scientific reports

OPEN



Cognitive underpinnings and ecological correlates of implicit bias against non-Americans in the United States

Benedek Kurdi^{1,2^I}, Keitaro Okura³, Eric Hehman⁴ & Melissa J. Ferguson¹

Of the 330 million residents of the United States, over 40 million were born abroad. Such individuals are routinely referred to using labels such as "alien," "foreigner," and "noncitizen." In this multimethod project relying on data from 5437 U.S. citizens in experimental studies and 193,649 U.S. citizens in archival studies, we examine implicit (automatic) evaluations of non-Americans in the United States, their effects on impression formation, and their ecological correlates in the form of real-life outcomes. In Studies 1A–1C, the labels "alien," "foreigner," and "noncitizen" were found to be highly and similarly implicitly negative. In Studies 2A–2D, applying these labels to specific individuals created immediate implicit evaluations predicted anti-immigrant policy positions at the level of individuals (Study 3A), and a conceptually and statistically related implicit White–American/Asian–foreign implicit stereotype predicted anti-immigrant voting patterns in 18 relevant ballot initiatives at the level of U.S. counties (Study 3B). Across studies, implicit anti-foreigner bias generalized across participant demographics but was somewhat stronger among men and political conservatives. Together, this work highlights the cognitive underpinnings and real-world correlates of robust and pervasive anti-foreigner biases in the United States.

The founding ethos of the United States emphasizes adherence to certain cultural and political values rather than membership in a specific ethnoracial group as the main criterion for belonging to the community of Americans. This idea is often expressed by referring to the United States using monikers such as a multiracial and multicultural "melting pot"¹ or a "country of immigrants." Indeed, more than 46 million people currently living in the United States were born abroad. About 21 million of these individuals are naturalized U.S. citizens, 12 million permanent residents (green card holders), 11 million undocumented immigrants, and 2 million temporary residents (e.g., those on student visas)². However, despite the large number of immigrants in the United States today, combined with the immigrant history of the country's majority groups, those born and raised abroad have and continue to face myriad forms of marginalization and exclusion in U.S. society, including economic inequality³, residential segregation⁴, and hate crime victimization⁵.

Social group-based inequalities are multiply determined^{6,7}, including via a host of historical, political, sociological, and other factors. However, uncovering how human minds represent and apply social group-relevant information can make an important contribution toward understanding and potentially mitigating such inequalities⁸. As such, the studies reported below use a multimethod approach, relying both on experimental studies and an archival study of regional voting patterns to investigate three separate but interrelated aspects of the psychology of anti-foreigner bias in the United States. We pursue this multipronged approach because we believe that the study of societal inequality has much to gain from treating individual and structural levels of analysis as mutually informative and reinforcing^{8,9}.

Specifically, in Studies 1A–1C we ask whether the choice of label used to denote non-Americans ("alien," "foreigner," or "noncitizen") can influence how Americans relate to and evaluate this large and diverse social group. Many Americans share the intuition that relatively subtle differences in labeling can have wide-ranging psychological repercussions. For example, in 2021, the Biden administration ordered immigration agencies to stop using the term "alien" — widely seen as dehumanizing — and to replace it with the ostensibly more neutral

¹Department of Psychology, Yale University, New Haven, CT 06511, USA. ²Department of Psychology, University of Illinois Urbana–Champaign, Champaign, IL 61820, USA. ³Department of Sociology, Yale University, New Haven, CT 06511, USA. ⁴Department of Psychology, McGill University, Montréal, QC H3A 1G1, Canada. ^{\Box}email: kurdi@illinois.edu

"noncitizen"¹⁰. The perceived importance of relevant labels is also illustrated by the public outcry following President Biden's use of the term "illegals" in his 2024 State of the Union address¹¹.

Indeed, the psychological effects of different labels used to refer to the same social groups are well documented in empirical research¹²⁻¹⁵, including in the context of non-American groups and immigration policy¹⁶⁻²⁰. The present project builds upon these existing findings in two ways. First, whereas past work has tended to investigate foreigner labeling effects in the context of undocumented immigration, here we probe the more general distinction between Americans vs. non-Americans, focusing on three labels: "alien," "noncitizen," and "foreigner." The former two labels were selected for inclusion because, as mentioned above, the Biden administration switched from using "alien" to using "noncitizen" in official communications in 2021. However, both of these labels are highly technical and thus not often used in everyday discourse. In addition, "noncitizen" contains a lexical negation, which may shift its evaluations in a negative direction²¹. As such, we also additionally included "foreigner," which is both more colloquial and monomorphemic, in Studies 1A–1C.

Second, prior research has relied exclusively on self-report measures to investigate relevant labeling effects. Such measures are highly informative with respect to participants' consciously endorsed values (e.g., about the inherent equality of different social groups). However, they are less well suited to index more automatic responses, which may be misaligned with such egalitarian views due to social desirability²² or a lack of introspective access to less controlled aspects of social thought and behavior²³. As such, each study in the present project measured both self-reported (explicit) and automatic (implicit) evaluations.

In Studies 2A–2D we turned to investigating the evaluative consequences of applying these labels to particular individuals. Across experiments, we additionally manipulated these individuals' other social group memberships, both to ensure the generalizability²⁴ of the results and to potentially document emergent intersectional biases²⁵. Investigating whether targets' racial group membership moderates foreigner labeling effects is especially important given that social representations, including implicit stereotypes, of who is seen as a (prototypical) member of the category "American" are heavily racialized in the United States^{26–32}. As such, the negative evaluative effects of applying a foreigner label might be exacerbated when such labels are used to refer to non-White individuals. This effect might be particularly strong in the context of Asian targets^{29,31,32}, especially against the backdrop of increased anti-Asian bias as a result of the COVID-19 pandemic³³.

