PNAS Nexus, 2024, 3, 1–10
https://doi.org/10.1093/pnasnexus/pgad485
Advance access publication 2 January 2024
Research Report

America's racial framework of superiority and Americanness embedded in natural language

Messi H. J. Lee [a,*], Jacob M. Montgomery [b] and Calvin K. Lai [c]

[a]Division of Computational and Data Sciences, Washington University in St. Louis, St. Louis, MO 63130-4899, USA
[b]Department of Political Science, Washington University in St. Louis, St. Louis, MO 63130-4899, USA
[c]Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, MO 63130-4899, USA
*To whom correspondence should be addressed: Email: hojunlee@wustl.edu
Edited By: Noshir Contractor

Abstract

America's racial framework can be summarized using two distinct dimensions: superiority/inferiority and Americanness/foreignness. We investigated America's racial framework in a corpus of spoken and written language using word embeddings. Word embeddings place words on a low-dimensional space where words with similar meanings are proximate, allowing researchers to test whether the positions of group and attribute words in a semantic space reflect stereotypes. We trained a word embedding model on the Corpus of Contemporary American English—a corpus of 1 billion words that span 30 years and 8 text categories—and compared the positions of racial/ethnic groups with respect to superiority and Americanness. We found that America's racial framework is embedded in American English. We also captured an additional nuance: Asian people were stereotyped as more American than Hispanic people. These results are empirical evidence that America's racial framework is embedded in American English.

Keywords: natural language processing, race, ethnicity, stereotypes, word embeddings

Significance Statement

An important mechanism for the transmission and perpetuation of stereotypes is communication through spoken and written language. In this study, we investigated the communication of racial/ethnic stereotypes with respect to two dimensions that define America's racial framework: superiority/inferiority and Americanness/foreignness. In a word embedding model trained on a billion-word corpus consisting of various texts of American English, we found that racial/ethnic stereotypes along these two dimensions were consistent and robust across different text categories. This work highlights the role that spoken and written language play in sustaining America's racial framework and may inform theory-driven approaches to debias word embeddings.

Introduction

Racial and ethnic minorities in the United States experience distinct forms of racial oppression. Black people have unequal access to health care (1) and are subject to disproportionate rates of police violence and incarceration (2). Asian people in the United States have historically been subject to the uniform profiling as a "model minority" and have been denied the diversity within the Asian American community (3). More recently, they have been subject to discrimination and physical violence as a result of the COVID-19 pandemic (4). Hispanic people in the United States report experiences of discrimination in employment, health care, housing, and police interactions (5) and are often criticized for not speaking English and asked to go back to their country (6). These divergent experiences of racial oppression needed to be accounted for in a general theory of American race relations.

One effort to parsimoniously explain the many distinct forms of stereotyping and discrimination is the Racial Position Model, which understands racial/ethnic hierarchy through two dimensions (7). One dimension is of superiority/inferiority, which is characterized by the perception of racial/ethnic groups in terms of status and competence. In this dimension, the racial/ethnic hierarchy is understood as White > Asian > Black ≈ Hispanic people, in which Black and Hispanic people are stereotyped as relatively lazy and incompetent than their White and Asian counterparts (8) and Asian people are situated between Black and White people (9). The other dimension is of Americanness/foreignness, which is characterized by the perception of racial/ethnic groups as more or less "American." In this dimension, the racial/ethnic hierarchy is understood as White > Black > Asian ≈ Hispanic people, in which Asian and Hispanic people are seen as unassimilable foreigners in the United States (10, 11), with Black people not being viewed as truly American as White people (12).

An important mechanism for the transmission and perpetuation of stereotypes is spoken and written language. As people characterize others on the basis of racial/ethnic group membership or express

Competing Interest: The authors declare no competing interest.
Received: June 6, 2023. **Accepted:** December 26, 2023

thoughts and feelings about relations with a racial/ethnic group, these expressions collide, merge, and influence each other and eventually form a collective image of the group, setting the way in which society perceives and judges that group (13). The communicative processes that produce these collective images happen in conversations and messages, but they also happen in societal products such as movies and magazines. For instance, Asian people have historically been portrayed as both model minorities and unassimilable foreigners in magazines, news articles, and television commercials (10).

Despite the crucial role that spoken and written language play in racial stereotyping, quantitative scientific studies on stereotyping have predominantly centered around self-reported measures (e.g. (8)) or cognitive tasks (e.g. the Implicit Association Test (14)). While these methods provide valuable insight into individuals' attitudes and beliefs, less is known about how stereotypes are communicated (cf. (15)). Given prevailing egalitarian norms, people are often motivated to suppress stereotyping in communication (16, 17). Hence, the study of racial stereotypes within natural language bridges the gap between traditional psychological approaches to the broader social context in which these biases operate, providing a stronger understanding of how stereotypes are formed and sustained. In this article, we looked at a word embedding model trained on a general collection of American English to examine America's racial framework as it is embedded in language. Word embedding models use co-occurrence statistics of words in a text corpus to determine their semantic and syntactic similarities (18). In these models, words are represented as low-dimensional vectors where words frequently used in similar contexts, presumably sharing meaning, are located near each other in the embedding space. This property of word embeddings allows us to measure social groups' associations with stereotypic attributes. For instance, word embeddings trained on language from the Internet show that men are more associated with work and science compared to women (19), and word embeddings show that White people are more associated with pleasantness than Black people in legal text (20, 21).

In this work, we investigated racial/ethnic stereotypes that define the hierarchies of America's racial framework in a general collection of American English. We trained a word embedding model on the Corpus of Contemporary American English (COCA)—a large corpus of American English that spans 30 years (1990–2019) and eight text categories, including academic articles, blogs, fiction, magazines, newspapers, spoken language, TV/movie subtitles, and the Internet (22). Notably, the corpus primarily consisted of language of professional nature where the potential for bias expression was attenuated. Consequently, this choice of corpus allowed us to carry out a more conservative assessment of racial stereotypes in natural language. We first conducted Single-Category Word Embedding Association Tests (SC-WEATs) to assess the extent to which individual racial/ethnic groups were associated with superiority and Americanness. Then, we conducted a series of Word Embedding Association Tests (WEATs) to compare the racial/ethnic groups' associations with the two stereotype dimensions. Finally, we conducted random-effects meta-analyses and meta-regressions using effect sizes derived from à la carte (ALC) embeddings, or representations of groups and attributes specific to each text category, to investigate the consistency of these stereotypes across text categories.

Results

Single-Category Word Embedding Association Tests

Using the SC-WEAT *D*s, we positioned racial/ethnic groups on a 2D plane defined by superiority/inferiority (*y*-axis) and Americanness/foreignness (*x*-axis). As shown in Fig. 1, the four racial/ethnic groups formed a quadrilateral where Black people were positioned on the bottom right-hand vertex, Asian people on the upper left-hand vertex, Hispanic people on the lower left-hand vertex, and White people on the upper right-hand vertex.

White people were most strongly associated with superiority than inferiority ($D = 2.26$, 95% CI = [1.61, 2.91]), followed by Asian people ($D = 1.18$, 95% CI = [0.82, 1.54]), Hispanic people ($D = 0.70$, 95% CI = [0.33, 1.08]), and Black people ($D = 0.60$, 95% CI = [0.30, 0.91]). These findings suggested that all racial/ethnic groups were more strongly associated with superiority than inferiority and that no group was more commonly discussed using words related to inferiority than superiority.

