself, male, brother, sons, fathers, men, boys, males, brothers, uncle, uncles, nephew, nephews*

Woman words *she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, femen, sisters, aunt, aunts, niece, nieces*

White last names *harris, nelson, robinson, thompson, moore, wright, anderson, clark, jackson, taylor, scott, davis, allen, adams, lewis, williams, jones, wilson, martin, johnson*

Hispanic last names *ruiz, alvarez, vargas, castillo, gomez, soto, gonzalez, sanchez, rivera, mendoza, martinez, torres, rodriguez, perez, lopez, medina, diaz, garcia, castro, cruz*

Asian last names *cho, wong, tang, huang, chu, chung, ng, wu, liu, chen, lin, yang, kim, chang, shah, wang, li, khan, singh, hong*

Russian last names *gurin, minsky, sokolov, markov, maslow, novikoff, mishkin, smirnov, orloff, ivanov, sokoloff, davidoff, savin, romanoff, babinski, sorokin, levin, pavlov, rodin, agin*

Chinese last names *chung, liu, wong, huang, ng, hu, chu, chen, lin, liang, wang, wu, yang, tang, chang, hong, li*

Islam words *allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad*

Christianity words *baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church*

Starting with a breakdown of ethnicity by last name compiled by [3][1], we identify 20 last names for each Whites, Asians, and Hispanics as follows: 1) Start with list of top 50 last names by percent of that ethnicity, conditioned on being top 5000 surnames overall, as well as the top 50 last names by total number in that ethnicity (i.e., multiplied count of that last name by percent in that ethnicity). 2) Choose the 20 names that appeared most on average in the Google Books/COHA vectors over time (with a minimum number for each time period). This second step ensures that an accurate ethnicity vector is identified each time period, with minimal distortions. Russian last names are collated from various sources online.

[1] available `https://raw.githubusercontent.com/fivethirtyeight/data/master/most-common-name/surnames.csv`

A.3 Neutral Words

Occupations *janitor, statistician, midwife, bailiff, auctioneer, photographer, geologist, shoemaker, athlete, cashier, dancer, housekeeper, accountant, physicist, gardener, dentist, weaver, blacksmith, psychologist, supervisor, mathematician, surveyor, tailor, designer, economist, mechanic, laborer, postmaster, broker, chemist, librarian, attendant, clerical, musician, porter, scientist, carpenter, sailor, instructor, sheriff, pilot, inspector, mason, baker, administrator, architect, collector, operator, surgeon, driver, painter, conductor, nurse, cook, engineer, retired, sales, lawyer, clergy, physician, farmer, clerk, manager, guard, artist, smith, official, police, doctor, professor, student, judge, teacher, author, secretary, soldier*

Professional Occupations[2] *statistician, auctioneer, photographer, geologist, accountant, physicist, dentist, psychologist, supervisor, mathematician, designer, economist, postmaster, broker, chemist, librarian, scientist, instructor, pilot, administrator, architect, surgeon, nurse, engineer, lawyer, physician, manager, official, doctor, professor, student, judge, teacher, author*

Occupations with Human Stereotype Scores from [4] *teacher, author, mechanic, broker, baker, surveyor, laborer, surgeon, gardener, painter, dentist, janitor, athlete, manager, conductor, carpenter, housekeeper, secretary, economist, geologist, clerk, doctor, judge, physician, lawyer, artist, instructor, dancer, photographer, inspector, musician, soldier, librarian, professor, psychologist, nurse, sailor, accountant, architect, chemist, administrator, physicist, scientist, farmer*

Adjectives from [5, 6] *headstrong, thankless, tactful, distrustful, quarrelsome, effeminate, fickle, talkative, dependable, resentful, sarcastic, unassuming, changeable, resourceful, persevering, forgiving, assertive, individualistic, vindictive, sophisticated, deceitful, impulsive, sociable, methodical, idealistic, thrifty, outgoing, intolerant, autocratic, conceited, inventive, dreamy, appreciative, forgetful, forceful, submissive, pessimistic, versatile, adaptable, reflective, inhibited, outspoken, quitting, unselfish, immature, painstaking, leisurely, infantile, sly, praising, cynical, irresponsible, arrogant, obliging, unkind, wary, greedy, obnoxious, irritable, discreet, frivolous, cowardly, rebellious, adventurous, enterprising, unscrupulous, poised, moody, unfriendly, optimistic, disorderly, peaceable, considerate, humorous, worrying, preoccupied, trusting, mischievous, robust, superstitious, noisy, tolerant, realistic, masculine, witty, informal, prejudiced, reckless, jolly, courageous, meek, stubborn, aloof, sentimental, complaining, unaffected, cooperative, unstable, feminine, timid, retiring, relaxed, imaginative, shrewd, conscientious, industrious, hasty, commonplace, lazy, gloomy, thoughtful, dignified, wholesome, affectionate, aggressive, awkward, energetic, tough, shy, queer, careless, restless, cautious, polished, tense, suspicious, dissatisfied, ingenious, fearful, daring, persistent, demanding, impatient, contented, selfish, rude, spontaneous, conventional, cheerful, enthusiastic, modest, ambitious, alert, defensive, mature, coarse, charming, clever, shallow, deliberate, stern, emotional, rigid, mild, cruel, artistic, hurried, sympathetic, dull, civilized, loyal, withdrawn, confident, indifferent, conservative, foolish, moderate, handsome, helpful, gentle, dominant, hostile, generous, reliable, sincere, precise, calm, healthy, attractive, progressive, confused, rational, stable, bitter, sensitive, initiative, loud, thorough, logical, intelligent, steady, formal, complicated, cool, curious, reserved, silent, honest, quick, friendly, efficient, pleasant, severe, peculiar, quiet, weak, anxious, nervous, warm, slow, dependent, wise, organized, affected, reasonable, capable, active, independent, patient, practical, serious, understanding, cold, responsible, simple, original, strong, determined, natural, kind*

Larger adjective list, mostly from [7] *disorganized, devious, impressionable, circumspect, impassive, aimless, effeminate, unfathomable, fickle, unprincipled, inoffensive, reactive, providential, resentful, bizarre, impractical, sarcastic, misguided, imitative, pedantic, venomous, erratic, insecure, resourceful, neurotic, forgiving, profligate, whimsical, assertive, incorruptible, individualistic, faithless, disconcerting, barbaric, hypnotic, vindictive, observant, dissolute, frightening, complacent, boisterous, pretentious, disobedient, tasteless, sedentary, sophisticated, regimental, mellow, deceitful, impulsive, playful, sociable, methodical, willful, idealistic, boyish, callous, pompous, unchanging, crafty, punctual, compassionate, intolerant, challenging, scornful, possessive, conceited, imprudent, dutiful, lovable, disloyal, dreamy, appreciative, forgetful, unrestrained, forceful, submissive, predatory, fanatical, illogical, tidy, aspiring, studious, adaptable, conciliatory, artful, thoughtless, deceptive, frugal, reflective, insulting, unreliable, stoic, hysterical, rustic, inhibited, outspoken, unhealthy, ascetic, skeptical, painstaking, contemplative, leisurely, sly, mannered, outrageous, lyrical, placid, cynical, irresponsible,*

[2]These were hand-coded from the overall list of occupations; follow-on work should study this more systematically.

vulnerable, arrogant, persuasive, perverse, steadfast, crisp, envious, naive, greedy, presumptuous, obnoxious, irritable, dishonest, discreet, sporting, hateful, ungrateful, frivolous, reactionary, skillful, cowardly, sordid, adventurous, dogmatic, intuitive, bland, indulgent, discontented, dominating, articulate, fanciful, discouraging, treacherous, repressed, moody, sensual, unfriendly, optimistic, clumsy, contemptible, focused, haughty, morbid, disorderly, considerate, humorous, preoccupied, airy, impersonal, cultured, trusting, respectful, scrupulous, scholarly, superstitious, tolerant, realistic, malicious, irrational, sane, colorless, masculine, witty, inert, prejudiced, fraudulent, blunt, childish, brittle, disciplined, responsive, courageous, bewildered, courteous, stubborn, aloof, sentimental, athletic, extravagant, brutal, manly, cooperative, unstable, youthful, timid, amiable, retiring, fiery, confidential, relaxed, imaginative, mystical, shrewd, conscientious, monstrous, grim, questioning, lazy, dynamic, gloomy, troublesome, abrupt, eloquent, dignified, hearty, gallant, benevolent, maternal, paternal, patriotic, aggressive, competitive, elegant, flexible, gracious, energetic, tough, contradictory, shy, careless, cautious, polished, sage, tense, caring, suspicious, sober, neat, transparent, disturbing, passionate, obedient, crazy, restrained, fearful, daring, prudent, demanding, impatient, cerebral, calculating, amusing, honorable, casual, sharing, selfish, ruined, spontaneous, admirable, conventional, cheerful, solitary, upright, stiff, enthusiastic, petty, dirty, subjective, heroic, stupid, modest, impressive, orderly, ambitious, protective, silly, alert, destructive, exciting, crude, ridiculous, subtle, mature, creative, coarse, passive, oppressed, accessible, charming, clever, decent, miserable, superficial, shallow, stern, winning, balanced, emotional, rigid, invisible, desperate, cruel, romantic, agreeable, hurried, sympathetic, solemn, systematic, vague, peaceful, humble, dull, expedient, loyal, decisive, arbitrary, earnest, confident, conservative, foolish, moderate, helpful, delicate, gentle, dedicated, hostile, generous, reliable, dramatic, precise, calm, healthy, attractive, artificial, progressive, odd, confused, rational, brilliant, intense, genuine, mistaken, driving, stable, objective, sensitive, neutral, strict, angry, profound, smooth, ignorant, thorough, logical, intelligent, extraordinary, experimental, steady, formal, faithful, curious, reserved, honest, busy, educated, liberal, friendly, efficient, sweet, surprising, mechanical, clean, critical, criminal, soft, proud, quiet, weak, anxious, solid, complex, grand, warm, slow, false, extreme, narrow, dependent, wise, organized, pure, directed, dry, obvious, popular, capable, secure, active, independent, ordinary, fixed, practical, serious, fair, understanding, constant, cold, responsible, deep, religious, private, simple, physical, original, working, strong, modern, determined, open, political, difficult, knowledge, kind

Competence Adjectives[3] *precocious, resourceful, inquisitive, sagacious, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, luminous, venerable, imaginative, shrewd, thoughtful, sage, smart, ingenious, clever, brilliant, logical, intelligent, apt, genius, wise*

Physical Appearance Adjectives[4] *alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong*

Terrorism related words *terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target, conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold, wreckage, aggression, slaughter, execute, overthrow, casualties, massacre, retaliation, proliferation, militia, hostility, debris, acid, execution, militant, rocket, guerrilla, sacrifice, enemy, soldier, terrorist, missile, hostile, revolution, resistance, shoot*

Outsider Adjectives *devious, bizarre, venomous, erratic, barbaric, frightening, deceitful, forceful, deceptive, envious, greedy, hateful, contemptible, brutal, monstrous, calculating, cruel, intolerant, aggressive, monstrous*

Princeton Trilogy Stereotypes from [8, 9, 10] For 1 analysis, all scores in tables used (where enough embedding data exists). For the second, only those for which scores over time were available were used.