Finally, in Studies 3A–3B we examine the association between implicit (and explicit) anti-foreigner evaluations and social behavior. Specifically, in Study 3A, we focus on the level of individual participants by correlating the extent of American–good/foreign–bad evaluative biases and White–American/Asian–foreign stereotypic biases with participants' anti-immigrant policy views, such as the extent to which they oppose the existence of sanctuary cities or support states suing the federal government over stricter enforcement of immigration regulations. We examine evaluative biases in this study because these biases are the focus of Studies 1A–2D; we additionally include measures of related stereotypic biases because — due to data availability — these biases are the focus of Study 3B, described in more detail below.

Importantly, Study 3A has limitations consistent with much empirical work in this domain. Notably, participants may be responding strategically to the policy items. Moreover, the policy items themselves are hypothetical and, as such, their generalizability beyond the online study setting may be limited³⁴. To address this concern, in Study 3B we draw inspiration from the recent bias of crowds^{35,36} and regional intergroup bias^{37,38} approaches to probe whether regional aggregates of implicit White–American/Asian–foreign stereotypes predict anti-immigrant voting patterns across 18 relevant ballot initiatives.

As such, Studies 3A and 3B have complementary strengths and limitations. The former allows for inferences about individual participants, but its setting is relatively contrived and does not preclude strategic responding on the voting preference items. By contrast, the latter is not suited for individual-level inferences³⁹, but its criterion behaviors are both naturalistic and have obvious external validity given the direct and tangible repercussions of the relevant ballot initiatives for millions of non-Americans (and Americans) living in the United States.

Results and discussion

Study 1A

Study 1A measured implicit (Implicit Association Test; IAT) and explicit (self-reported) evaluations of the labels "American" vs. "alien," "foreigner," and "noncitizen" in a sample of U.S. citizens. The results are shown in Fig. 1. Means, standard deviations, and correlations between implicit and explicit evaluations for this and all remaining studies are reported in Supplementary Tables 1–4.

Overall, participants exhibited a statistically significant and very strong implicit preference for the label "American" relative to the labels "alien," "foreigner," and "noncitizen," t(356) = 28.07, p < 0.001, Cohen's d = 1.49, BF₁₀ = 3.30×10^{88} . This result is unsurprising given robust and near-ubiquitous findings of ingroup favoritism on implicit evaluation measures, especially among members of dominant groups^{40,41}.

Counter to the widespread intuition that such labels are meaningfully different from each other, the specific category label used on the IAT ("alien," "foreigner," or "noncitizen") produced no significant effect, F(2, 354) = 0.21, p = 0.812, $\eta^2 < 0.01$, BF₀₁ = 26.86. But perhaps importantly, in Study 1A we manipulated only the IAT category labels between conditions; category stimuli were the same across conditions and included all three labels ("alien," "foreigner," and "noncitizen"). As such, the lack of a condition effect may have been due to the relatively weak manipulation. We revisit this issue in Study 1B below.

The pattern for explicit evaluations was distinct in two ways. First, participants exhibited a statistically significant and small explicit preference for the labels "alien," "foreigner," and "noncitizen" relative to the label "American," t(338) = -5.12, p < 0.001, Cohen's d = -0.28, BF₁₀ = 1.64×10^4 . One potential interpretation of the discrepancy between implicit and explicit evaluations is that the latter were indicative of the social sensitivity of the domain²² and, relatedly, pressures to appear nonprejudiced⁴².



Fig. 1. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 1A (N=357) by label condition (American/alien, American/foreigner, American/noncitizen). Positive values indicate relative preference for "American" over the other label. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals. *n.s.* = nonsignificant, ** = p < 0.01.

Second, unlike for implicit evaluations, this result was significantly modulated by the specific label used to denote the non-American category, F(2, 336)=4.78, p=0.009, $\eta^2=0.03$, BF₁₀=2.53. Whereas participants showed no significant preference in the American/alien condition, they exhibited an outgroup preference in the American/foreigner and American/noncitizen conditions. Accordingly, the American/alien condition was significantly different from the American/foreigner, t(336)=2.49, p=0.013, and American/noncitizen conditions, t(336)=2.85, p=0.005. The American/foreigner and American/noncitizen conditions did not differ from each other, t(336)=0.32, p=0.748.

The data thus suggest that, at least on the explicit evaluation measure, the "alien" label was perceived relatively more negatively than the "foreigner" and "noncitizen" labels, but not different from the "American" label. Importantly, overall, explicit evaluations did not suggest any negativity toward these three labels relative to "American."

Study 1B

Study 1B followed the setup of Study 1A, but the manipulation of non-American labels was strengthened by varying not only IAT category labels but also IAT category stimuli between participants.

The results are shown in Fig. 2. Despite the stronger manipulation involving both category labels and category stimuli, implicit evaluations perfectly mirrored the results from Study 1A. Specifically, there was, again, a statistically significant and very strong implicit preference for the label "American" relative to the labels "alien," "foreigner," or "noncitizen," t(287) = 20.16, p < 0.001, Cohen's d = 1.19, BF₁₀ = 2.64×10⁵³. Like in Study 1A, the IAT label (and in this case, the IAT stimuli; "alien," "foreigner," or "noncitizen") produced no significant effect, F(2, 285) = 2.08, p = 0.127, $\eta^2 = 0.01$, BF₀₁ = 4.07. This result suggests that implicit preferences reflect evaluations of the labels' shared referents (i.e., non-Americans) rather than connotations of the specific label.