White people were most strongly associated with Americanness than foreignness ($D = 0.87$, 95% CI = [0.55, 1.20]), followed by Black people ($D = 0.26$, 95% CI = [−0.03, 0.56]), Asian people ($D = -0.18$, 95% CI = [−0.47, 0.10]), and Hispanic people ($D = -0.90$, 95% CI = [−1.27, −0.54]). Whereas White people were strongly associated with Americanness and Hispanic people were strongly associated with foreignness, Black and Asian people were not consistently associated with either.

The 2D plot of racial positions identified from SC-WEAT *D* scores provided a high-level overview of the groups' positions with respect to America's racial framework. Now, we turn to WEATs to directly compare the positions of two groups with respect to the same stereotype dimension. WEAT *D* aggregates the scores of two different groups by taking the difference of one group's score from the other. This may cancel out measurement error in the associations and yield an effect size that more precisely estimates the differences in associations between two groups than the comparison of SC-WEAT *D*s.

Word Embedding Association Tests

WEATs in the superiority/inferiority dimension revealed that the superiority/inferiority dimension of America's racial framework was embedded in American English. As hypothesized, White people were stereotyped as more superior than Black people ($D = 1.16$, 95% CI = [0.76, 1.56]), Asian people ($D = 0.69$, 95% CI = [0.29, 1.09]), and Hispanic people ($D = 1.64$, 95% CI = [1.23, 2.04]). Moreover, Asian people were stereotyped as more superior than Black people ($D = 0.48$, 95% CI = [0.08, 0.87]) and Hispanic people ($D = 0.73$, 95% CI = [0.33, 1.13]). The only nonsignificant difference in superiority was between Black and Hispanic people ($D = 0.15$, 95% CI = [−0.24, 0.53]; see Fig. 2).

WEATs in the Americanness/foreignness dimension revealed that the Americanness/foreignness dimension of America's racial framework was embedded in American English. As hypothesized, White people were stereotyped as more American than Black people ($D = 0.51$, 95% CI = [0.11, 0.91]), Asian people ($D = 1.02$, 95% CI = [0.63, 1.41]), and Hispanic people ($D = 1.79$, 95% CI = [1.40, 2.19]). Moreover, Black people were stereotyped as more American than Asian people ($D = 0.45$, 95% CI = [0.04, 0.87]) and Hispanic people ($D = 1.14$, 95% CI = [0.73, 1.55]).

In addition to the hypothesized findings, WEATs revealed a pattern of stereotyping that deviated from the predictions of the Racial Position Model (7). Asian people were stereotyped as more American than Hispanic people ($D = 0.70$, 95% CI = [0.29, 1.11]). This observation contrasted with a prior study in which a comparison of these groups, based on self-reported foreignness ratings, yielded no statistically significant differences (7).

Analyses using Relational Inner Product Association (RIPA) scores, an alternative measure that is more robust to word

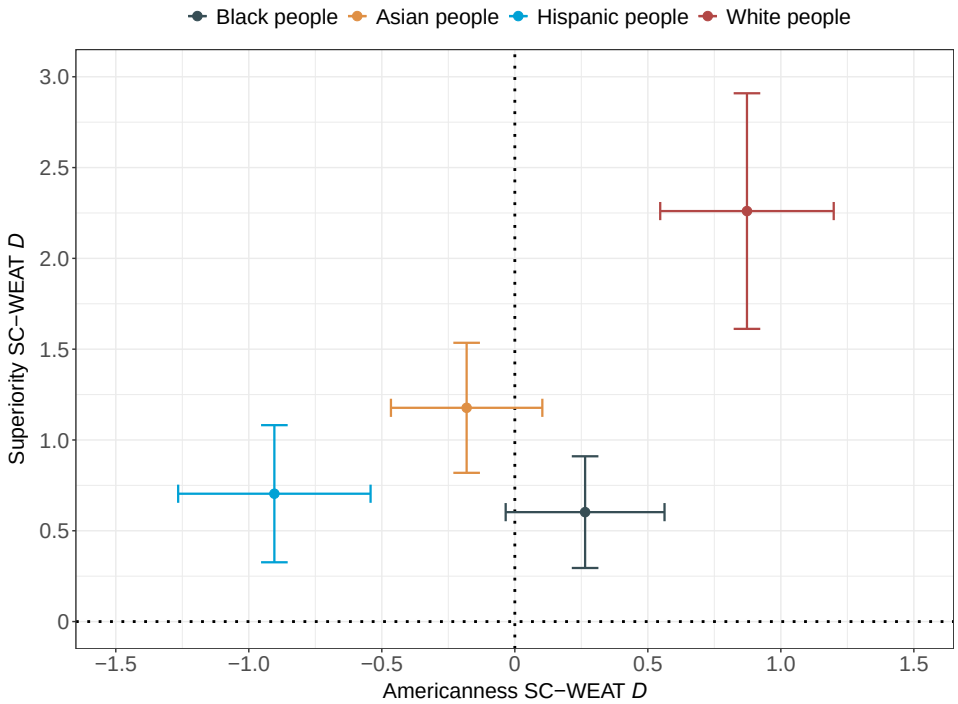

Fig. 1. The 2D plot of racial positions identified from SC-WEAT *D* scores derived from a single word embedding model trained on the Corpus of Contemporary American English. Superiority SC-WEAT *D* scores are used as y-coordinate values, and Americanness SC-WEAT *D* scores are used as x-coordinate values. Bigger y-coordinate value indicates that the group is stereotyped as more superior than inferior, and bigger x-coordinate value indicates that the group is stereotyped as more American than foreign. Error bars represent 95% CIs computed using the standard deviation of the bootstrap distribution of SC-WEAT *D*s.

frequency, replicated the direction of all WEAT findings in the Americanness/foreignness dimension and most of the WEAT findings in the superiority/inferiority dimension. However, two of the findings diverged from WEAT: White people were not stereotyped as more superior than Asian people ($d = 0.05$, 95% CI = [−0.01, 0.11]) and Black people were stereotyped as more superior than Hispanic people ($d = 0.08$, 95% CI = [0.02, 0.14]). For details on the analysis and full reporting on these analyses, see Section S11.

Consistency of WAET *D*s across text categories

The meta-analyses of WEAT *D*s revealed overall consistency across text categories. The meta-analytic estimates of the five group comparisons in the superiority/inferiority dimension that returned significant WEAT *D*s were all large and significant (all *D*s ≥ 0.54; see Fig. 3 and Table S12). Furthermore, the meta-analytic estimates of all six group comparisons in the Americanness/foreignness dimension revealed significant and large overall WEAT *D*s (all *D*s ≥ 0.29; see Fig. 4 and Table S13).