All static scores in tables *courteous, deceitful, meditative, artistic, conservative, reserved, nationalistic, loyal, ignorant, superstitious, quiet, sly, religious, traditional, industrious*

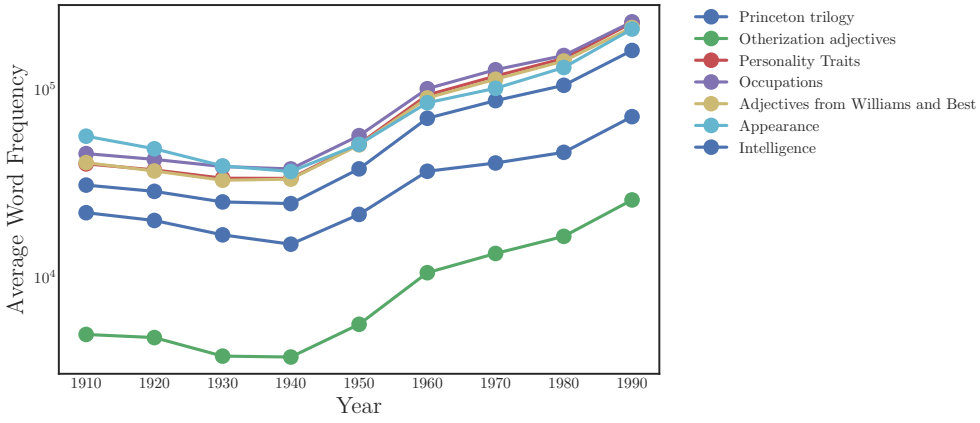
Scores over time *deceitful, meditative, conservative, reserved, loyal, ignorant, superstitious, quiet, sly, religious, traditional, industrious*

[3] mostly from `https://www.e-education.psu.edu/writingrecommendationlettersonline/node/151`,`https://www.macmillandictionary.com/us/thesaurus-category/american/words-used-to-describe-intelligent-or-wise-people`

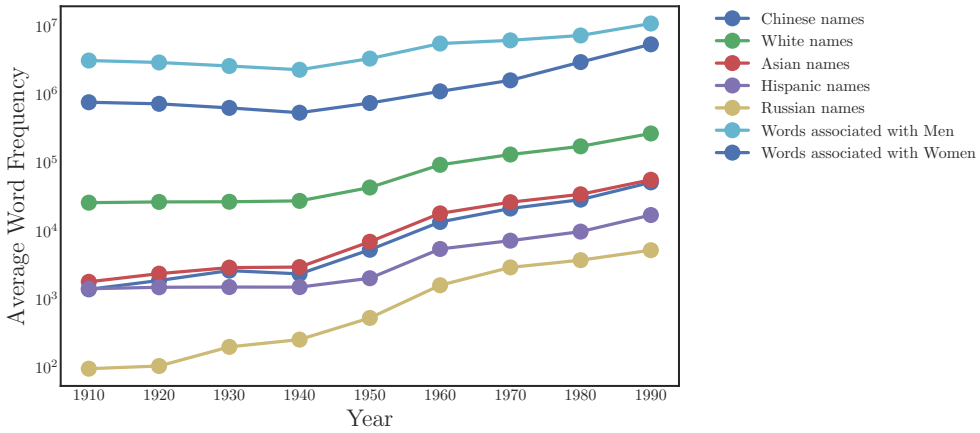
A.4 Embedding Quality

Here we show that the quality of the average vector constructed for each group does not appreciably change over time and thus cannot explain the overtime trends.

Figure A.1 shows the average frequency of words in each list across time in the COHA embeddings. Counts in general increase for all lists throughout time as datasets get larger. However, vector quality remains about the same; Figure A.2 shows the average variance on each dimension across time for each group. Except for Russian names (which are used for a plot in the appendix), these variances remain relatively steady throughout time, with small potential decreases as the size of the training datasets increase, indicating that average vector quality does not change appreciably.



(a) Neutral words



(b) Group words

Figure A.1: Average counts of words in a list in the COHA embeddings over time

[4]mostly from http://usefulenglish.ru/vocabulary/appearance-and-character, http://www.sightwordsgame.com/parts-of-speech/adjectives/appearance/, http://www.stgeorges.co.uk/blog/physical-appearance-adjectives-the-bald-and-the-beautiful

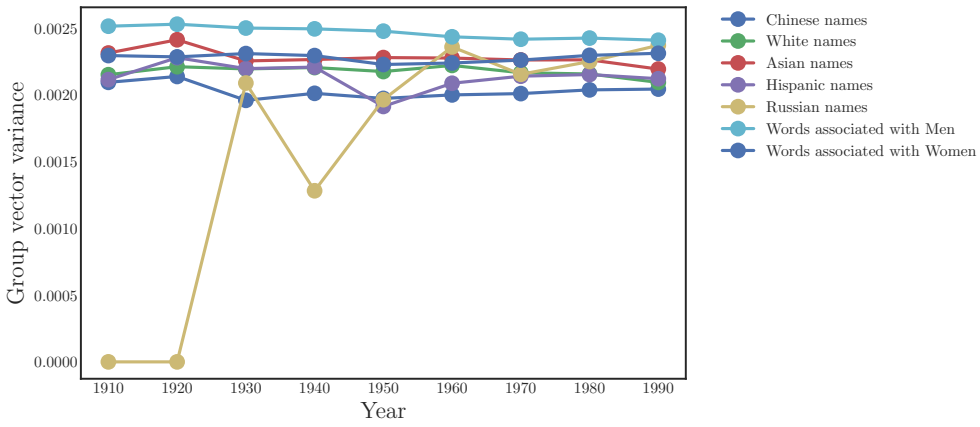


Figure A.2: Average variance in each embedding dimension over time for each group in the COHA embeddings

A.5 Similarity metrics

In the main exposition, we use the relative norm bias metric, $\sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2$. Using relative cosine similarity instead, $\sum_{v_m \in M} v_m \cdot v_2 - v_m \cdot v_1$, makes no difference – the metrics have Pearson correlation $> .95$ in general (note that Pearson correlation of 1 is a perfectly linear relationship). The following table shows their correlation for a few embedding/neutral word/group combinations, and the pattern holds generally.

Table A.1: Pearson correlation of two possible bias metrics

Embedding	Neutral Words	Groups	Pearson correlation
Google News	Occupations	Men, Women	.998
Google News	Personality Traits	Men, Women	.998
SGNS 1990	Occupations	Men, Women	.997
SGNS 1990	Personality Traits	Men, Women	.994
Google News	Occupations	Whites, Asians	.973
Google News	Personality Traits	Whites, Asians	.993
SGNS 1990	Occupations	Whites, Asians	.991
SGNS 1990	Personality Traits	Whites, Asians	.999

We thus primarily report results using only the relative norm difference bias metric.

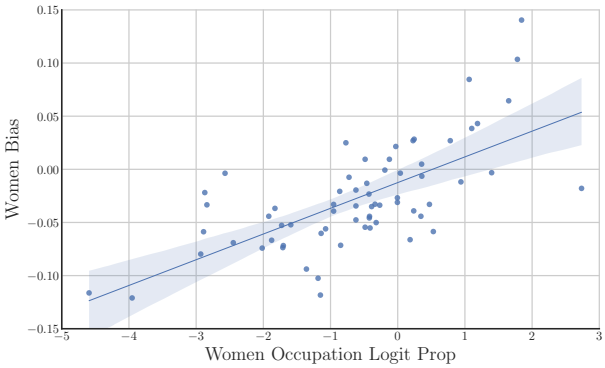
A.6 Occupation percentage transformation

In the main exposition, we use the relative percent difference of participation statistics in an occupation as a representation of the census data. [11], whose preprocessing steps we follow for the rest of the census data, instead uses the following transformation (for gender) occupation statistics:

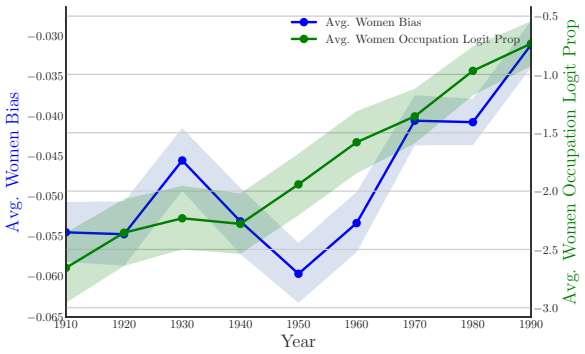
$$\text{logit-prop}(p) = \log \frac{p}{1-p} \tag{1}$$
$$\text{where } p = \%\text{ of women in occupation}$$

For ethnicity, an analogous metric, the conditional logit-proportion in Equation (2) can be used.

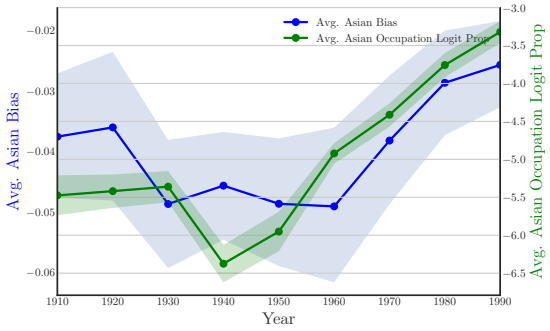
$$\text{cond-logit-prop}(\text{group 1}, \text{group 2}) = \log \frac{p}{1-p} \tag{2}$$
$$\text{where } p = \frac{\%\text{ of group 1}}{\%\text{ of group 1} + \%\text{ of group 2}}$$



(a) Analogue of Figure 1.



(b) Analogue of Figure 2.



(c) Analogue of Figure 3.

Figure A.3: Analogue of occupation verification figures using logit-proportion for occupation data.

The results do not qualitatively change when these transformations are used instead, as our (linear) models already are rough estimates for any true relationship. For brevity, we reproduce just the relevant figures from the main exposition using these logit proportion metrics. Statistical tests and equivalent plots from the rest of this appendix can also be reproduced, and results are qualitatively similar.

Appendix B Gender analysis

B.1 Additional Validation Analysis

B.1.1 Occupations

Table B.1 shows the top occupations and adjectives by gender in the Google News embedding.

Below, we first show the regression table corresponding to Figure 1, and then show the same plot for other embeddings, as well as the subset of occupations corresponding to *professional* occupations.

Occupations		Adjectives	
Man	Woman	Man	Woman
carpenter	nurse	honorable	maternal
mechanic	midwife	ascetic	romantic
mason	librarian	amiable	submissive
blacksmith	housekeeper	dissolute	hysterical
retired	dancer	arrogant	elegant
architect	teacher	erratic	caring
engineer	cashier	heroic	delicate
mathematician	student	boyish	superficial
shoemaker	designer	fanatical	neurotic
physicist	weaver	aimless	attractive

Table B.1: Top occupations and adjectives by gender in the Google News embedding.