Consistent with Study 1A, participants exhibited a statistically significant and small explicit preference for the labels "alien," "foreigner," and "noncitizen" relative to the label "American," t(287) = -7.08, p < 0.001, Cohen's d = -0.42, BF₁₀ = 5.83 × 10⁸. This result was again significantly moderated by the specific label used to denote the non-American category, F(2, 285) = 4.19, p = 0.016, η^2 = 0.03, BF₁₀ = 1.62. Specifically, the American/alien condition was significantly different from the American/foreigner, t(285) = 2.06, p = 0.041, and American/noncitizen conditions, t(285) = 2.80, p = 0.006, reflecting less positive evaluations of "alien" relative to "foreigner" and "noncitizen." The American/foreigner and American/noncitizen conditions did not differ from each other, t(285) = 0.85, p = 0.398. These small differences notwithstanding, similar to Study 1A, explicit evaluations did not suggest any negativity toward these three labels relative to "American."

Study 1C

Given the relative nature of the measures used in Studies 1A–1B, it is conceivable that differences across non-American labels may have been obscured in those studies given the overwhelming positivity of the American label in comparison. As such, in Study 1C, we directly contrasted the non-American labels with each other. That is, depending on participants' condition assignment, the IAT labels and stimuli were "alien" vs. "foreigner," "alien" vs. "noncitizen," or "foreigner" vs. "noncitizen."

The results are shown in Fig. 3. Unlike in Studies 1A–1B, implicit evaluations significantly differed from each other across the three label conditions, F(2, 303) = 5.07, p = 0.007, $\eta^2 = 0.03$, $BF_{10} = 3.49$. Follow-up analyses indicated that the significant omnibus test was due to a difference between the noncitizen/foreigner and foreigner/ alien, t(303) = 2.89, p = 0.004, and the noncitizen/foreigner and noncitizen/alien conditions, t(303) = 2.62, p = 0.009. The remaining two conditions did not differ from each other t(303) = 0.38, p = 0.707. This pattern of between-condition differences is indicative of the fact that whereas foreigner and alien as well as noncitizen and alien were evaluatively equivalent to each other, foreigner was somewhat more negative than noncitizen. However, even the noncitizen–foreigner comparison produced only a relatively modest effect ($\beta = 0.35$).

The pattern of cross-condition differences was more pronounced on the explicit evaluation measure, F(2, 303) = 21.65, p < 0.001, $\eta^2 = 0.13$, BF₁₀ = 6.10×10^6 . Each pairwise comparison was significant (ps ≤ 0.011). These significant differences were driven by the fact that foreigner was significantly preferred to alien ($\beta = 0.55$) but dispreferred to noncitizen ($\beta = 0.32$). In contrast, reflecting a non-transitive preference ordering, noncitizen and alien did not differ from each other.

Study 2A

After examining implicit evaluations of the labels "alien," "foreigner," and "noncitizen" in isolation in Studies 1A–1C, in Study 2A we turned to studying the evaluative consequences of applying such labels to specific targets. Participants in this study were introduced to two novel individuals (both White men). In an attribute conditioning paradigm^{43,44}, one of these individuals was repeatedly paired with the label "American," whereas the other was repeatedly paired with the labels "alien," "foreigner," and "noncitizen." The goal was to measure implicit and explicit evaluations of the two individuals following this minimal learning manipulation. Given that Studies 1A–1C generally found the three labels to be similarly implicitly negative, in Studies 2A–2D we used them as a set to induce implicit evaluations toward novel targets.

The results are shown in Fig. 4. On the implicit evaluation measure, participants exhibited a small but statistically significant preference for the American over the non-American target, t(301)=3.81, p<0.001, Cohen's d=0.22, BF₁₀=73.48. We obtained a similar result in a Bayesian mixed-effects model, which additionally included a random intercept for the particular images used to represent the two individuals in the learning task and on the IAT, $\beta_0=0.21$ [-0.05; 0.46]; however, given that the 95-percent highest density interval (HDI) overlapped with zero, the condition difference in this study should be interpreted with some caution. As such, Studies 2B–2D were conducted, in part, to examine the robustness of the conditioning effect obtained in Study 2A.



Fig. 2. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 1B (N=288) by label condition (American/alien, American/foreigner, American/noncitizen). Positive values indicate relative preference for "American" over the other label. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals. *n.s.* = nonsignificant, * = p < 0.05.



Fig. 3. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 1C (N=306) by label condition (foreigner/alien, noncitizen/alien, noncitizen/foreigner). Positive values indicate relative preference for the first over the second label. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals. ** = p < 0.01, *** = p < 0.001.



Fig. 4. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 2A (N=302). Positive values indicate relative preference for the American over the foreign target. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals.

Mirroring Studies 1A–1C, explicit evaluations were dissociated from implicit evaluations in Study 2A. Specifically, participants did not show any explicit preference between the American and non-American targets, t(292) = 0.91, p = 0.362, Cohen's d=0.05, BF₀₁=10.12. The same result also emerged in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0 = 0.05$ [-0.09; 0.19]. Similar to the previous studies, this result is likely indicative of the social sensitivity of the domain²² and pressures to appear nonprejudiced⁴².