Meta-regression results added strength to the conclusion that racial/ethnic stereotypes are generally consistent across text categories. In the superiority/inferiority dimension, 5 of 48 group comparisons in individual text categories stood out as significantly different from the other text categories (Fig. 3). Stereotyping of White people as more superior than Asian people was smaller in fiction than in other text categories ($b = -0.77$, $z = -2.40$, $P = 0.017$). Stereotyping of White people as more superior than Hispanic people and stereotyping of Black people as more superior than Hispanic people were larger in newspapers than in other text categories ($bs = 1.52$, 0.65, $zs = 2.01$, 2.08, $P = 0.044$, 0.038). Stereotyping of Asian people as more superior than Black people was smaller on the Internet than in other text categories ($b = -0.62$, $z = -2.40$, $P = 0.017$), and stereotyping of Asian people as more superior than Hispanic people was larger in academic articles than in other text categories ($b = 0.83$, $z = 2.21$, $P = 0.027$).

In the Americanness/foreignness dimension, 6 of 48 group comparisons in individual text categories stood out as significantly different from the other text categories (Fig. 4). Stereotyping of White people as more American than Black people was larger in academic articles than in other text categories ($b = 0.65$, $z = 2.06$, $P = 0.040$). Stereotyping of White people as more American than Asian people, stereotyping of White people as more American than Hispanic people, stereotyping of Black people as more American than Asian people, and stereotyping of Black people as more American than Hispanic people were larger in newspapers than in other text categories ($bs = 1.47$, 1.97, 0.54, 0.75, $zs = 6.08$, 3.55, 2.54, 3.50, $P < 0.012$). Stereotyping of Asian people as more American than Hispanic people was smaller in fiction than in other text categories ($b = -0.58$, $z = -1.99$, $P = 0.047$).

We observed greater variability in effect sizes across text categories when comparing Hispanic and White people than other comparisons. The 95% CI of the effect comparing Hispanic and White people was 1.93 to 2.65 times larger than other comparisons in the superiority/inferiority dimension and 1.51 to 3.14 times larger than other comparisons in the Americanness/foreignness dimension. This difference was partially attributable to stereotyping of White people as more superior and American than Hispanic people being much larger in newspapers than in other text categories.

Discussion

In this project, we found empirical evidence that America's racial framework is embedded in American English. Racial/ethnic groups were differentially associated with superiority and Americanness, suggesting that the groups were stereotyped along these dimensions in written and spoken language. Specifically, we found a hierarchy of

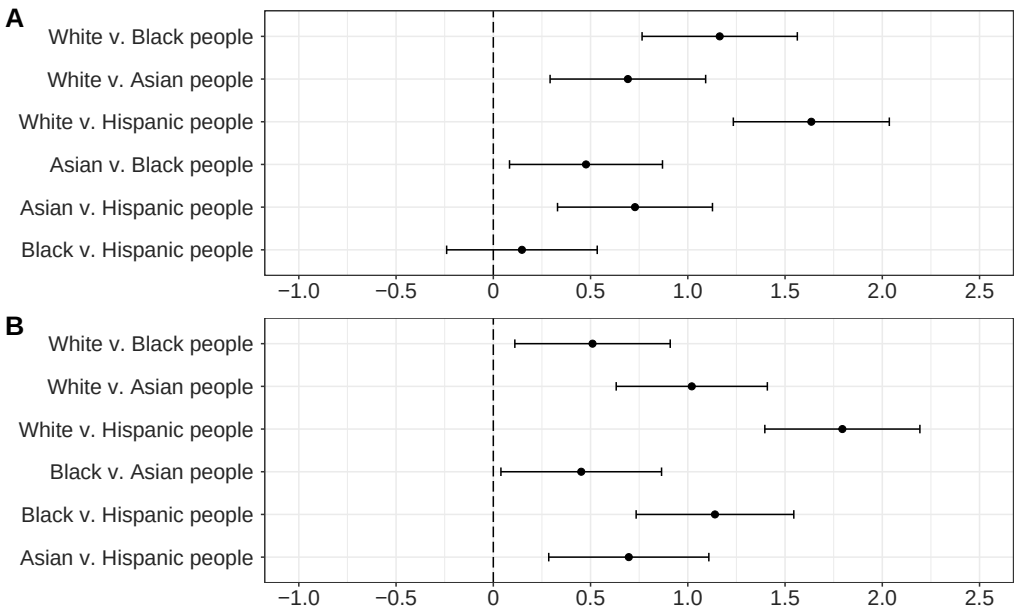

Fig. 2. Pairwise comparisons of racial/ethnic groups with respect to A) superiority/inferiority and B) Americanness/foreignness attributes using WEAT *D* scores. More positive scores indicate stronger associations between the first-labeled group and superiority/Americanness than the second-labeled group. Error bars represent 95% CIs derived from the standard deviation of the permutation distribution of WEAT *D*s.

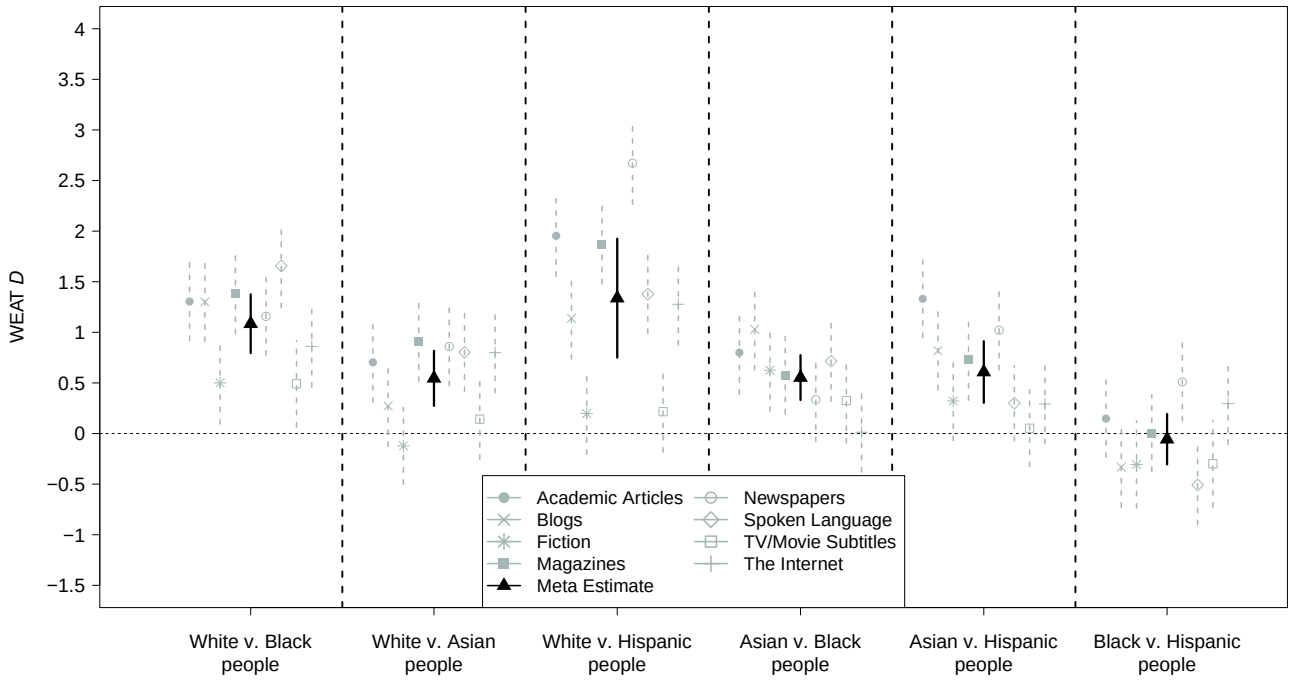

Fig. 3. Comparison of superiority/inferiority WEAT *D*s across group comparisons and text categories. Error bars represent 95% CIs computed from the standard error (i.e. the standard deviation of the permutation distribution of WEAT *D*s).

groups' associations with superiority in the order of White > Asian > Black ≈ Hispanic people and another hierarchy of groups' associations with Americanness in the order of White > Black > Asian > Hispanic people. We found these hierarchies to be robust and consistent across different text categories, highlighting the role language plays in sustaining America's racial framework.