Dep. Variable:	Women Bias	**R-squared:**	0.499
Model:	OLS	**Adj. R-squared:**	0.491
Method:	Least Squares	**F-statistic:**	63.65
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	3.53e-11
Time:	19:59:00	**Log-Likelihood:**	128.37
No. Observations:	66	**AIC:**	-252.7
Df Residuals:	64	**BIC:**	-248.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. $%$ Difference	0.0007	9.13e-05	7.978	0.000	0.001	0.001
const	-0.0118	0.005	-2.487	0.015	-0.021	-0.002

Omnibus:	1.265	**Durbin-Watson:**	1.652
Prob(Omnibus):	0.531	**Jarque-Bera (JB):**	0.829
Skew:	0.268	**Prob(JB):**	0.661
Kurtosis:	3.119	**Cond. No.**	56.7

Table B.2: OLS Regression Results corresponding to Figure 1, regressing women occupation percent difference with word vector relative distance in Google News vectors in 2015.

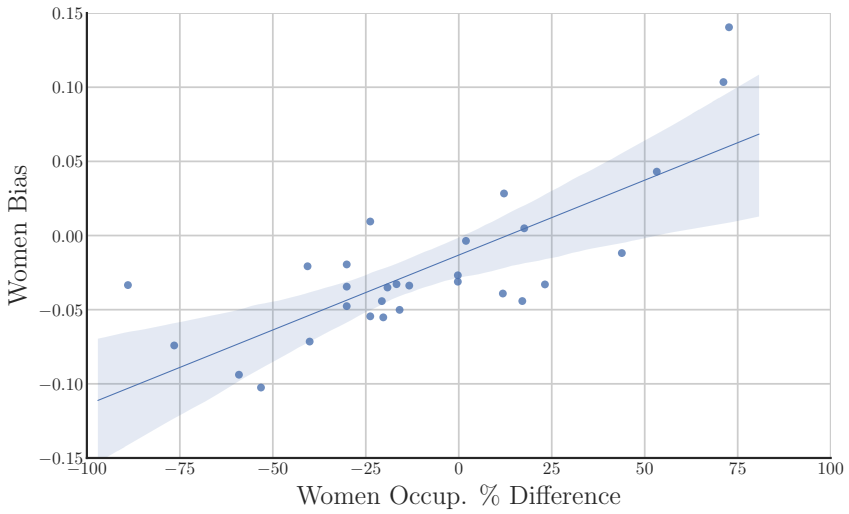


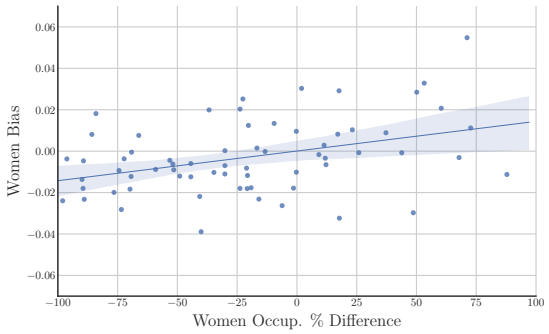
Figure B.1: Percent woman in an occupation vs relative norm distance from *professional* occupations in Google News vectors. $p < 10^{-5}$, r-squared$= .595$. Regression bias coefficient confidence interval: $(-.026, 0)$. Note that there is little difference in model from all occupations.

Dep. Variable:	Women Bias	**R-squared:**	0.595
Model:	OLS	**Adj. R-squared:**	0.580
Method:	Least Squares	**F-statistic:**	39.72
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	9.57e-07
Time:	19:59:11	**Log-Likelihood:**	58.614
No. Observations:	29	**AIC:**	-113.2
Df Residuals:	27	**BIC:**	-110.5
Df Model:	1		

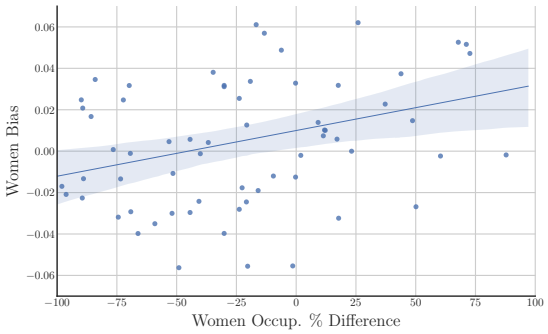
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0010	0.000	6.303	0.000	0.001	0.001
const	-0.0132	0.006	-2.072	0.048	-0.026	-0.000

Omnibus:	3.544	**Durbin-Watson:**	1.514
Prob(Omnibus):	0.170	**Jarque-Bera (JB):**	2.644
Skew:	0.739	**Prob(JB):**	0.267
Kurtosis:	3.034	**Cond. No.**	40.9

Table B.3: OLS Regression Results corresponding to Figure B.1, regressing women occupation percent difference with word vector relative distance in Google News vectors in 2015 for *professional* occupations.



(a) Common Crawl GloVe. $p = .001$, r-squared$= .149$.



(b) Wikipedia GloVe. $p = .008$, r-squared$= .105$.

Figure B.2: Percent woman in an occupation vs relative norm distance from occupations to the respective gender words in Common Crawl and Wikipedia.

Dep. Variable:	Women Bias	**R-squared:**	0.149
Model:	OLS	**Adj. R-squared:**	0.136
Method:	Least Squares	**F-statistic:**	11.19
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.00138
Time:	19:59:16	**Log-Likelihood:**	178.31
No. Observations:	66	**AIC:**	-352.6
Df Residuals:	64	**BIC:**	-348.2
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0001	4.28e-05	3.345	0.001	5.77e-05	0.000
const	6.248e-05	0.002	0.028	0.978	-0.004	0.004

Omnibus:	1.013	**Durbin-Watson:**	2.405
Prob(Omnibus):	0.603	**Jarque-Bera (JB):**	0.578
Skew:	0.217	**Prob(JB):**	0.749
Kurtosis:	3.148	**Cond. No.**	56.7

Table B.4: OLS Regression Results corresponding to Figure B.2a, regressing women occupation percent difference with word vector relative distance in Common Crawl vectors.

Dep. Variable:	Women Bias	**R-squared:**	0.105
Model:	OLS	**Adj. R-squared:**	0.091
Method:	Least Squares	**F-statistic:**	7.546
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.00780
Time:	19:59:22	**Log-Likelihood:**	136.80
No. Observations:	66	**AIC:**	-269.6
Df Residuals:	64	**BIC:**	-265.2
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	8.03e-05	2.747	0.008	6.02e-05	0.000
const	0.0100	0.004	2.395	0.020	0.002	0.018

Omnibus:	1.316	**Durbin-Watson:**	2.481
Prob(Omnibus):	0.518	**Jarque-Bera (JB):**	1.073
Skew:	0.057	**Prob(JB):**	0.585
Kurtosis:	2.386	**Cond. No.**	56.7

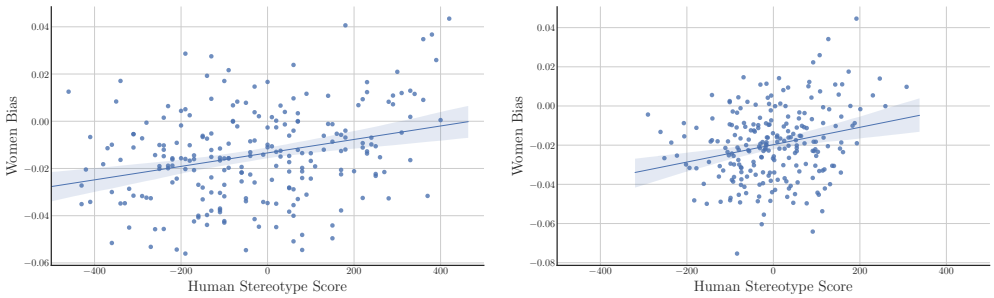
Table B.5: OLS Regression Results corresponding to Figure B.2b, regressing women occupation percent difference with word vector relative distance in Wikipedia vectors.

Table B.6: Most Man and Woman *residuals* when regressing embedding bias vs occupation gender proportion. The more Woman (Man) a residual the more biased the embedding is toward Women (Men) above what occupation proportions would suggest.

Man	Woman
secretary	nurse
mechanic	housekeeper
musician	gardener
architect	clerk
janitor	librarian
carpenter	sailor
broker	judge
geologist	artist
accountant	dancer
economist	painter

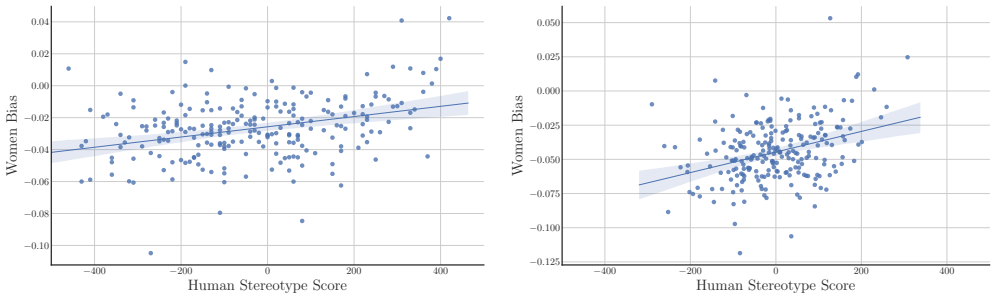
B.1.2 Adjectives

This subsection contains the supplementary material related to adjectives. In particular, we show the scatter plots and fit information corresponding the stereotype scores from [5] and [6] to the SGNS and SVD embeddings from the respective decades.



(a) 1990 SGNS vectors woman bias vs subjective scores from [6]. $p < 10^{-5}$, r-squared= .086.

(b) 1970 SGNS vectors woman bias vs subjective scores from [5]. $p < .0002$, r-squared= .095.



(c) 1990 SVD vectors woman bias vs subjective scores from [6]. $p < 10^{-6}$, r-squared= .116.

(d) 1970 SVD vectors woman bias vs subjective scores from [5]. $p < 10^{-7}$, r-squared= .127.

Figure B.3: Human stereotype score vs relative norm distance from adjectives to the respective gender words in 1970 and 1990, using both SGNS and SVD vectors

Note the stereotype scores used in the plots are linear transformations of the scores reported in the original papers, done to standardize figures. Given a score r from 1990 [6], the transformed score $r' = 500 - 10r$. Given a score from 1970 [5], the transformed score $r' = 500 - r$.