Study 2B

Study 2B was procedurally identical to Study 2A, with an additional between-participant manipulation of target gender. As such, the male target condition was a direct replication of Study 2A, whereas the female target condition constituted a test of robustness and generalizability²⁴. In addition, the gender contrast is also of theoretical interest given that men are often assumed to be the main targets (as well as the main perpetrators) of intergroup conflict and prejudice⁴⁵.

The results are shown in Fig. 5. Similar to Study 2A, on the implicit evaluation measure, participants exhibited a small but statistically significant preference for the American over the non-American target, t(1039)=5.96, p<0.001, Cohen's d=0.18, $BF_{10}=1.31\times10^6$. The same result was also confirmed in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0=0.19$ [0.06; 0.32]. Target gender produced no significant effect, t(1022.6)=0.77, p=0.443, Cohen's d=0.05, $BF_{01}=10.76$, attesting to the generalizability of the findings from Study 2A.

Mirroring Study 2A, explicit evaluations were dissociated from implicit evaluations. Specifically, participants did not show any explicit preference between the American and non-American targets, t(1008) = -0.16, p = 0.875, Cohen's d < 0.01, BF₀₁ = 27.86. The same result also emerged in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0 = -0.01$ [-0.09; 0.08]. Unlike on the implicit evaluation measure, target gender produced a statistically significant effect, t(1006.8) = 2.37, p = 0.018, Cohen's d = 0.15, BF₁₀ = 1.11. However, given that the effect was small, and the Bayes Factor remained inconclusive, we refrain from interpreting this result.

Study 2C

In the United States, who is perceived to be American or non-American is heavily racialized^{29,31,46}. As such, Study 2C was designed to investigate whether the social category of race moderates the effects observed in Studies 2A–2B. Whereas Studies 2A–2B included only White targets, participants in this study were randomly assigned to learn about two Asian, two Black, two multiracial, or two White targets. However, we did not make the racial group membership of the two targets explicit, and a manipulation check item administered at the end of the study revealed that 58% of participants did not categorize both faces as intended. Therefore, we treat the present study as a further test of generalizability across stimulus materials and revisit the issue of race effects in Study 2D. We additionally report analyses by target race in Supplementary Results.

The results are shown in Fig. 6. Consistent with prior studies, on the implicit evaluation measure, participants exhibited a small but statistically significant preference for the American over the non-American target, t(986)=11.39, p < 0.001, Cohen's d=0.36, BF₁₀=8.37×10²⁴. The same result was also confirmed in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0 = 0.34$ [0.26; 0.44]. This finding attests to the robustness of the implicit preference for American over non-American targets, even following a minimal learning manipulation.

Unlike in previous studies, explicit and implicit evaluations were characterized by similar mean levels. Specifically, mirroring implicit evaluations, participants exhibited an explicit preference for the American over the non-American target, t(967) = 3.97, p < 0.001, Cohen's d = 0.13, BF₁₀ = 86.80. The same result also emerged in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0 = 0.12$ [0.03; 0.20]. However, we note that the explicit evaluation effect was one third in size of the parallel implicit evaluation effect and small by conventional standards.

Study 2D

Study 2D was similar to Study 2C but also included an initial racial categorization task designed to explicitly teach participants the two focal targets' racial category membership. Thanks to the inclusion of this categorization task, manipulation check accuracy improved to 77% from 42% in Study 2C. As such, in the present study, we are able to investigate the effect of the target race variable, as intended.

The results are shown in Fig. 7. Consistent with Studies 2A–2C, on the implicit evaluation measure, participants exhibited a small but statistically significant preference for the American over the non-American target, t(1190) = 10.22, p < 0.001, Cohen's d = 0.30, $BF_{10} = 1.65 \times 10^{20}$. The same result was also confirmed in a Bayesian mixed-effects model accounting for stimulus effects, $\beta_0 = 0.29$ [0.21; 0.36]. In addition, the present results also underscore the generalizability of the pro-American/anti-foreigner bias across target race, given the very strong evidence that we obtained for the lack of any effect associated with this variable, F(3, 1187) = 0.02, p = 0.996, $\eta^2 < 0.01$, $BF_{01} = 420.55$. The same results emerged among the subset of participants with perfect manipulation check performance (see Supplementary Results).

Unlike in previous studies but like in Study 2C, explicit and implicit evaluations were characterized by similar mean levels. Specifically, mirroring implicit evaluations, participants exhibited an explicit preference for the American over the non-American target, t(1155)=4.28, p < 0.001, Cohen's d=0.13, BF₁₀=287.55. The same result also emerged in a Bayesian mixed-effects model accounting for stimulus effects, β_0 =0.13 [0.06; 0.19]. Also similar to implicit evaluations, target race did not modulate the labeling effect, F(3, 1152)=0.67, p=0.571, $\eta^2 < 0.01$, BF₀₁=159.60.



Fig. 5. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 2B (N=1,040) by target gender condition (female vs. male). Positive values indicate relative preference for the American over the foreign target. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals. *n.s.* = nonsignificant, * = p < 0.05.



Fig. 6. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 2C (N=987). Positive values indicate relative preference for the American over the foreign target. Given the unsuccessful manipulation of target race, the plot collapses across target race conditions. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals.



Fig. 7. The distribution of implicit evaluations (top pane) and explicit evaluations (bottom pane) from Study 2D (N=1,191) by target race condition (Asian, Black, multiracial, White). Positive values indicate relative preference for the American over the foreign target. Both implicit and explicit evaluations were standardized for comparability. The dots show condition means, and the error bars represent 95-percent confidence intervals. *n.s.* = nonsignificant.