Implications

We view the main contribution of our work as bridging the gap between traditional psychological approaches to the broader social context in which racial stereotypes operate. We show that, despite prevailing egalitarian norms, racial stereotypes are consistently and similarly embedded in various forms of spoken and written language, potentially sustaining America's racial hierarchy.

Building on this understanding of the broader social context in which racial stereotypes operate, our work further enriches the expanding body of research on racial bias in natural language. Earlier works predominantly explored positive vs. negative and/or pleasant vs. unpleasant associations between White and Black people (e.g. (19, 23, 24)). However, recent research,

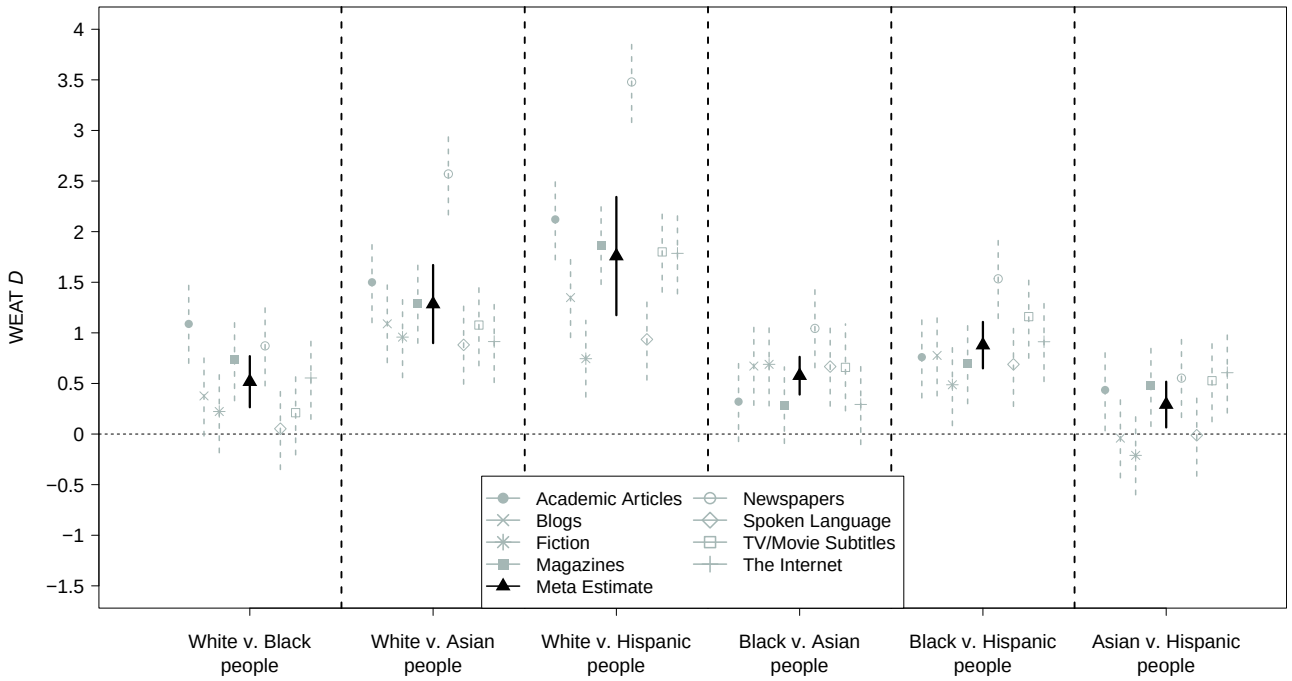

Fig. 4. Comparison of Americanness/foreignness WEAT *D*s across group comparisons and text categories. Error bars represent 95% CIs computed from the standard error (i.e. the standard deviation of the permutation distribution of WEAT *D*s).

exemplified by Charlesworth et al. (25) and Nicolas et al. (26), has expanded the analysis to a wider array of racial/ethnic groups and has ventured beyond simple binary associations to evaluate more complex attributes such as competence and warmth, as outlined in the Stereotype Content Model (8). In line with these advancements, our work not only extends the focus to the four largest racial/ethnic groups in the United States but also evaluates their representations with respect to a contemporary model of race relations. This approach enables us to provide a deeper, more nuanced understanding of the unique experiences of these groups, thereby offering a more comprehensive view of America's racial dynamics.

Linguistic positivity bias

The SC-WEATs revealed that the positions of the groups with respect to the superiority/inferiority dimension were biased towards superiority. We provide three plausible explanations for this finding. First, it may be that stereotypes, as measured using word embeddings, are influenced by the linguistic positivity bias. English words frequently used in everyday communication tend to lean toward positive connotations (27). Consequently, the names selected to represent racial/ethnic groups are more likely to appear in the context of the positively valenced attribute words. Our investigation confirmed this speculation by revealing a positivity bias in COCA. Specifically, superiority attribute words appeared 0.85 million times, in contrast to the 0.29 million occurrences of inferiority attribute words. Given the relatively higher prevalence of superiority attribute words in COCA, it is reasonable to infer that the representations of group words are more likely to exhibit a universal bias towards superiority.

Second, it may be that stereotypes, as measured using word embeddings, reflect contemporary forms of racism. Contemporary racism is less reliant on overt expressions of negative stereotypes or name-calling. Rather, subordinate groups in the United States are otherized in "…a subtle and apparently nonracial way" ((28), p. xvii). In this context, stereotyping may operate as relative differences in positivity rather than relative differences in overt negativity.

Third, it may be that COCA is particularly more positively biased. COCA encompasses diverse text categories such as academic articles, magazines, newspapers, and TV/movie subtitles. These texts are primarily intended for consumption by the general public and undergo editorial processes aimed at filtering out both explicit and subtle expressions of racial stereotypes. Consequently, representations of groups derived from such language sources may be more favorable and thus more superior.

Word embeddings revealed additional nuances

We found an additional nuance in the communication of racial/ethnic stereotypes that was not consistently predicted in prior research: Asian people were stereotyped as more American than Hispanic people. This finding implies that the dynamics of stereotyping might diverge within the context of natural language, thereby revealing nuanced facets of racial stereotyping that traditional psychological approaches may fail to capture.

One explanation for this divergence is the interplay between the two dimensions. While superiority/inferiority and Americanness/foreignness are distinct stereotype dimensions, past research shows that lower status ethnic minorities tend to be considered less prototypical of American identity (29). Supporting this account, the superiority and Americanness scores of the names used to represent Black, Asian, and Hispanic people were moderately correlated in our data ($r(148) = 0.31$, $P < 0.001$). This may explain why Asian people, who are racially stereotyped as more superior, were more associated with Americanness and/or less associated with foreignness than Hispanic people despite both groups being stereotyped as more foreign than the two other groups.