Dep. Variable:	Women Bias	**R-squared:**	0.095
Model:	OLS	**Adj. R-squared:**	0.087
Method:	Least Squares	**F-statistic:**	11.87
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.25e-05
Time:	19:59:27	**Log-Likelihood:**	592.34
No. Observations:	230	**AIC:**	-1179.
Df Residuals:	227	**BIC:**	-1168.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	2.842e-05	6.16e-06	4.612	0.000	1.63e-05	4.06e-05
counts	4.005e-09	2.63e-09	1.525	0.129	-1.17e-09	9.18e-09
const	-0.0143	0.001	-10.492	0.000	-0.017	-0.012

Omnibus:	0.981	**Durbin-Watson:**	1.856
Prob(Omnibus):	0.612	**Jarque-Bera (JB):**	1.095
Skew:	0.129	**Prob(JB):**	0.578
Kurtosis:	2.783	**Cond. No.**	5.69e+05

Table B.7: Regression table associated with Figure B.3a

Dep. Variable:	Women Bias	**R-squared:**	0.062
Model:	OLS	**Adj. R-squared:**	0.053
Method:	Least Squares	**F-statistic:**	7.464
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.000725
Time:	19:59:29	**Log-Likelihood:**	601.86
No. Observations:	230	**AIC:**	-1198.
Df Residuals:	227	**BIC:**	-1187.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	4.434e-05	1.15e-05	3.863	0.000	2.17e-05	6.7e-05
counts	2.932e-10	4.64e-09	0.063	0.950	-8.85e-09	9.43e-09
const	-0.0198	0.001	-15.407	0.000	-0.022	-0.017

Omnibus:	0.434	**Durbin-Watson:**	1.827
Prob(Omnibus):	0.805	**Jarque-Bera (JB):**	0.223
Skew:	0.051	**Prob(JB):**	0.894
Kurtosis:	3.113	**Cond. No.**	3.03e+05

Table B.8: Regression table associated with Figure B.3b

Dep. Variable:	Women Bias	**R-squared:**	0.116
Model:	OLS	**Adj. R-squared:**	0.108
Method:	Least Squares	**F-statistic:**	14.86
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	8.59e-07
Time:	19:59:32	**Log-Likelihood:**	600.10
No. Observations:	230	**AIC:**	-1194.
Df Residuals:	227	**BIC:**	-1184.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	3.197e-05	5.96e-06	5.366	0.000	2.02e-05	4.37e-05
counts	2.304e-09	2.54e-09	0.907	0.365	-2.7e-09	7.31e-09
const	-0.0262	0.001	-19.937	0.000	-0.029	-0.024

Omnibus:	12.746	**Durbin-Watson:**	1.810
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	28.422
Skew:	-0.160	**Prob(JB):**	6.73e-07
Kurtosis:	4.692	**Cond. No.**	5.69e+05

Table B.9: Regression table associated with Figure B.3c

Dep. Variable:	Women Bias	**R-squared:**	0.127
Model:	OLS	**Adj. R-squared:**	0.119
Method:	Least Squares	**F-statistic:**	16.48
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	2.08e-07
Time:	19:59:34	**Log-Likelihood:**	571.07
No. Observations:	230	**AIC:**	-1136.
Df Residuals:	227	**BIC:**	-1126.
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Human Stereotype Score	7.525e-05	1.31e-05	5.734	0.000	4.94e-05	0.000
counts	-1.464e-09	5.3e-09	-0.276	0.783	-1.19e-08	8.99e-09
const	-0.0445	0.001	-30.277	0.000	-0.047	-0.042

Omnibus:	16.771	**Durbin-Watson:**	1.840
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36.244
Skew:	0.317	**Prob(JB):**	1.35e-08
Kurtosis:	4.839	**Cond. No.**	3.03e+05

Table B.10: Regression table associated with Figure B.3d

B.2 Amazon Mturk Stereotypes and Embedding bias

Dep. Variable:	Women Bias	**R-squared:**	0.452
Model:	OLS	**Adj. R-squared:**	0.439
Method:	Least Squares	**F-statistic:**	34.69
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	5.72e-07
Time:	20:59:25	**Log-Likelihood:**	80.806
No. Observations:	44	**AIC:**	-157.6
Df Residuals:	42	**BIC:**	-154.0
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0007	0.000	5.890	0.000	0.000	0.001
const	-0.0140	0.006	-2.240	0.030	-0.027	-0.001

Omnibus:	1.880	**Durbin-Watson:**	1.589
Prob(Omnibus):	0.391	**Jarque-Bera (JB):**	1.453
Skew:	0.445	**Prob(JB):**	0.484
Kurtosis:	2.960	**Cond. No.**	52.9

Table B.11: Women embedding bias vs occupation percent difference for occupations for which census data, MTurk data, and embedding bias is available.

Dep. Variable:	Women Bias	**R-squared:**	0.655
Model:	OLS	**Adj. R-squared:**	0.647
Method:	Least Squares	**F-statistic:**	79.88
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	2.87e-11
Time:	20:59:29	**Log-Likelihood:**	90.999
No. Observations:	44	**AIC:**	-178.0
Df Residuals:	42	**BIC:**	-174.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Stereotype Score	0.0611	0.007	8.938	0.000	0.047	0.075
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.12: Women embedding bias vs MTurk Stereotype score for occupations for which census data, MTurk data, and embedding bias is available.

We next perform an additional joint regression, with the crowdsource scores and the occupation percent difference as covariates and the embedding bias as the outcome. The crowdsource scores remain significantly associated with the embedding bias while the occupation percent difference do not (at $p < 10^{-5}$, versus $p = .203$ for occupation percentage, $r^2 = .669$, intercept confidence interval $(-.281, .027)$). This result indicates that the embedding bias is more closely aligned with human stereotypes than with actual occupation participation.

Dep. Variable:	Women Bias	**R-squared:**	0.669
Model:	OLS	**Adj. R-squared:**	0.653
Method:	Least Squares	**F-statistic:**	41.42
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.44e-10
Time:	20:59:29	**Log-Likelihood:**	91.880
No. Observations:	44	**AIC:**	-177.8
Df Residuals:	41	**BIC:**	-172.4
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	0.000	1.294	0.203	-0.000	0.000
Stereotype Score	0.0517	0.010	5.179	0.000	0.032	0.072
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.13: Women embedding bias vs occupation percent difference and MTurk stereotype data for occupations for which census data, MTurk data, and embedding bias is available.

Dep. Variable:	Stereotype Score	**R-squared:**	0.537
Model:	OLS	**Adj. R-squared:**	0.526
Method:	Least Squares	**F-statistic:**	48.78
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	1.53e-08
Time:	20:59:27	**Log-Likelihood:**	-29.168
No. Observations:	44	**AIC:**	62.34
Df Residuals:	42	**BIC:**	65.90
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0106	0.002	6.984	0.000	0.008	0.014
const	-0.1271	0.076	-1.666	0.103	-0.281	0.027

Omnibus:	1.704	**Durbin-Watson:**	1.390
Prob(Omnibus):	0.427	**Jarque-Bera (JB):**	1.140
Skew:	0.392	**Prob(JB):**	0.565
Kurtosis:	3.092	**Cond. No.**	52.9

Table B.14: MTurk Stereotype score vs occupation percent difference and MTurk stereotype data for occupations for which census data, MTurk data, and embedding bias is available.

B.3 Dynamic Analysis

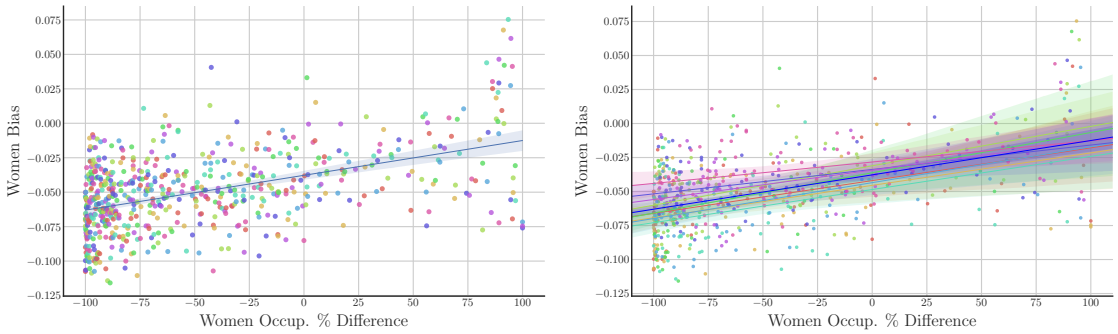
This section contains additional information for gender bias associated with the "dynamic" or over-time analysis. We first present additional information such as regression tables and plots with SVD embeddings for robustness. We then show the occupations and adjectives most associated with men and women, respectively, for each decade.

B.3.1 Model from occupation percentage to embedding bias over time

In this section, we show that the relationship between occupation relative percentage to embedding bias is approximately *consistent* over time, i.e. a *single* linear model performs well across datasets/time. This consistency allows us to extract meaning from trends in the embedding associations over time.

We first train a single model for all (occupation percentage, embedding bias) pairs across time. We compare its performance to a single model where there are additional term for each year. Next, we compare its performance to models trained separately for each year, showing that the single model both has similar parameters and performance to such separate models. Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation). Finally, we repeat the entire analysis using the SVD embeddings.

We note that the separate models and the joint model are similar both in their parameters and their performance.



(a) Single model trained for all time

(b) Separate model trained for each decade

Figure B.4: Scatter plot for occupation proportion vs embedding bias all years of SVD embeddings, along with individually trained regression lines.

Dep. Variable:	Women Bias	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	196.5
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	4.17e-39
Time:	17:12:19	**Log-Likelihood:**	1445.6
No. Observations:	638	**AIC:**	-2887.
Df Residuals:	636	**BIC:**	-2878.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.8e-05	14.020	0.000	0.000	0.000
const	-0.0378	0.001	-28.576	0.000	-0.040	-0.035

Omnibus:	7.243	**Durbin-Watson:**	1.744
Prob(Omnibus):	0.027	**Jarque-Bera (JB):**	7.296
Skew:	0.218	**Prob(JB):**	0.0260
Kurtosis:	3.291	**Cond. No.**	97.4

Table B.15: Single model for Embedding bias vs occupation percentage difference across years

Next, here is the regression with an individual term for each year.

Dep. Variable:	Women Bias	**R-squared:**	0.298
Model:	OLS	**Adj. R-squared:**	0.288
Method:	Least Squares	**F-statistic:**	29.59
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	4.45e-43
Time:	17:12:19	**Log-Likelihood:**	1472.5
No. Observations:	638	**AIC:**	-2925.
Df Residuals:	628	**BIC:**	-2880.
Df Model:	9		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0002	1.77e-05	13.342	0.000	0.000	0.000
const	-0.0346	0.001	-29.916	0.000	-0.037	-0.032
yr_1910.0	-0.0076	0.003	-2.813	0.005	-0.013	-0.002
yr_1920.0	-0.0109	0.003	-4.068	0.000	-0.016	-0.006
yr_1930.0	0.0006	0.003	0.222	0.824	-0.005	0.006
yr_1940.0	-0.0053	0.003	-1.815	0.070	-0.011	0.000
yr_1950.0	-0.0143	0.003	-5.275	0.000	-0.020	-0.009
yr_1960.0	-0.0101	0.003	-3.735	0.000	-0.015	-0.005
yr_1970.0	0.0033	0.003	1.226	0.221	-0.002	0.009
yr_1980.0	0.0011	0.003	0.395	0.693	-0.004	0.006
yr_1990.0	0.0085	0.003	3.099	0.002	0.003	0.014

Omnibus:	23.364	**Durbin-Watson:**	1.746
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27.127
Skew:	0.404	**Prob(JB):**	1.29e-06
Kurtosis:	3.608	**Cond. No.**	8.37e+17

Table B.16: Embedding bias vs occupation percentage difference with additional term for each year

Next, we show the performance and model values when models are trained separately for each embedding year.