Study 3A

Studies 1A–2D investigated the cognitive underpinnings of implicit bias against non-Americans, including the evaluative effects of different labels (Studies 1A–1C) and the downstream consequences of applying such labels to specific individuals (Studies 2A–2D). In the final set of studies, we turned to probing the correlates of the American/foreign evaluative bias, and a related bias preferentially linking White Americans to the concept of Americanness and Asian Americans to the concept of foreignness^{29,31,32}, at the level of individual participants (Study 3A) and at the level of U.S. counties (Study 3B). Together, these studies probe whether, and to what extent, the relevant measures of explicit and implicit evaluation, whose basic cognitive properties we investigated in Studies 1A–2D, are predictive of relevant social behaviors.

The aims of Study 3A, which we conducted at the level of individual participants, were threefold. First, we examined whether the American/foreign evaluative IAT used in Studies 1A–2D was associated with antiimmigrant policy positions (e.g., opposition to the existence of sanctuary cities), thus providing a measure of criterion validity. To ensure correspondence with Study 3B, we asked participants to respond to 12 policy items modeled after the 18 real-world ballot initiatives featured in that study. Second, the archival data used to index regional anti-foreigner bias in Study 3B did not include a measure of American/foreign implicit evaluations but rather featured an IAT measuring White/Asian–American/foreign stereotypes. As such, we also included this stereotype IAT as a potential predictor of anti-immigrant policy preferences in the present study. Third, Study 3A also allowed us to investigate whether and to what extent the American/foreign=good/bad evaluation IAT included in Studies 1A–2D and the conceptually related White/Asian–American/foreign stereotype IAT available in the archival data used in Study 3B are related to each other.

The evaluative IAT used in Studies 1A–2D was significantly related to immigration policy views such that a stronger American–good/foreign–bad bias predicted more anti-immigrant policy preferences, $\beta = 0.23$ [0.17; 0.29], t(941)=7.30, p < 0.001. Although the evaluative IAT and the stereotype IAT were significantly correlated with each other, r = 0.15, t(964)=4.59, p < 0.001, performance on the stereotype IAT did not predict policy preferences at the individual level, $\beta = 0.05$ [-0.01; 0.12], t(941)=1.58, p = 0.114. The effects remained virtually unchanged when both IATs were used to predict policy preferences simultaneously.

Attesting to the unique predictive validity of the evaluative IAT, the effects of this variable on immigration policy views remained significant in a model that additionally included two measures of American–good/ foreign–bad explicit evaluations and two measures of Asian–foreign/White–American explicit stereotypes, $\beta = 0.16 \ [0.10; \ 0.22], \ t(872) = 5.33, \ p < 0.001$. The two explicit evaluation measures also had unique effects of $\beta = 0.29 \ [0.22; \ 0.35], \ t(872) = 8.11, \ p < 0.001$, and $\beta = 0.24 \ [0.18; \ 0.32], \ t(872) = 6.93, \ p < 0.001$, respectively, whereas neither the implicit nor the explicit stereotype measures were significantly associated with ballot preferences. The fact that the evaluative IAT showed any incremental predictive validity over and above the explicit evaluation and stereotype measures is noteworthy given that, unlike those measures, the IAT does not share any method variance with the policy preference variable, which was measured via self-report⁴⁷. In addition, given that self-reports are highly controllable, participants had ample opportunity to ensure that their responses across the explicit evaluation and policy preference items were internally consistent with each other⁴⁸.

Together, these analyses suggest that the American/foreign–good/bad evaluation IAT has a unique effect in predicting relevant policy preferences above and beyond parallel explicit evaluation items and the White/Asian–American/foreign stereotype IAT. Although the stereotype IAT did not produce any effects at the individual level, in Study 3B we turn to investigating whether this IAT predicts actual voting patterns (rather than hypothetical policy preferences) at the level of U.S. counties. The evaluative IAT could not be included in this study because it was not available at the regional level.

Study 3B

Although Study 3A had the benefit of allowing for inferences about individual participants, the policy preferences measured were hypothetical, thus limiting the external validity of the design. As such, in Study 3B, we turned to investigating the ecological correlates of anti-foreigner bias in the United States using a real-world outcome as the criterion measure. Specifically, this study probed the relationship between county-level aggregates of the White/Asian–American/foreign stereotype IAT, obtained using archival data from the Project Implicit educational website (http://implicit.harvard.edu/)^{40,41}, and anti-immigrant vote shares in an exhaustive set of 18 real-world ballot initiatives from ten different states over a 28-year period between 1994 and 2022. Descriptive statistics for this study are available in Supplementary Table 5.

Aggregating across the 18 ballot initiatives, we found a significant meta-analytic relationship between the White/Asian–American/foreign IAT and anti-immigrant vote share (see Fig. 8), β =0.24 [0.16; 0.32], z=5.86, p<0.001. In contrast, county-level explicit Asian–foreign/White–American bias did not have a significant meta-analytic effect, β = -0.07 [-0.16; 0.01], z = -1.75, p=0.081. The relationship between county-level implicit bias and anti-immigrant vote share remained significant after controlling for county-level explicit bias, and the size of the effect was virtually unchanged, β =0.28 [0.20; 0.37], z=6.84, p<0.001. These results provide evidence for the predictive value of anti-foreigner implicit bias for real-world outcomes.