Consistency of racial stereotypes across text categories

Both the meta-analyses and meta-regressions of WEAT *D*s derived from ALC embeddings of group and attribute words revealed

surprising consistency in the strength of racial/ethnic stereotypes in American English. All 11 group comparisons that had yielded large and significant WEAT *D*s revealed large and significant meta-analytic estimates, and meta-regressions revealed that a vast majority (85 out of 96) of WEAT *D*s for individual text categories did not significantly differ from other text categories, suggesting that racial/ethnic stereotypes are consistently embedded in different types of American English. Supplementary tests using ALC embeddings of group words yielded similar results; see Section S9.

While the consistency of WEAT *D*s across text categories indicates the robustness of stereotyping in natural language, the 11 of 96 meta-regressions that found significantly different WEAT *D*s are suggestive of how stereotyping in language varies. We first found that eight meta-regressions indicated that newspapers or academic articles showed greater WEAT *D*s than other types of text. That may not be because of greater stereotyping, but because these forms of text are uniquely characterized by their tendency to abstractly discuss the existence of stereotypes and real-world disparities that link groups with terms related to superiority and Americanness (e.g. "He pointed out that the few images of black Americans that circulate in some countries center on Hollywood films that often depict Black people as criminals." from a 2013 newspaper article). Of the three meta-regressions that found less stereotyping than other types of text, two meta-regressions found less stereotyping in fiction, a form of text conventionally associated with imaginative storytelling and perspective-taking.

Limitations and future directions

One limitation is the use of common last names indicative of racial/ethnic identity to represent groups within the embedding space. Specifically, when selecting names to represent Black and White people, we excluded the more common last names that were shared by other racial/ethnic groups. Consequently, names of White people were more likely to be of Anglo-Saxon (e.g. Clark, Roberts) or German origin (e.g. Wagner, Schmidt), and names of Black people were more likely to be of French origin (e.g. Pierre, Francois). We recognize that such names may present limitations in accurately representing racial/ethnic groups in the United States. To account for this limitation, we performed supplementary analyses using various threshold values to compile group word stimuli. These supplementary analyses indicated that our findings are robust to the threshold values used to compile group word stimuli, suggesting that idiosyncrasies in name selection do not explain our findings. Furthermore, we performed a supplementary analysis using RIPA scores instead of WEAT to account for the possible imbalance in word frequency of group word stimuli. The analysis indicated that our findings are generally robust to the association score used. Nonetheless, future work would benefit from a more robust list of group word stimuli to represent racial/ethnic groups in the embedding space.

Another significant limitation pertains to the corpus used. Despite the diversity of texts the COCA captures, the corpus only represents a small subset of all American English. For example, the text categories that are included in COCA are not natural forms of communication. Instead, most text categories—fiction, newspapers, and TV/movie subtitles—come from professional sectors where texts are disproportionately produced by White people and go through robust editorial processes. As such, the text we assessed may be more representative of the views of the people and groups who hold power. Future work should investigate racial/ethnic stereotypes in a more diverse corpora that proportionately represent language in everyday communication.

Finally, future work should take bottom-up approaches to study racial/ethnic stereotypes in natural language as opposed to a top-down approach. Natural language processing offers the flexibility and granularity to test whether the dimensions explored in this work are in fact the most relevant dimensions that characterize America's racial framework. By employing data-driven discovery methods, researchers could find emergent dimensions of stereotyping that had not previously been explored, allowing for a more comprehensive understanding of how racial/ethnic stereotypes manifest in contemporary American English. This approach would not only enhance the generalizability of our findings but also provide a more robust assessment of the Racial Position Model's suitability as the primary framework for this work.

Conclusion

We used a word embedding model to study the representations of racial/ethnic groups in a corpus of American English in terms of the dimensions that define America's racial framework—superiority/inferiority and Americanness/foreignness. These tests not only revealed that the framework is embedded in American English, but they also shed light on an additional nuance that had not been discussed in this literature as much: Asian people are stereotyped as more American than Hispanic people. Furthermore, we established that these racial/ethnic stereotypes are consistent across text categories. We take this as empirical evidence that America's racial framework is embedded in American English.

Materials and methods

The word embedding model

Prior to training the word embedding model, we preprocessed the COCA. These steps included the removal of nontext characters and lower-casing which are often sufficient for training reliable word embedding models (30). In addition, we removed stop words, identified negations, and preserved common phrases. We discuss these preprocessing steps further in Section S1. We then used the word2vec skip-gram with negative sampling (SGNS) to train the word embedding model. As detailed in Section S2, this model specification yields better representations of rare words and saves computation time (18). We set the context window size to six and the dimensionality of the embedding space to 300 as these values have been shown to provide a good balance of performance and computation time (30).

Word embedding models enable the evaluation of semantic and syntactic similarity between two words through cosine similarity measurements. When two words share meaning, they are positioned closely in the embedding space and share a large cosine similarity value. Conversely, when two words seldom appear together and lack shared contextual terms, they share a small cosine similarity value. This property of word embeddings allows the assessment of stereotypes—associations of social groups with semantic attributes. By comparing the extent to which words chosen to represent a social group are associated with a set of semantic attributes, as opposed to those chosen to represent another social group, researchers can study stereotypes of racial/ethnic groups within natural language. We introduce two tests to quantify stereotypes in word embeddings: The SC-WEAT (31) and the WEAT (19).

The SC-WEATs

Both the SC-WEAT and WEAT use the score value to quantify stereotypes in word embedding models. The score value of a group word w is the difference in mean cosines between w and words used to represent A and mean cosines between the word w and words used to represent B (see Eq. 1).

$$s(w, A, B) = \frac{1}{n(A)} \sum_{a \in A} \frac{w \cdot a}{||w|| \cdot ||a||} - \frac{1}{n(B)} \sum_{b \in B} \frac{w \cdot b}{||w|| \cdot ||b||} \quad (1)$$

In the above equation, w represents the word embedding of a group word, a and b represent the word embedding of attribute words each used to represent attributes A and B, $n()$ represents the number of elements in each set, and $||\cdot||$ is the Euclidean norm. The SC-WEAT summarizes the extent to which a single group is associated with attribute A over attribute B by standardizing the score values derived for all words used to represent the group—the SC-WEAT D (see Eqs. 2 and 3). A large positive SC-WEAT D of group P with respect to attributes A and B indicates that group P is more strongly associated with attribute A than it is associated with attribute B.

$$\text{SC-WEAT } D\ (P, A, B) = \frac{\text{mean}_{p \in P} s(p, A, B)}{\sigma_{s(p,A,B)}} \quad (2)$$

$$\sigma_{s(p,A,B)} = \sqrt{\frac{1}{n(P) - 1} \sum_{p \in P} (s(p, A, B) - \text{mean}_{p \in P} s(p, A, B))^2} \quad (3)$$

To estimate the variability of SC-WEAT Ds, we performed bootstrapping. We randomly sampled, with replacement, the 50 names that were used to represent each group, and derived SC-WEAT D calculations. Repeating this step 1,000 times yielded bootstrap distributions of effect sizes, and the standard deviations of the bootstrap distributions were used to estimate standard errors and the 95% CIs. To summarize these findings, we positioned the four racial/ethnic groups on a 2D plane defined by superiority/inferiority (y-axis) and Americanness/foreignness (x-axis) using their SC-WEAT Ds and 95% CIs.