Model Year	r^2	coefficient p-value	coefficient value	intercept p-value	intercept value
1910	0.2332	$1.508e-05$	$0.0002408 \pm 5.181e-05$	$4.565e-15$	-0.04197 ± 0.004226
1920	0.2236	$2.398e-05$	$0.0002385 \pm 5.275e-05$	$1.3e-16$	-0.04547 ± 0.004215
1930	0.2865	$8.952e-07$	$0.0002896 \pm 5.386e-05$	$1.904e-10$	-0.03109 ± 0.004195
1940	0.205	0.0002197	$0.000296 \pm 7.525e-05$	$9.189e-08$	-0.03622 ± 0.005962
1950	0.2099	$5.191e-05$	$0.0002437 \pm 5.651e-05$	$9.101e-18$	-0.04853 ± 0.004224
1960	0.2821	$1.586e-06$	$0.0002639 \pm 5.031e-05$	$2.247e-18$	-0.04347 ± 0.003672
1970	0.1827	0.0001801	$0.0001795 \pm 4.536e-05$	$2.905e-16$	-0.03378 ± 0.003176
1980	0.2567	$7.626e-06$	$0.0002158 \pm 4.454e-05$	$6.358e-19$	-0.03429 ± 0.002792
1990	0.1231	0.002905	$0.0001572 \pm 5.088e-05$	$1.706e-14$	-0.02832 ± 0.002913

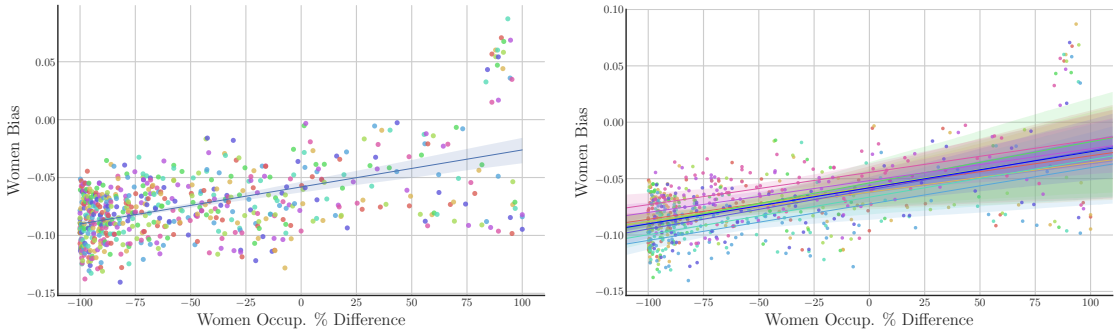
Table B.17: Performances of models trained separately for each decade.

Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation).

Year	MSE using own model	MSE using model from other years
1910	0.000587	0.0005998
1920	0.0006938	0.000743
1930	0.0006222	0.0006474
1940	0.0008265	0.0008325
1950	0.0006428	0.0007494
1960	0.0005713	0.0006107
1970	0.0004408	0.0005077
1980	0.0003628	0.000389
1990	0.0004424	0.0006109

Table B.18: Comparative performance of model trained using a given year's data versus one using all other years

SVD embeddings We repeat the above using SVD embeddings.



(a) Single model trained for all time

(b) Separate model trained for each decade

Figure B.5: Scatter plot for occupation proportion vs embedding bias all years of SVD embeddings, along with individually trained regression lines.

Dep. Variable:	Women Bias	**R-squared:**	0.292
Model:	OLS	**Adj. R-squared:**	0.290
Method:	Least Squares	**F-statistic:**	261.7
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	1.46e-49
Time:	17:44:46	**Log-Likelihood:**	1386.6
No. Observations:	638	**AIC:**	-2769.
Df Residuals:	636	**BIC:**	-2760.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.98e-05	16.178	0.000	0.000	0.000
const	-0.0581	0.001	-40.112	0.000	-0.061	-0.055

Omnibus:	37.968	**Durbin-Watson:**	1.608
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	57.798
Skew:	0.461	**Prob(JB):**	2.81e-13
Kurtosis:	4.151	**Cond. No.**	97.4

Table B.19: Single model for Embedding bias vs occupation percentage difference across years

Next, here is the regression with an individual term for each year.

Dep. Variable:	Women Bias	**R-squared:**	0.353
Model:	OLS	**Adj. R-squared:**	0.343
Method:	Least Squares	**F-statistic:**	38.02
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	6.19e-54
Time:	17:44:46	**Log-Likelihood:**	1415.4
No. Observations:	638	**AIC:**	-2811.
Df Residuals:	628	**BIC:**	-2766.
Df Model:	9		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Women Occup. % Difference	0.0003	1.94e-05	15.873	0.000	0.000	0.000
const	-0.0528	0.001	-41.662	0.000	-0.055	-0.050
yr_1910.0	-0.0036	0.003	-1.218	0.224	-0.009	0.002
yr_1920.0	-0.0067	0.003	-2.277	0.023	-0.012	-0.001
yr_1930.0	-0.0044	0.003	-1.511	0.131	-0.010	0.001
yr_1940.0	-0.0046	0.003	-1.444	0.149	-0.011	0.002
yr_1950.0	-0.0150	0.003	-5.066	0.000	-0.021	-0.009
yr_1960.0	-0.0204	0.003	-6.924	0.000	-0.026	-0.015
yr_1970.0	-0.0088	0.003	-2.998	0.003	-0.015	-0.003
yr_1980.0	0.0019	0.003	0.645	0.519	-0.004	0.008
yr_1990.0	0.0089	0.003	2.952	0.003	0.003	0.015

Omnibus:	74.864	**Durbin-Watson:**	1.536
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	152.277
Skew:	0.687	**Prob(JB):**	8.58e-34
Kurtosis:	4.960	**Cond. No.**	8.37e+17

Table B.20: Embedding bias vs occupation percentage difference with additional term for each year

Next, we show the performance and model values when models are trained separately for each embedding year.

Model Year	r^2	coefficient p-value	coefficient value	intercept p-value	intercept value
1910	0.252	$6.057e-06$	$0.0002816 \pm 5.759e-05$	$2.572e-19$	-0.05791 ± 0.004697
1920	0.1995	$7.468e-05$	$0.000261 \pm 6.205e-05$	$1.333e-19$	-0.06195 ± 0.004958
1930	0.2769	$1.471e-06$	$0.000308 \pm 5.866e-05$	$9.807e-20$	-0.05716 ± 0.004568
1940	0.3064	$3.071e-06$	$0.0003647 \pm 7.084e-05$	$1.026e-13$	-0.05381 ± 0.005612
1950	0.3167	$2.678e-07$	$0.0003328 \pm 5.843e-05$	$6.926e-24$	-0.06643 ± 0.004367
1960	0.3159	$2.78e-07$	$0.0003227 \pm 5.675e-05$	$2.783e-27$	-0.07251 ± 0.004141
1970	0.3701	$1.441e-08$	$0.0003472 \pm 5.414e-05$	$8.662e-25$	-0.05991 ± 0.003791
1980	0.3071	$6.469e-07$	$0.0002819 \pm 5.136e-05$	$7.529e-25$	-0.05174 ± 0.003219
1990	0.2841	$2.029e-06$	$0.0002859 \pm 5.502e-05$	$6.646e-22$	-0.04449 ± 0.00315

Table B.21: Performances of models trained separately for each decade.

Finally, for each embedding year, we compare performance of the model trained for that embedding versus a model trained using all *other* data (leave-one-out validation).

Year	MSE using own model	MSE using model from other years
1910	0.0007251	0.0007358
1920	0.0009599	0.0009724
1930	0.0007379	0.000741
1940	0.0007324	0.0007397
1950	0.000687	0.0007683
1960	0.0007269	0.0009379
1970	0.0006279	0.000639
1980	0.0004824	0.0005458
1990	0.0005174	0.0007325

Table B.22: Comparative performance of model trained using a given year's data versus one using all other years

B.3.2 Average Occupation bias over time

The following is the analogue of Figure 2 with the SVD embeddings.

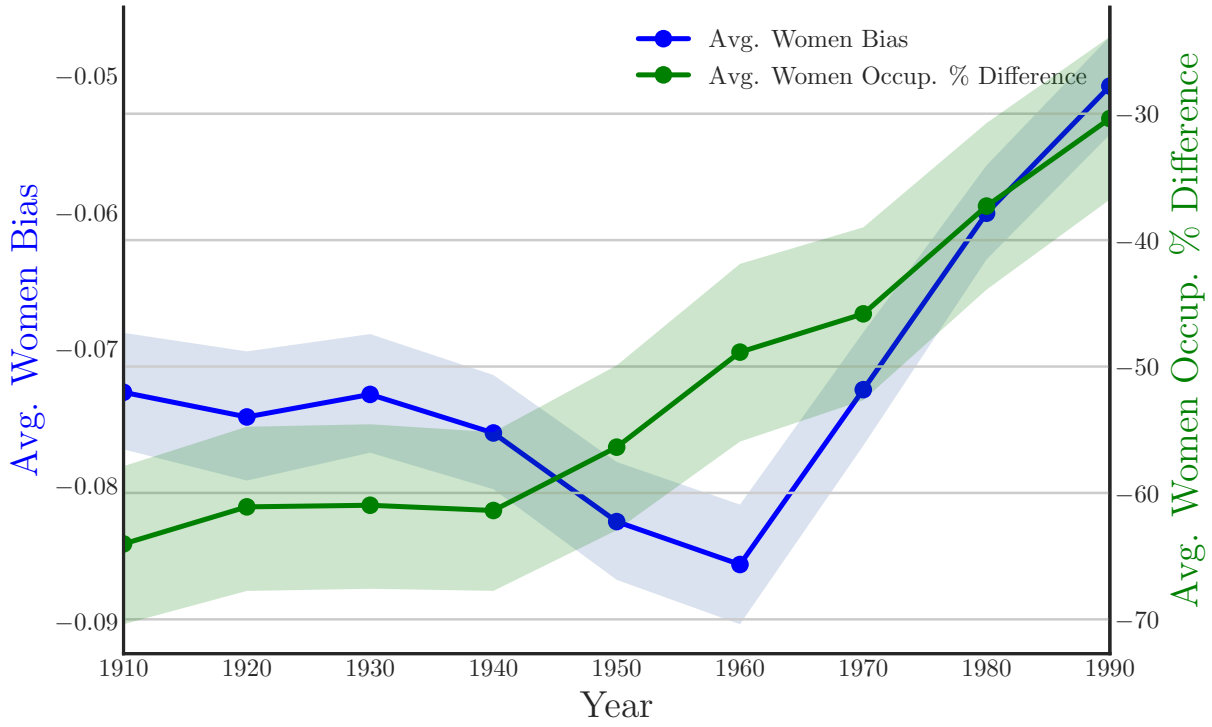


Figure B.6: Gender bias over time in COHA dataset in occupations vs the average relative percentage. In blue is the woman bias in the SVD embeddings, while the in green is the average relative percentage women in the occupation.