We found that all heterogeneity in the effect of implicit bias on anti-immigrant votes could be accounted for by the methodological strength of the data pertaining to each individual ballot initiative, including restriction of range issues in the dependent variable⁴⁹ and the precision with which the independent variable was measured. Specifically, the strength of the relationship increased as a result of more county-level variability in antiimmigrant vote share, b=4.65 [1.30; 8.01], z=2.72, p=0.007, as well as the median by-county sample size, b=0.0009 [0.0002; 0.0017], z=2.38, p=0.017.



Explicit bias and anti-immigrant votes



Fig. 8. Forest plots showing the meta-analytic relationship between county-level White/Asian–American/ foreign implicit stereotypes and anti-immigrant vote share (top pane) and White/Asian–American/foreign explicit stereotypes and anti-immigrant vote share (bottom pane) from Study 3B. County-level bias was calculated from a total of 125,126 participants; a total of 40,488,991 voters participated in the relevant ballot initiatives. Each row represents one ballot initiative. The squares show standardized regression coefficients, and the error bars represent 95-percent confidence intervals. The meta-analytic effect size is depicted using a diamond. AZ = Arizona, CA = California, CO = Colorado, FL = Florida, MA = Massachusetts, MD = Maryland, MT = Montana, NM = New Mexico, OR = Oregon.

Demographic Variability (Studies 1A, 1B, 2A-2D, and 3A)

Finally, we combined data from Studies 1A, 1B, 2A–2D, and 3A to probe demographic correlates of American/ foreign implicit and explicit evaluations measured in those studies. To this end, we fit mixed-effects models to the data, with random intercepts for studies, separately for the two dependent measures. We used likelihood ratio tests to determine improvements in model fit. Relevant descriptive statistics are reported in Supplementary Table 6.

Participants' race, age, place of birth (US vs. non-US), parents' place of birth (US vs. non-US), and the language spoken at home (English, English and some other language, or only non-English) did not have significant effects on implicit bias against non-Americans. Participant gender had a significant effect $\chi^2(2)=30.95$, p<0.001, such that male participants exhibited a stronger bias than did both female participants, $\beta=0.14$, t(5076)=5.30, p<0.001, and participants of other genders, $\beta=0.27$, t(5075)=2.58, p=0.010. Women and participants of other genders did not differ from each other, $\beta=0.14$, t(5074)=1.30, p=0.194. We also observed a significant effect of participant ideology, $\chi^2(1)=46.30$, p<0.001, such that conservative participants exhibited a stronger bias than did liberal participants, b=0.09, t(4983)=6.82, p<0.001. These demographic effects are in line with well-established trends from the relevant literature⁴¹. At the same time, we note that the pro-American/anti-foreigner bias remained significant even among female and strongly liberal participants.

Similar to implicit evaluations, explicit evaluations were moderated by participant ideology, $\chi^2(1) = 102.40$, p < 0.001, such that conservative participants exhibited stronger anti-foreigner biases than did liberal participants, b = 0.15, t(4850) = 10.17, p < 0.001. Unlike for implicit evaluations, highly liberal participants expressed an outgroup preference, whereas highly conservative participants expressed an ingroup preference of equivalent size. No other demographic effects on explicit evaluations were significant.

General discussion

In Studies 1A–1C we demonstrated a significant and very large implicit (automatic) preference for the label "American" over non-American labels including "alien," "foreigner," and "noncitizen." In Studies 2A–2D, we applied these labels to specific targets and showed that an experimental manipulation consisting of as little as 30 stimulus pairings was sufficient to induce implicit negativity toward these targets relative to control targets paired with the label "American." Study 3A provided evidence for predictive validity at the individual level by showing that the American-good/foreign-bad implicit bias was significantly associated with anti-immigrant policy preferences. Finally, in Study 3B, a conceptually and statistically related bias preferentially linking White Americans over Asian Americans to Americanness significantly predicted actual anti-immigrant voting patterns at the level of U.S. counties across ten states.

Across all nine studies reported here, we obtained evidence of robust implicit anti-foreigner biases in the United States. Although these results are in line with theoretical perspectives from the intergroup relations literature emphasizing the ubiquity of ingroup preference as a fundamental motive of human social cognition and behavior^{50,51}, the pervasive nature of implicit anti-foreigner bias documented here is still noteworthy: The bias emerged both toward abstract labels (Studies 1A–1C) and specific individuals (Studies 2A–2D) as well as both at the level of individual participants (Study 3A) and U.S. counties (Study 3B). The result also generalized across different versions of the label that intuitively differ from each other in valence ("alien," "foreigner," and "noncitizen"), across target demographics, including White male (Study 2A), White female (Study 2B), and Asian, Black, and multiracial male targets (Studies 2C–2D), and participants demographics, despite some heterogeneity in results by gender and political ideology. Remarkably, participants' personal and family history of immigration also did not moderate these effects, possibly indicative of the quick assimilation of individuals into U.S. society whose cultural values heavily emphasize the idea of American exceptionalism⁵².

By contrast, the results involving explicit evaluations were more variable, likely reflecting the social sensitivity of the domain²² and pressures to appear nonprejudiced⁴². For example, whereas participants expressed an outgroup preference in Studies 1A–1C, explicit evaluations were neutral in Studies 2A–2B (which involved White targets) and exhibited an ingroup preference in Studies 2C–2D (which involved targets of multiple races). Finally, although explicit anti-foreigner biases were moderately predictive of policy views at the individual level (Study 3A), they were uncorrelated with actual voting patterns at the regional level (Study 3B). Together, these findings attest to the value of using a combination of self-report and indirect measures to understand the antecedents and correlates of human social behavior⁵³, especially against the backdrop of theoretical perspectives that center ingroup preference as an essential and consistent driver of human intergroup cognition and behavior^{50,51}.