The WEAT

Whereas SC-WEATs quantify the differential association of one group with respect to two attributes, WEATs can directly compare the differential associations of two different groups in word embedding models.[a]

$$\text{WEAT } D\ (P, Q, A, B) = \frac{\text{mean}_{p \in P} s(p, A, B) - \text{mean}_{q \in Q} s(q, A, B)}{\sqrt{\frac{(n(P) - 1)\sigma^2_{s(P,A,B)} - (n(Q) - 1)\sigma^2_{s(Q,A,B)}}{n(P) + n(Q) - 2}}} \quad (4)$$

In the above equation, p and q represent the word embedding of group words each used to represent groups P and Q. The WEAT summarizes the extent to which a group (P) is associated with attribute A over attribute B compared with the other group (Q)—the WEAT D. It is the difference in mean scores of groups P and Q divided by the pooled standard deviation of the scores of both groups (see Eq. 4). Hence, a large positive WEAT D indicates that group P is more strongly associated with attribute A and/or less associated with attribute B compared with group Q.

To estimate the variability of WEAT Ds, we performed permutation tests. While previous studies utilizing WEATs often employed shuffling of attribute word stimuli to establish a permutation distribution of effect sizes (e.g. (19, 32)), we randomly shuffled group word stimuli instead. This was necessary because the attribute word stimuli were chosen to represent distinct domains of superiority and Americanness, and this property of the list rendered the words nonexchangeable.[b] By shuffling group words, we implicitly enforced a null distribution for effect sizes where the groups were indistinguishable. We permuted the group word stimuli of two different groups and derived WEAT D calculations. Repeating this step 1,000 times yielded a permutation distribution of effect sizes, and the standard deviations of the permutation distributions were used to estimate the standard errors and the 95% CIs.

Using WEATs, we tested the following group comparisons that corresponded to the two hierarchies of the Racial Position Model (7). In the superiority/inferiority dimension, we predicted that White people would be stereotyped as more superior than Black, Asian, and Hispanic people and that Asian people would be stereotyped as more superior than Black and Hispanic people. In the Americanness/foreignness dimension, we predicted that White people would be stereotyped as more American than Black, Asian, and Hispanic people and that Black people would be stereotyped as more American than Asian and Hispanic people. Then, we used WEATs to explore patterns of stereotyping that had not been elaborated upon in the original conception of the Racial Position Model. Specifically, we compared Black and Hispanic peoples' associations with superiority and Asian and Hispanic peoples' associations with Americanness.

Despite the common use of WEATs to quantify stereotypes in word embeddings, some studies have shown that WEATs systematically overestimate bias, particularly when word frequencies in the training corpus are unbalanced (33, 34). To account for this limitation, we conducted robustness checks using an alternative measure called RIPA that is more robust to word frequency (33).

Meta-analyses and meta-regressions

To test the consistency of racial stereotypes across different text categories (e.g. academic articles, newspapers, the Internet) inside COCA, we performed WEATs within each text category and contrasted them with effect sizes from the rest of the corpus. Previous research studying the consistency of stereotypes across smaller collections of text within a corpus trained separate word embedding models on each collection and performed WEATs (e.g. (32)). However, there are limitations to this approach. One limitation is the misalignment of coordinate axes of the word embedding models. Due to the stochastic nature of SGNS and the difference in the vocabulary used in each subcorpus, coordinate axes of word embedding models trained on distinct text collections do not align (35). Furthermore, word embeddings trained on smaller collections of texts are less precise than those trained on a larger corpus.

We addressed these limitations by leveraging the COCA word embedding model to induce ALC embeddings that are specific to each text category (36). We first induced additive embeddings for all words that appear inside a specific text category. The additive embedding of a word was induced by taking the average of all context words as the word appears inside the specific text category. We then performed linear transformations on the additive embeddings. These transformations minimized the disparity between all additive embeddings induced for that specific text category and their COCA word embeddings, thus aligning the coordinate axes of the additive embeddings with that of the COCA word embeddings. This alignment allowed for a reliable comparison of bias across text categories while maintaining embedding quality.

We induced ALC embeddings for all group and attribute words within each text category by adapting the ALC embedding implementation of the *conText* package (R version 4.2.2; 37). Using these embeddings, we performed WEATs. Then, we conducted two types of analyses: First, we performed random-effects meta-analyses to succinctly summarize the effects across text categories. Second, we performed meta-regressions comparing the effect size of an individual text category to those of the rest (e.g. effect size of academic articles vs. effect sizes of all other text categories). The power of the meta-regressions was limited because the total number of effect sizes used to perform the analysis was small ($k = 8$). Thus, the results were interpreted alongside the patterns identified from the meta-analyses. Both the meta-analyses and the meta-regressions were performed using the *meta* package (38) in R. Notably, we performed supplementary analysis to test the robustness of the results when inducing ALC embeddings for group words but not for attribute words. Most major conclusions held regardless of the approach (see Section S9).

Now we introduce the words that were chosen to represent, first, the racial/ethnic groups of interest and, second, the relevant semantic attributes such as superiority and Americanness.

Group word stimuli selection

Racial/ethnic groups were represented using common last names in the United States. Using data from the 2010 Census, we looked for the fifty most common last names that were representative of each racial/ethnic group. For each racial/ethnic group, we ordered the names in descending order of frequency, so the top name was the most frequently occurring name for that racial/ethnic group. We crossed out names that did not appear regularly enough to have a projection in the word embedding model, names that were homonyms for other frequently used words (e.g. Park, Baker), and names common to multiple racial/ethnic groups so that the name would be a reliable indicator of only one category. Specifically, we included names where 70% of people with that name belonged to the group of interest. After removing names that did not meet these criteria, we then selected the top 50 remaining names (Table 1). Given the range of threshold values that we could have selected when choosing group word stimuli, we show that our results are largely robust to the choice of threshold values in Section S4.

Table 1. Racial/ethnic group word stimuli.