B.3.3 Cross-time correlation plots

Here, we give supporting information for the cross time correlation plot for gender bias over time, Figure 4. After describing and reporting results from a statistical test for phase shifts, we provide the same plot but with occupations, and with SVD embeddings.

We first run a test to determine whether the changes in how adjectives are associated with women significantly change between corresponding decades. In particular, we do the following: from the correlation heatmap in the figure, we take the differences between each adjacent column (excluding the changes between the diagonal elements and their neighbors); these differences quantify the change in the associations more robustly than just looking at the correlations

between adjacent embeddings would. Then, we test whether a given set of differences is distributed differently than the rest through a Kolmogorov-Smirnov 2 sample test, which quantifies the difference between 2 empirical distribution functions; for each transition interval (difference between 2 columns), we run a K-S test between the differences in that interval and those between every other interval.

Transition Interval	Test statistic	p value
1910-1920	0.449	0.1206
1920-1930	0.3265	0.4475
1930-1940	0.449	0.1206
1940-1950	0.5306	0.03958
1950-1960	0.3469	0.3714
1960-1970	0.8571	7.173e-05
1970-1980	0.551	0.0291
1980-1990	0.3265	0.4475

Table B.23: Kolmogorov-Smirnov tests for phase change for Figure 4.

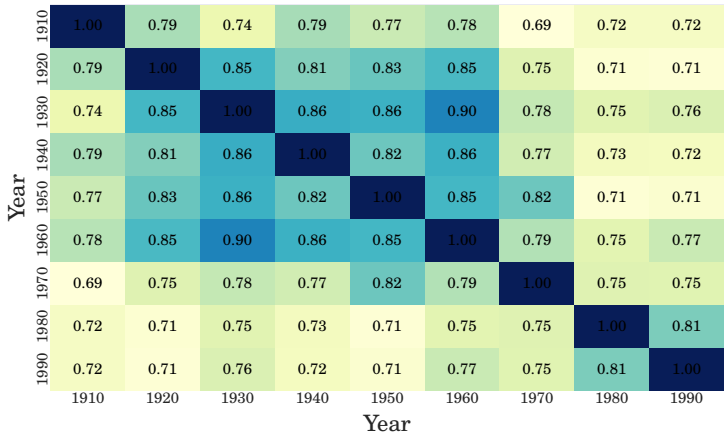


Figure B.7: Pearson correlation in embedding bias scores for occupations over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.3673	0.3041
1920-1930	0.4082	0.1962
1930-1940	0.2857	0.6203
1940-1950	0.5918	0.01519
1950-1960	0.2653	0.7109
1960-1970	0.5306	0.03958
1970-1980	0.4898	0.07071
1980-1990	0.8163	0.0001858

Table B.24: Kolmogorov-Smirnov tests for phase change for Figure B.7.

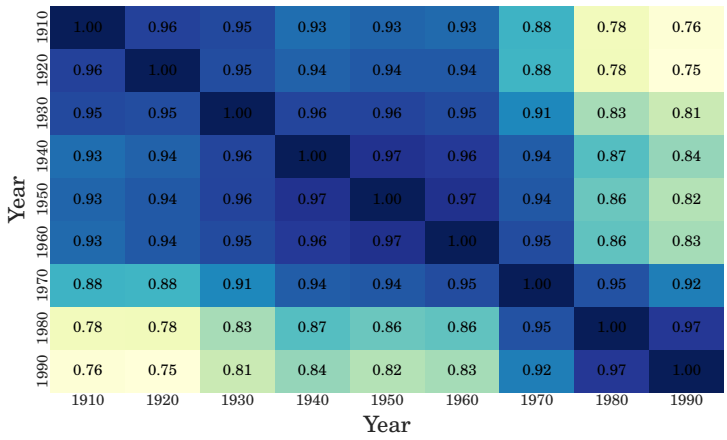


Figure B.8: Pearson correlation in SVD embedding bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.6735	0.003603
1920-1930	0.3265	0.4475
1930-1940	0.4286	0.1547
1940-1950	0.7551	0.0007097
1950-1960	0.7755	0.0004593
1960-1970	0.6939	0.002443
1970-1980	0.8776	4.38e-05
1980-1990	0.5306	0.03958

Table B.25: Kolmogorov-Smirnov tests for phase change for Figure B.8.

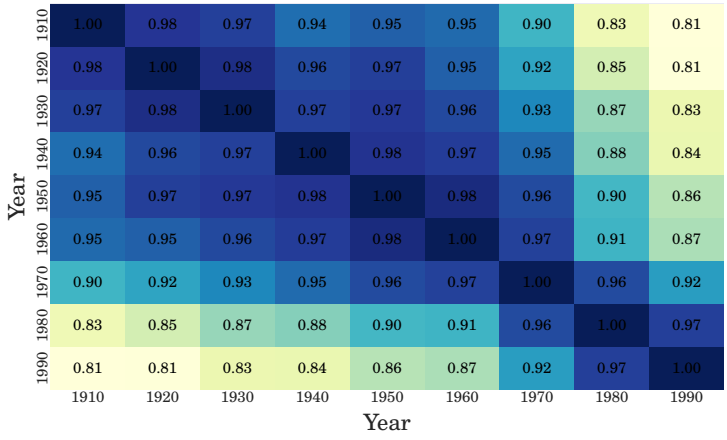


Figure B.9: Pearson correlation in SVD embedding bias scores for occupations over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.5306	0.03958
1920-1930	0.449	0.1206
1930-1940	0.449	0.1206
1940-1950	0.3673	0.3041
1950-1960	0.6122	0.01079
1960-1970	0.6122	0.01079
1970-1980	0.9592	5.423e-06
1980-1990	0.6531	0.005254

Table B.26: Kolmogorov-Smirnov tests for phase change for Figure B.9.

B.3.4 Tables with top occupations and adjectives for each gender

This subsection contains the top occupations and adjectives for each gender for each decade. We caution that due to the noisy nature of embeddings, these tables must be analyzed in the aggregate rather than focusing on individual associations.

1910	1920	1930	1940	1950	1960	1970	1980	1990
mathematician	accountant	engineer	surveyor	architect	lawyer	architect	architect	architect
soldier	surveyor	architect	architect	engineer	architect	engineer	auctioneer	mathematician
architect	architect	lawyer	engineer	mathematician	surveyor	judge	judge	surveyor
surveyor	lawyer	surveyor	smith	lawyer	soldier	economist	surveyor	engineer
administrator	mathematician	manager	sheriff	sheriff	engineer	soldier	sheriff	pilot
lawyer	sheriff	pilot	lawyer	postmaster	pilot	author	author	lawyer
judge	engineer	author	scientist	surveyor	scientist	surveyor	engineer	author
scientist	statistician	scientist	author	scientist	economist	administrator	broker	judge
author	mason	mathematician	economist	author	author	mason	inspector	soldier
economist	scientist	accountant	mason	soldier	mason	mathematician	police	blacksmith

Table B.27: Most Man occupations in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse
attendant	housekeeper	housekeeper	attendant	housekeeper	attendant	dancer	dancer	housekeeper
housekeeper	attendant	attendant	janitor	attendant	dancer	housekeeper	attendant	midwife
cashier	dancer	dancer	housekeeper	dancer	housekeeper	attendant	housekeeper	dentist
cook	teacher	janitor	midwife	cook	photographer	conductor	midwife	student
bailiff	supervisor	midwife	dentist	gardener	midwife	dentist	statistician	dancer
porter	cook	clerical	cook	cashier	dentist	statistician	student	supervisor
operator	doctor	dentist	clerical	midwife	janitor	baker	conductor	bailiff
supervisor	dentist	cook	clergy	musician	cook	clerical	dentist	physician
clergy	mechanic	teacher	sailor	sailor	porter	sailor	supervisor	doctor

Table B.28: Most Woman occupations in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
honorable	regimental	honorable	honorable	knowledge	gallant	honorable	honorable	honorable
gallant	honorable	trusting	conservative	gallant	honorable	wise	loyal	regimental
regimental	stoic	courageous	ambitious	honorable	sage	knowledge	petty	unreliable
skillful	political	gallant	shrewd	directed	regimental	gallant	gallant	skillful
disobedient	sage	confident	regimental	regimental	knowledge	insulting	lyrical	gallant
faithful	ambitious	adventurous	knowledge	efficient	wise	trusting	honest	honest
wise	reserved	experimental	destructive	sage	conservative	honest	faithful	loyal
obedient	progressive	efficient	misguided	wise	honest	providential	obedient	wise
obnoxious	unprincipled	predatory	gallant	faithful	adventurous	modern	wise	directed
steadfast	shrewd	modern	petty	creative	efficient	regimental	hostile	courageous

Table B.29: Most Man adjectives in each decade in the SGNS embeddings.

1910	1920	1930	1940	1950	1960	1970	1980	1990
charming	charming	charming	delicate	delicate	sweet	attractive	maternal	maternal
placid	relaxed	delicate	placid	sweet	charming	maternal	attractive	morbid
delicate	delicate	soft	sweet	charming	soft	charming	masculine	artificial
passionate	amiable	hysterical	gentle	transparent	relaxed	sweet	impassive	physical
sweet	hysterical	transparent	soft	placid	attractive	caring	emotional	caring
dreamy	placid	sweet	warm	childish	placid	venomous	protective	emotional
indulgent	soft	relaxed	charming	soft	delicate	silly	relaxed	protective
playful	gentle	shy	childish	colorless	maternal	neat	charming	attractive
mellow	attractive	maternal	irritable	tasteless	indulgent	delicate	naive	soft
sentimental	sweet	smooth	maternal	agreeable	gentle	sensitive	responsive	tidy

Table B.30: Most Woman adjectives in each decade in the SGNS embeddings.

B.3.5 Trend analysis for competence and physical related adjectives

The following contains statistical tests associated with Section *Individual words whose biases changed over time*.

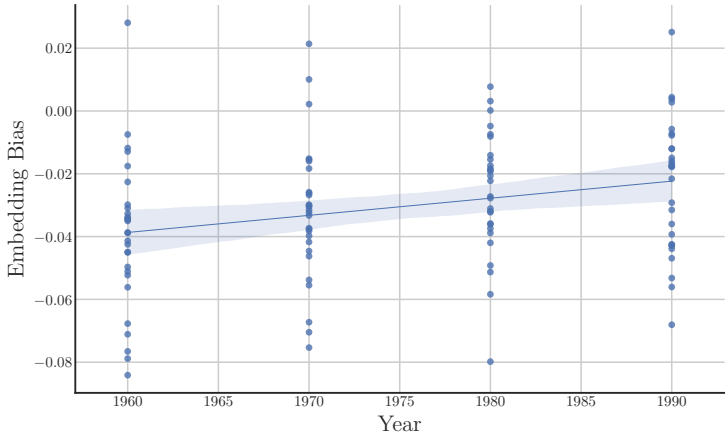


Figure B.10: Embedding bias of competence related words by decade after 1960.