The present work also raises some theoretical puzzles and opens up new avenues for empirical inquiry. First, given the relative nature of the Implicit Association Test (IAT) used to measure implicit evaluations and stereotypes in this work, it is not entirely clear to what extent the present results were driven by ingroup preference (i.e., positivity toward Americans), outgroup derogation (i.e., negativity toward non-Americans), or a combination of both. We chose to use the shorthand "anti-foreigner bias" to refer to the pattern of results reported above because Studies 3A and 3B provide clear evidence for the predictive validity of the relevant IATs in the context of anti-immigrant (rather than pro-American) behaviors. Nonetheless, future work may be conducted to obtain more direct evidence on the relative contributors of ingroup preference versus outgroup derogation to the patterns of implicit evaluation and stereotyping obtained here.

Second, implicit evaluations of both the abstract labels "alien," "foreigner," and "noncitizen" and the individuals to whom those labels had been applied were negative. However, the average effect size was an order of magnitude larger for the former than for the latter. Why and how applying a label to a particular target decreases the evaluative strength of that label is an intriguing open question. Resolving this open question may be informed by past social–cognitive work on the "dilution effect"⁵⁴ suggesting that stereotypes are often stronger in the context of abstract groups than they are when applied to specific individuals.

Third, the correlation between the American/foreign-good/bad evaluation IAT and the White/Asian-American/foreign stereotype IAT was relatively modest. This pattern of results is unexpected given previous findings of robust correlation between implicit evaluations and stereotypes^{55,56} and thus ripe for further exploration. It should be noted, however, that whereas past work has compared IATs with the same categories but different attributes (e.g., White/Asian-good/bad and White/Asian-smart/dumb), in the present studies both the categories (American/foreign vs. White American/Asian American) and the attributes (good/bad vs. American/foreign) differed from each other. This feature of the design may have further reduced correspondence⁵⁷ between the two IATs and thus depressed the correlation between them.

Fourth, some of the patterns that emerged with respect to predictive validity might be worth empirical follow-up work. Specifically, it is presently unclear why explicit bias was predictive of anti-immigrant policy views at the level of individuals but not at the level of geographic regions. Similarly, more work is needed to understand why the White/Asian–American/foreign stereotype IAT had no predictive validity at the individual level while being significantly (and uniquely) predictive of actual voting patterns at the regional level. Notably, the American/foreign stereotype IAT and the good/bad evaluative IAT were both available and thus allowed for a direct comparison only in the individual-level Study 3A. Given that the outcome measure in Study 3B consisted of anti-immigrant voting behavior in general (rather than toward Asian Americans in particular), the general evaluative IAT may have been even more predictive of such behaviors than the stereotype IAT that is specific to Asian Americans⁵⁷, had it been available in the aggregate-level archival data.

Such future work may also be able to contribute to the more general theoretical question of the relationship between (implicit) social group evaluations and stereotypes. Whereas implicit evaluations are often conceptualized as mental links between social group targets (such as White Americans and Asian Americans) with positive and negative valence, implicit stereotypes are usually thought of as containing additional information on specific semantic dimensions (such as smart vs. dumb, safe vs. dangerous, or American vs. foreign). Although these two constructs are conceptually distinct, the empirical relationship between them has been repeatedly investigated, with conflicting findings^{55,56,58}. Understanding under what conditions the evaluative and stereotype IATs included in the present project are relatively more or less highly related to each other may help move this debate forward. However, as noted above, unlike most relevant work, the evaluative and stereotype IATs included in the present studies as a result of data availability differed both in their categories (American vs. non-American and White American vs. Asian American) and in their attributes (good vs. bad and American vs. foreign), thus potentially creating an unfair test of the evaluation–stereotype relationship.

These open questions notwithstanding, the present work provides robust evidence for both the cognitive underpinnings and ecological correlates of pervasive anti-foreigner biases among U.S. citizens. These biases emerged both in the abstract and toward particular targets and both at the level of individual participants and at the level of geographic units. Notably, the implicit White–American/Asian–foreign bias was significantly and uniquely associated with the consequential real-world outcome of anti-immigrant vote share in ballot initiatives over the past 30 years. Among other goals, these ballot initiatives have aimed — and often succeeded — to eliminate sanctuary cities, to exclude non-Americans from social services, and to further criminalize undocumented immigrants.

The present findings also dovetail with several recent reviews of the prejudice reduction literature^{8,59,60}, which have concluded that single-shot, light-touch interventions are unlikely to produce meaningful change in entrenched intergroup negativity. Specifically, in the context of the present studies, simply using different labels to refer to non-Americans did not eliminate or even significantly decrease the corresponding biases. As such, despite the good intentions leading to the removal of the term "alien" from the vocabulary of federal agencies, minimal steps of this kind are unlikely to yield meaningful reductions in anti-foreigner attitudes and behaviors. Rather, the prospect of positive change is likely predicated on a joint consideration of the dynamic interplay between the cognitive constraints characterizing human minds⁶¹ and the myriad forms of structural disadvantage and exclusion characterizing non-Americans' everyday social environments in the United States³⁻⁵.

Method

Study 1A

372 participants with U.S. citizenship and U.S. residence were recruited from the Project Implicit educational website (http://implicit.harvard.edu/). In line with the standard Implicit Association Test (IAT)⁶² scoring algorithm⁶³, and as preregistered, participants who did not complete the IAT (n=13) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=2), were excluded from all analyses. IAT trials with response latencies above 10,000 ms were discarded. Participant exclusions resulted in a final sample of 357 participants. Information on the distribution of demographic variables for this and all remaining studies is available in Supplementary Tables 6 and in the open data.