Group	Word stimuli
Black people	Pierre, Alston, Bolden, Ruffin, Hairston, Chatman, Francois, Smalls, Lockett, Myles, Bethea, Braxton, Artis, Hollins, Jean-Baptiste, Antoine, Diallo, Bowens, Stallworth, Edmond, Abdi, Baptiste, McKoy, Etienne, Faison, Armstead, Drayton, Kamara, Batiste, Toussaint, Pinkney, Archie, Alexis, Pinckney, McCants, Cobbs, Jean-Louis, Mickens, Broadnax, Bah, Weatherspoon, McClinton, Merriweather, Pettway, Crayton, Thompkins, Mensah, Heyward, Rayford, Desir
Asian people	Nguyen, Kim, Patel, Tran, Chen, Le, Wang, Yang, Singh, Wong, Pham, Lin, Liu, Chang, Huang, Wu, Zhang, Chan, Khan, Shah, Huynh, Yu, Lam, Choi, Kaur, Vang, Ho, Chung, Truong, Xiong, Phan, Vu, Vo, Lim, Lu, Tang, Cho, Ngo, Cheng, Kang, Ng, Dang, Hoang, Hong, Han, Bui, Ma, Chu, Sharma, Xu
Hispanic people	Garcia, Rodriguez, Hernandez, Martinez, Lopez, Gonzalez, Perez, Sanchez, Ramirez, Torres, Flores, Rivera, Gomez, Diaz, Cruz, Morales, Reyes, Gutierrez, Ortiz, Chavez, Ramos, Ruiz, Mendoza, Alvarez, Jimenez, Castillo, Vasquez, Romero, Moreno, Gonzales, Herrera, Aguilar, Medina, Vargas, Castro, Guzman, Mendez, Fernandez, Munoz, Salazar, Garza, Soto, Vazquez, Alvarado, Contreras, Delgado, Pena, Rios, Guerrero, Sandoval
White people	Smith, Miller, Anderson, Martin, Clark, Nelson, Adams, Roberts, Campbell, Phillips, Murphy, Collins, Peterson, Morris, Rogers, Morgan, Cox, Kelly, Bailey, Reed, Myers, Sullivan, Bennett, Hughes, Russell, Reynolds, Olson, Stevens, Snyder, Cole, Wagner, Meyer, Hamilton, Graham, Schmidt, Murray, Gibson, Ellis, Ryan, Wells, Hansen, Webb, Hoffman, Weaver, Johnston, Nichols, Kelley, Mills, Palmer, Tucker

Attribute word stimuli selection

Superiority/inferiority and Americanness/foreignness attributes were represented using word lists curated based on prior stereotyping research. For superiority and inferiority, we first compiled a primary list of words to encompass three distinct domains of superiority as outlined in Zou and Cheryan's qualitative coding scheme (2017): intellectual/mental, moral, and social/cultural superiority. For instance, words like "intelligent," "capable," "competent," "hardworking," and "skilled" were chosen to represent intellectual/mental superiority. Upon confirming that the chosen primary words had projections within the word embedding model, two additional words sharing meaning and exhibiting high cosine similarity values in the trained word embedding model were identified for each primary word (e.g. "truthful" and "candid" for "honest"). The same approach was used to select words to represent inferiority. The full list of words used to represent the two attributes is summarized in Table 2. The primary words are in bold.

For Americanness, we compiled a primary list of words to represent qualities that Americans believe are central to the American identity or what Americans consider to be "True American." Devos and Banaji (12) identified three components that are central to the American identity: civic values such as democracy, equality, or striving for self-improvement; emotional attachment to the nation such as patriotism and defending America when it is criticized; and being born or having spent most of one's life in America. We added words representing each of these components to our primary list, and then for each primary word, we identified two additional words sharing meaning and exhibiting high cosine similarity values with each primary word in the trained word embedding model.

For foreignness, we compiled a primary list of words from research on the perpetual foreigner stereotype. The stereotype postulates that ethnic minority group members will always be treated as foreigners in American society, regardless of their citizenship or birthplace. In this literature, ethnic minority groups are discussed using expressions like "foreigner(s)," "immigrant(s)," "tourist(s)," "not American," "not belong," "outsider(s)," and "noncitizen(s)" (see (10, 39–41). We selected seven commonly used expressions in this literature to match the number of primary words used to represent Americanness. Then, we identified two additional words sharing meaning and exhibiting high cosine similarity values in the trained word embedding model. The full list of words used to represent Americanness and foreignness is summarized in Table 3. Furthermore, we assessed the robustness of the results to an alternative set of foreignness attribute word stimuli, removing words that were just negations of Americanness (i.e. "not_american," "not_americans," and "not_america"). We show that our findings are robust to foreignness attribute word stimuli in Section S5.

Table 2. Words used to represent superiority and inferiority attributes.

Attribute	Word Stimuli
Superior	**intelligent**, smart, highly_intelligent, **capable**, perfectly_capable, adept, **competent**, qualified, competence, **hardworking**, industrious, conscientious, **skilled**, highly_skilled, skillful, **civilized**, civility, enlightened, **disciplined**, discipline, methodical, **law_abiding**, law_abiding_citizens, law_abiding_citizen, **honest**, truthful, candid, **reliable**, dependable, trustworthy, **rich**, wealthy, richer, **powerful**, potent, power, **educated**, highly_educated, informed, **respectable**, decent, respectability, **employed**, hired, recruited
Inferior	**unintelligent**, ignorant, uninformed, **incapable**, not_capable, ineffectual, **incompetent**, inept, incompetence, **lazy**, shiftless, unmotivated, **unskilled**, low_wage, unskilled_labor, **uncivilized**, barbaric, savages, **undisciplined**, irresponsible, reckless, **criminal**, crimes, criminals, **dishonest**, disingenuous, deceitful, **unreliable**, untrustworthy, not_reliable, **poor**, impoverished, poorer, **powerless**, helpless, impotent, **uneducated**, undereducated, poorly_educated, **disreputable**, unsavory, seedy, **unemployed**, jobless, underemployed

Bolded words are the primary words we selected to represent the attributes. Unbolded words are synonyms that are proximate to the primary words in the embedding space.

Table 3. Words used to represent Americanness and foreignness attributes.

Attribute	Word Stimuli
American	**democracy**, democracies, democratic, **equality**, equal_rights, equality_opportunity, **self_reliant**, self_sufficient, self_reliance, **patriotic**, patriotism, nationalistic, **belong**, belongs, belonging, **resident**, residents, longtime_resident, **citizen**, citizens, citizenry
Foreign	**foreign**, foreigners, foreigner, **immigrant**, immigrants, immigration, **tourist**, tourists, visitor, **not_american**, not_americans, not_america, **not_belong**, not_welcome, not_part, **outsiders**, outsider, outcasts, **noncitizen**, noncitizens, not_citizens

Bolded words are the primary words we selected to represent the attributes. Unbolded words are synonyms that are proximate to the primary words in the embedding space.

When compiling the attribute word stimuli, we also made sure that each word used to represent the four attributes occurred at least once inside every text category as ALC embeddings require the word to appear at least once inside the subcorpus in which the embedding is induced. Finally, to assess the generalizability of our results, we performed supplementary tests focusing on different domains of superiority/inferiority (i.e. intellectual/mental, moral, and social/cultural superiority) and a domain of Americanness/foreignness (i.e. legal status). We show that our results are largely consistent across domains of attributes in Section S8.

Notes

[a] Despite recent advances in the field of Natural Language Processing, we use WEATs to assess stereotypes due to their accessibility and transparency. Unlike contextual word embeddings (e.g. BERT) and text generative models (e.g. ChatGPT) whose training data is opaque, we can locally train word embedding models using corpora that are fully accessible to researchers. This approach empowers researchers to make inferences about the language in which the model is trained on as opposed to the model itself.

[b] Exchangeability refers to a requirement of permutation tests where the observations being permuted are equal except for the group they belong to.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

The authors declare no funding.

Author Contributions

Conceptualization: M.H.J.L., J.M.M., and C.K.L.; methodology: M.H.J.L., J.M.M., and C.K.L.; Investigation: M.H.J.L.; visualization: M.H.J.L.; supervision: J.M.M. and C.K.L.; writing—original draft: M.H.J.L.; writing—review and editing: M.H.J.L., J.M.M., and C.K.L.