Dep. Variable:	Embedding Bias	**R-squared:**	0.074
Model:	OLS	**Adj. R-squared:**	0.066
Method:	Least Squares	**F-statistic:**	8.844
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	0.00361
Time:	23:45:30	**Log-Likelihood:**	271.25
No. Observations:	112	**AIC:**	-538.5
Df Residuals:	110	**BIC:**	-533.1
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Year	0.0005	0.000	2.974	0.004	0.000	0.001
const	-1.1062	0.362	-3.058	0.003	-1.823	-0.389

Omnibus:	1.434	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.488	**Jarque-Bera (JB):**	0.943
Skew:	0.092	**Prob(JB):**	0.624
Kurtosis:	3.410	**Cond. No.**	3.49e+05

Table B.31: Regression table associated with Figure B.10.

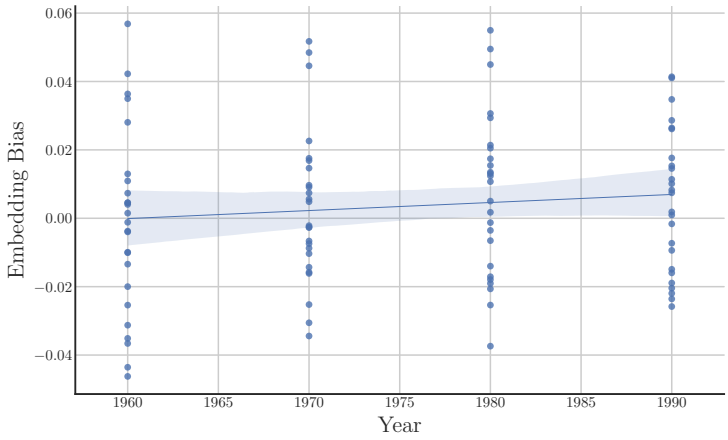


Figure B.11: Embedding bias of appearance related words by decade after 1960.

Dep. Variable:	Embedding Bias	**R-squared:**	0.013
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	1.247
Date:	Sat, 24 Feb 2018	**Prob (F-statistic):**	0.267
Time:	23:45:32	**Log-Likelihood:**	233.94
No. Observations:	100	**AIC:**	-463.9
Df Residuals:	98	**BIC:**	-458.7
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Year	0.0002	0.000	1.117	0.267	-0.000	0.001
const	-0.4612	0.416	-1.108	0.270	-1.287	0.365

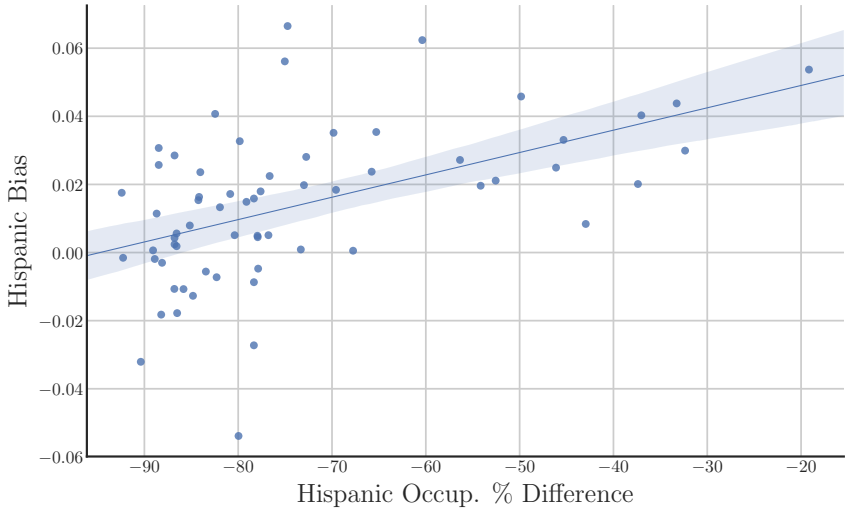
Omnibus:	2.269	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.322	**Jarque-Bera (JB):**	2.208
Skew:	0.296	**Prob(JB):**	0.332
Kurtosis:	2.577	**Cond. No.**	3.49e+05

Table B.32: Regression table associated with Figure B.11.

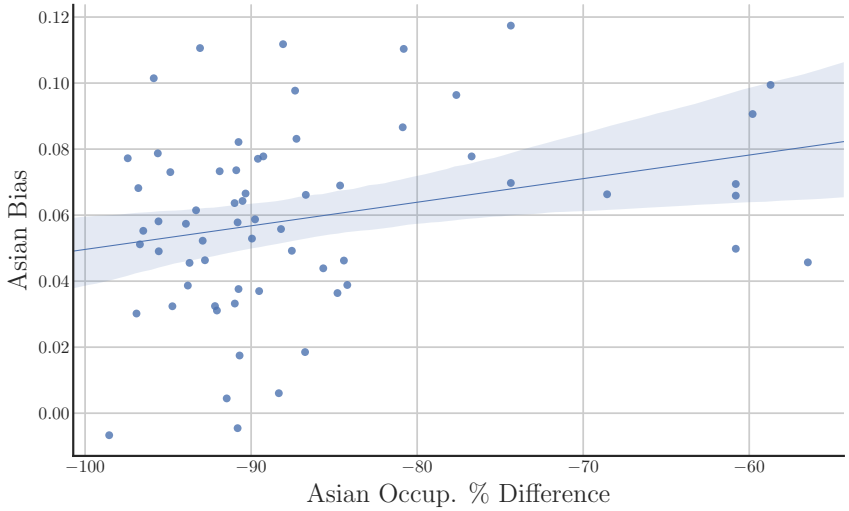
Appendix C Ethnic groups

This section provides additional information and figures related to Section *Quantifying ethnic stereotypes*.

C.1 Snapshot Analysis



(a) Percent difference of Hispanics (compared to Whites) in an occupation vs relative norm distance from occupations to the respective gender words in Google News vectors. More positive indicates more Hispanic associated, for both proportion of occupations and for relative distance. $p < 10^{-5}$ and with r-squared= .279.



(b) Percent difference of Asians (compared to Whites) in an occupation vs relative norm distance from occupations to the respective gender words in Google News vectors. More positive indicates more Asian associated, for both proportion of occupations and for relative distance. $p = .041$ and with r-squared= .065.

Dep. Variable:	Hispanic Bias	**R-squared:**	0.279
Model:	OLS	**Adj. R-squared:**	0.267
Method:	Least Squares	**F-statistic:**	23.93
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	7.42e-06
Time:	23:47:48	**Log-Likelihood:**	163.37
No. Observations:	64	**AIC:**	-322.7
Df Residuals:	62	**BIC:**	-318.4
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Hispanic Occup. % Difference	0.0006	0.000	4.892	0.000	0.000	0.001
const	0.0517	0.008	6.380	0.000	0.036	0.068

Omnibus:	4.172	**Durbin-Watson:**	2.261
Prob(Omnibus):	0.124	**Jarque-Bera (JB):**	4.783
Skew:	-0.065	**Prob(JB):**	0.0915
Kurtosis:	4.333	**Cond. No.**	239.

Table C.1: Regression table corresponding to Figure C.1a

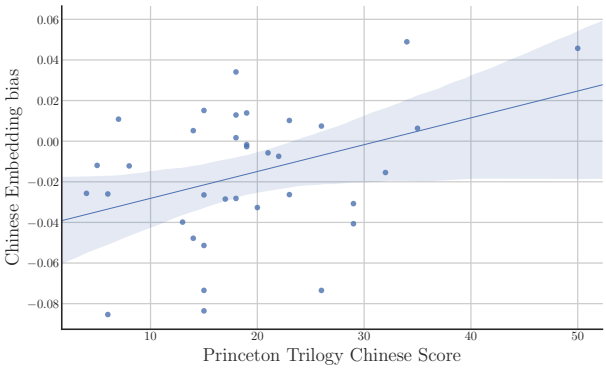
Dep. Variable:	Asian Bias	**R-squared:**	0.065
Model:	OLS	**Adj. R-squared:**	0.050
Method:	Least Squares	**F-statistic:**	4.338
Date:	Thu, 22 Feb 2018	**Prob (F-statistic):**	0.0414
Time:	23:47:45	**Log-Likelihood:**	141.23
No. Observations:	64	**AIC:**	-278.5
Df Residuals:	62	**BIC:**	-274.1
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Asian Occup. % Difference	0.0006	0.000	2.083	0.041	2.41e-05	0.001
const	0.1087	0.024	4.521	0.000	0.061	0.157

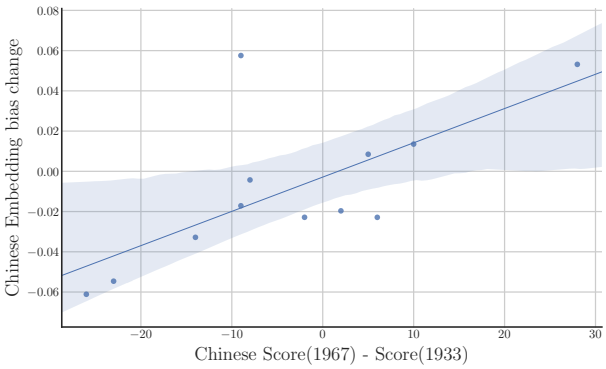
Omnibus:	0.103	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.950	**Jarque-Bera (JB):**	0.157
Skew:	-0.089	**Prob(JB):**	0.925
Kurtosis:	2.834	**Cond. No.**	594.

Table C.2: Regression table corresponding to Figure C.1b

C.2 Princeton Trilogy



(a) Chinese stereotype score vs embedding bias for the corresponding decade across all three trilogy studies. All stereotypes (which are present in the embeddings) in the Chinese portion of Table 1 in [10] are included.



(b) Change in stereotypes between 1933 and 1969

Figure C.2: Plots associated with Princeton Trilogy validation. Full word lists are in Section A.2.

Dep. Variable:	Chinese Embedding bias	**R-squared:**	0.146
Model:	OLS	**Adj. R-squared:**	0.120
Method:	Least Squares	**F-statistic:**	5.644
Date:	Fri, 23 Feb 2018	**Prob (F-statistic):**	0.0235
Time:	21:24:48	**Log-Likelihood:**	72.702
No. Observations:	35	**AIC:**	-141.4
Df Residuals:	33	**BIC:**	-138.3
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Princeton Trilogy Chinese Score	0.0013	0.001	2.376	0.023	0.000	0.002
const	-0.0413	0.012	-3.489	0.001	-0.065	-0.017

Omnibus:	1.808	**Durbin-Watson:**	1.846
Prob(Omnibus):	0.405	**Jarque-Bera (JB):**	1.693
Skew:	-0.468	**Prob(JB):**	0.429
Kurtosis:	2.465	**Cond. No.**	48.0

Table C.3: Regression Table associated with Figure C.2a.