Data Availability

Most data and all analysis scripts have been made publicly available via OSF and can be accessed at: https://osf.io/n5xyk/. The only exceptions were data files containing the Corpus of Contemporary American English (COCA) that is proprietary and must be purchased separately (https://www.english-corpora.org/coca/).

References

1 Pew Research Center. 2022. 3. *Black Americans' views about health disparities, experiences with health care*. Pew Research Center. https://www.pewresearch.org/science/2022/04/07/black-americans-views-about-health-disparities-experiences-with-health-care/.
2 Hinton E, LeShae H, Reed C. 2018. *An unjust burden: the disparate treatment of Black Americans in the criminal justice system*. New York: Vera Institute of Justice.
3 Yi V, Museus SD. 2015. Model minority myth. In: *The Wiley Blackwell encyclopedia of race, ethnicity, and nationalism*. John Wiley & Sons, Ltd. p. 1–2. https://doi.org/10.1002/9781118663202.wberen528.
4 Gover AR, Harper SB, Langton L. 2020. Anti-Asian hate crime during the COVID-19 pandemic: exploring the reproduction of inequality. *Am J Crim Just*. 45:647–667.
5 Findling MG, *et al.* 2019. Discrimination in the United States: experiences of Latinos. *Health Serv Res*. 54(Suppl 2):1409–1418.
6 Pew Research Center. 2018. 2. *Latinos and discrimination*. Pew Research Center. https://www.pewresearch.org/hispanic/2018/10/25/latinos-and-discrimination/.
7 Zou LX, Cheryan S. 2017. Two axes of subordination: a new model of racial position. *J Pers Soc Psychol*. 112:696–717.
8 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.
9 O'Brien LT, Major B. 2005. System-justifying beliefs and psychological well-being: the roles of group status and identity. *Pers Soc Psychol Bull*. 31:1718–1729.

10 Kim CJ. 1999. The racial triangulation of Asian Americans. *Politics Soc*. 27:105–138.
11 Lacayo CO. 2017. Perpetual inferiority: whites' racial ideology toward Latinos. *Sociol Race Ethn*. 3:566–579.
12 Devos T, Banaji MR. 2005. American = White? *J Pers Soc Psychol*. 88:447–466.
13 Blumer H. 1958. Race prejudice as a sense of group position. *Pac Sociol Rev*. 1:3–7.
14 Greenwald AG, McGhee DE, Schwartz JL. 1998. Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol*. 74:1464–1480.
15 Fiske ST, *et al.* 2015. Talking up and talking down: the power of positive speaking. *J Soc Issues*. 71:834–846.
16 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.
17 Kervyn N, Bergsieker HB, Fiske ST. 2012. The innuendo effect: hearing the positive but inferring the negative. *J Exp Soc Psychol*. 48:77–85.
18 Mikolov T, Chen K, Corrado G, Dean J. 2013. Efficient estimation of word representations in vector space. arXiv [Preprint] [accessed 2023 Apr 22]. http://arxiv.org/abs/1301.3781.
19 Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*. 356:183–186.
20 Matthews S, Hudzina J, Sepehr D. 2022. Gender and racial stereotype detection in legal opinion word embeddings. arXiv [Preprint] [accessed 22 Apr 2023].
21 Rice D, Rhodes JH, Nteta T. 2019. Racial bias in legal language. *Res Politics*. 6:205316801984893.
22 Davies M. 2008. Data from "The Corpus of Contemporary American English (COCA)". https://www.english-corpora.org/coca/.
23 Garg N, Schiebinger L, Jurafsky D, Zou J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A*. 115:E3635–E3644.
24 Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. 2019. Measuring bias in contextualized word representations. In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Florence, Italy: Association for Computational Linguistics. p. 166–172.
25 Charlesworth TES, Sanjeev N, Hatzenbuehler ML, Banaji MR. 2023. Identifying and predicting stereotype change in large language corpora: 72 groups, 115 years (1900–2015), and four text sources. *J Pers Soc Psychol*. 125:969–990.
26 Nicolas G, Bai X, Fiske ST. 2022. A spontaneous stereotype content model: taxonomy, properties, and prediction. *J Pers Soc Psychol*. 123:1243–1263.
27 Dodds PS, *et al.* 2015. Human language reveals a universal positivity bias. *Proc Natl Acad Sci U S A*. 112:2389–2394.
28 Bonilla-Silva E. 2006. *Racism without racists: color-blind racism and the persistence of racial inequality in the United States*. Lanham, Boulder, New York, Toronto, Oxford: Rowman & Littlefield Publishers. p. xvii.
29 Devos T, Mohamed H. 2014. Shades of American identity: implicit relations between ethnic and national identities. *Soc Pers Psychol Compass*. 8:739–754.
30 Rodriguez PL, Spirling A. 2022. Word embeddings: what works, what doesn't, and how to tell the difference for applied research. *J Politics*. 84:101–115.
31 Kurdi B, Mann TC, Charlesworth TES, Banaji MR. 2019. The relationship between implicit intergroup attitudes and beliefs. *Proc Natl Acad Sci U S A*. 116:5862–5871.
32 Charlesworth TES, Yang V, Mann TC, Kurdi B, Banaji MR. 2021. Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol Sci*. 32:218–240.
33 Ethayarajh K, Duvenaud D, Hirst G. 2019. Understanding undesirable word embedding associations. arXiv [Preprint] [accessed 2023 Sep 29]. https://arxiv.org/abs/1908.06361.
34 van Loon A, Giorgi S, Willer R, Eichstaedt J. 2022. Negative associations in word embeddings predict anti-black bias across regions—but only via name frequency. *Proc Int AAAI Conf Weblogs Soc Media*. 16:1419–1424.
35 Hamilton WL, Leskovec J, Jurafsky D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. p. 1489–1501.
36 Khodak M, *et al.* 2018. A La Carte embedding: cheap but effective induction of semantic feature vectors. arXiv [Preprint] [accessed 2023 Apr 22]. http://arxiv.org/abs/1805.05388.
37 Rodriguez PL, Spirling A, Stewart B, Barrie C. 2023. conText: "a la Carte" on text (ConText) embedding regression. [accessed 2023 Apr 22]. https://cran.r-project.org/web/packages/conText/index.html.
38 Balduzzi S, Rücker G, Schwarzer G. 2019. How to perform a meta-analysis with R: a practical tutorial. *BMJ Ment Health*. 22:153–160.
39 Huynh Q-L, Devos T, Smalarz L. 2011. Perpetual foreigner in one's own land: potential implications for identity and psychological adjustment. *J Soc Clin Psychol*. 30:133–162.
40 Kim SY, Wang Y, Deng S, Alvarez R, Li J. 2011. Accent, perpetual foreigner stereotype, and perceived discrimination as indirect links between English proficiency and depressive symptoms in Chinese American adolescents. *Dev Psychol*. 47:289–301.
41 Lee SJ, Wong N-WA, Alvarez AN. 2009. The model minority and the perpetual foreigner: Stereotypes of Asian Americans. In: Tewari N. Alvarez AN, editors. *Asian American psychology: current perspectives*. New York (NY): Routledge/Taylor & Francis Group. p. 69–84.