Dep. Variable:	Chinese Embedding bias change	**R-squared:**	0.472
Model:	OLS	**Adj. R-squared:**	0.419
Method:	Least Squares	**F-statistic:**	8.931
Date:	Fri, 23 Feb 2018	**Prob (F-statistic):**	0.0136
Time:	21:24:47	**Log-Likelihood:**	26.887
No. Observations:	12	**AIC:**	-49.77
Df Residuals:	10	**BIC:**	-48.80
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Chinese Score(1967) - Score(1933)	0.0017	0.001	2.988	0.014	0.000	0.003
const	-0.0029	0.008	-0.343	0.739	-0.021	0.016

Omnibus:	18.181	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	13.350
Skew:	1.914	**Prob(JB):**	0.00126
Kurtosis:	6.471	**Cond. No.**	15.1

Table C.4: Regression Table associated with Figure C.2b.

C.3 Dynamic Analysis

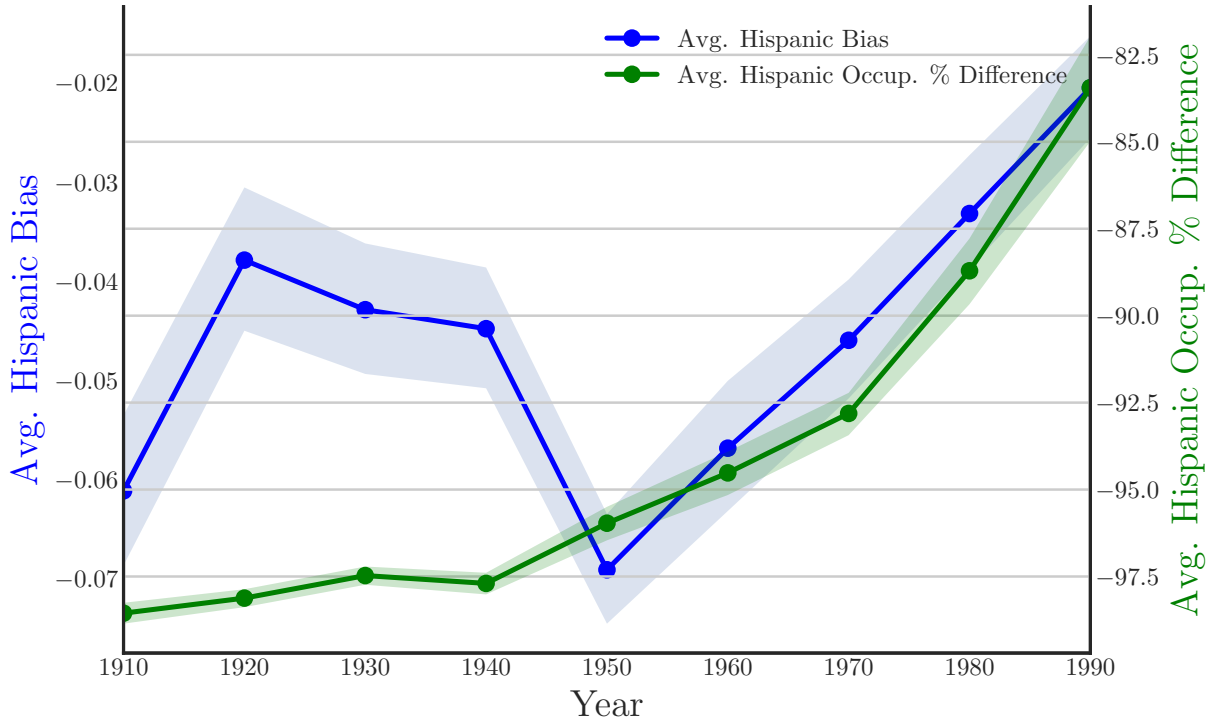


Figure C.3: Ethnic (Hispanic vs White) bias over time in COHA dataset in occupations vs the average percent difference. In blue is the relative Hispanic bias in the SGNS embeddings, while in green is the average percent difference of Hispanics in each occupation.

C.4 Cross-time Correlation plots

Here, we give supporting information for the cross time correlation plot for Asian bias over time. We first provide the results of the KolmogorovSmirnov tests for Figure 5 (as we performed and described in Appendix Section B.3.3). We then show the same plots and test results for Russian and Hispanic associations, respectively.

Transition Interval	Test Statistic	p-value
1910-1920	0.2653	0.7109
1920-1930	0.3469	0.3714
1930-1940	0.3265	0.4475
1940-1950	0.2653	0.7109
1950-1960	0.6122	0.01079
1960-1970	0.3673	0.3041
1970-1980	0.7551	0.0007097
1980-1990	0.4898	0.07071

Table C.5: KolmogorovSmirnov tests for phase change for Figure 5.

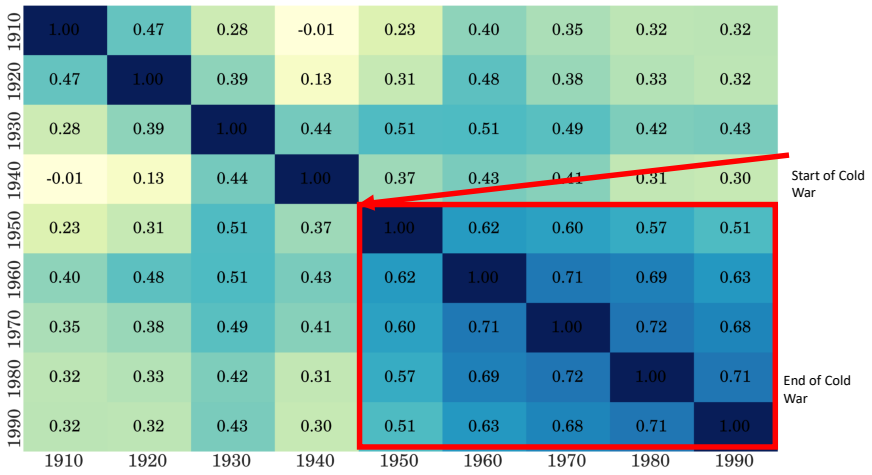


Figure C.4: Pearson correlation in SGNS embedding Russian bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.2653	0.7109
1920-1930	0.5102	0.05321
1930-1940	0.5102	0.05321
1940-1950	0.7347	0.001084
1950-1960	0.4286	0.1547
1960-1970	0.5102	0.05321
1970-1980	0.5306	0.03958
1980-1990	0.7143	0.001637

Table C.6: KolmogorovSmirnov tests for phase change for Figure C.4.

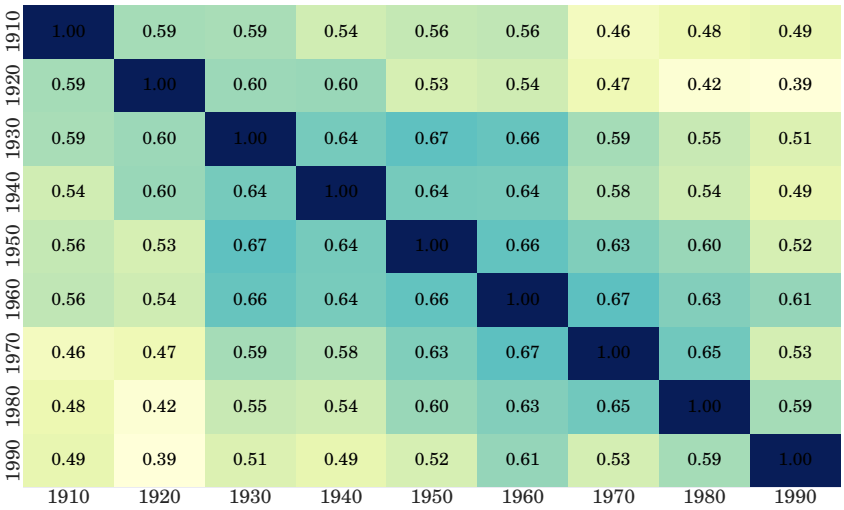


Figure C.5: Pearson correlation in SGNS embedding Hispanic bias scores for adjectives over time between embeddings for each decade.

Transition Interval	Test statistic	p value
1910-1920	0.1837	0.9729
1920-1930	0.7143	0.001637
1930-1940	0.5102	0.05321
1940-1950	0.2653	0.7109
1950-1960	0.5102	0.05321
1960-1970	0.4286	0.1547
1970-1980	0.3265	0.4475
1980-1990	0.1429	0.9989

Table C.7: KolmogorovSmirnov tests for phase change for Figure C.5.

1910	1920	1930	1940	1950	1960	1970	1980	1990
irresponsible	mellow	hateful	solemn	disorganized	imprudent	cynical	superstitious	inhibited
envious	relaxed	unchanging	reactive	outrageous	pedantic	solemn	upright	passive
barbaric	haughty	oppressed	outrageous	pompous	irresponsible	mellow	providential	dissolute
aggressive	tense	contemptible	bizarre	unstable	inoffensive	discontented	unstable	haughty
transparent	hateful	steadfast	fanatical	effeminate	sensual	dogmatic	forceful	complacent
monstrous	venomous	relaxed	assertive	unprincipled	venomous	aloof	appreciative	forceful
hateful	stubborn	cruel	unprincipled	venomous	active	forgetful	dry	fixed
cruel	pedantic	disorganized	barbaric	disobedient	inert	dominating	reactive	active
greedy	transparent	brutal	haughty	predatory	callous	disconcerting	fixed	sensitive
bizarre	compassionate	intolerant	disconcerting	boisterous	inhibited	inhibited	sensitive	hearty

Table C.8: Top Asian (vs White) Adjectives over time by relative norm difference.

References

[1] Parker R, Graff D, Kong J, Chen K, Maeda K (2011) English Gigaword Fifth Edition LDC2011t07.

[2] Pennington J, Socher R, Manning C (2014) Glove: Global Vectors for Word Representation in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, eds. Moschitti A, Pang B. (Association for Computational Linguistics, Doha, Qatar), pp. 1532–1543.

[3] Chalabi M, Flowers A (2014) Dear Mona, Whats The Most Common Name In America?

[4] Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings in *Advances in Neural Information Processing Systems 29*, eds. Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R. (Curran Associates, Inc.), pp. 4349–4357.

[5] Williams JE, Best DL (1977) Sex Stereotypes and Trait Favorability on the Adjective Check List. *Educational and Psychological Measurement* 37(1):101–110.

[6] Williams JE, Best DL (1990) *Measuring sex stereotypes: A multination study, Rev. ed.*, Measuring sex stereotypes: A multination study, Rev. ed. (Sage Publications, Inc, Thousand Oaks, CA, US).

[7] Gunkel P (1987) 638 Primary Personality Traits.

[8] Katz D, Braly K (1933) Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology* 28(3):280.

[9] Gilbert GM (1951) Stereotype persistence and change among college students. *The Journal of Abnormal and Social Psychology* 46(2):245.

[10] Karlins M, Coffman TL, Walters G (1969) On the fading of social stereotypes: Studies in three generations of college students. *Journal of personality and social psychology* 13(1):1.

[11] Levanon A, England P, Allison P (2009) Occupational Feminization and Pay: Assessing Causal Dynamics Using 19502000 U.S. Census Data. *Social Forces* 88(2):865–891.