All participants completed an Implicit Association Test (IAT) and self-report measures in the same fixed order, with the IAT administered first and the self-report measures administered last.

The IAT was a standard five-block IAT measuring implicit evaluations of the label "American" relative to the labels "alien," "foreigner," and "noncitizen." The category labels were "American" vs. "Alien," "Foreigner," or "Noncitizen" (the latter manipulated between participants). Irrespective of this between-participant manipulation, the category stimuli were the same for all participants. The category stimuli for the American category included "American," "AMERICAN," "Americans," "AMERICANS," "american," and "americans." The category stimuli for the non-American category included "Alien," "FOREIGNER," "Noncitizen," and "NONCITIZEN." To create further perceptual variation across stimuli, category stimuli were randomly selected to appear in one of the following fonts on each trial: purple, 3-em font size, bold; green, 2.5-em font size, italicized; maroon, 4-em font size, serif; yellow, 3.5-em font size, cursive; and blue, 2-em font

size. The attribute labels were "Good" and "Bad," and the attribute stimuli included "Fantastic," "Good," "Great," "Pleasant," and "Wonderful," and "Awful," "Bad," "Horrible," "Terrible," and "Unpleasant," respectively.

On each trial of the IAT, participants were asked to categorize the stimulus appearing on the screen using the E and I keys on their keyboards. If they made a mistake, a red X was displayed, and participants were asked to correct their response. For each trial, latency until the correct response was recorded. Block 1 (20 trials) consisted of attribute practice trials, whereas block 2 (20 trials) consisted of category practice trials. Block 3 (40 trials) was the first critical block, in which one category and one attribute (e.g., American–good) were assigned to the left response key, and the other category and attribute (e.g., Foreigner–bad) were assigned to the right response keys reversed relative to blocks 2–3. Finally, block 5 (40 trials) was the second critical block where the pairing of categories and attributes was reversed relative to block 3. The order of the two critical blocks was randomized.

IAT scores were calculated using the improved scoring algorithm⁶³ such that positive scores express a relative preference for the American over the non-American label.

To parallel the IAT, participants were asked to respond to 20 self-report items of the form "Cs are a," with C referring to one of the IAT category labels and a referring to one of the IAT attribute stimuli (e.g., "Noncitizens are wonderful"). Participants used 100-point slider scales to enter their responses. The endpoints of the scale were labeled "Completely inaccurate" and "Completely accurate," and the midpoint was labeled "Neither inaccurate nor accurate." Positive items and negative items were averaged separately for each target, the sum of negative items was subtracted from the sum of positive items, and then the composite score for the non-American target subtracted from the composite score for the American target. As such, like for the IAT, positive scores express a preference for the American over the non-American label.

Participants reported standard demographic information, including gender, political orientation, race, and age when registering on the Project Implicit website. At the end of the study, participants additionally indicated whether they were born in the United States, whether their parents were born in the United States, and whether they spoke English and/or some other language at home while growing up. Participants were then debriefed and thanked for their participation.

Study 1B

306 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=11) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=7), were excluded from all analyses. Participant exclusions resulted in a final sample of 288 participants.

The procedure of Study 1B was identical to the procedure of Study 1A, with the following exception: On the IAT, in addition to the category labels, we also manipulated category stimuli between conditions. For example, for a participant assigned to the American/foreigner condition, IAT category stimuli included only variations of the corresponding label (e.g., "Foreigner," "FOREIGNER," "Foreigners," "FOREIGNERS," and "foreigner"). In any given condition, the remaining two labels did not appear at all, thus making the manipulation stronger.

Study 1C

318 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=9) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=3), were excluded from all analyses. Participant exclusions resulted in a final sample of 306 participants.

The procedure of Study 1C was similar to Studies 1A–1B, with the following exception: On the IAT, participants were assigned to one of three between-participant conditions comparing two non-American labels with each other (i.e., alien vs. foreigner, alien vs. noncitizen, or foreigner vs. noncitizen). Participants then responded to the parallel self-report items for the same two non-American labels.

Study 2A

306 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=4) were excluded from all analyses. There were no participants with response latencies of 300 ms or below on 10% or more of IAT trials. Participant exclusions resulted in a final sample of 302 participants.

Study 2A consisted of a learning phase and a test phase. In the learning phase, participants were introduced to two White male targets, randomly selected from a set of four individuals drawn from the Chicago Face Database⁶⁴. Each individual was depicted using four unique images, each displaying a different emotional expression (neutral, happy closed mouth, happy open mouth, and angry). In an attribute conditioning^{43,44} procedure, one of the individuals was repeatedly paired with the label "American" (15 trials) whereas the other individual was repeatedly paired with the labels "alien," "foreigner," and "noncitizen" (5 trials each, resulting in a total of 15 trials). The order of trials was individually randomized. On each trial, an orienting stimulus (***) was displayed for 1000 ms, followed by the simultaneous presentation of a target face and a label for 3500. The intertrial interval was 500 ms.

The test phase was similar to Studies 1A–1C and consisted of an IAT and self-report measures. On the IAT, the neutral facial images of the two targets encountered in the learning phase served as category labels and the four unique facial images served as category stimuli. The IAT procedure and the attribute labels and stimuli were the same as in Studies 1A–1C. The self-report measures were also identical to the ones used in Studies 1A–1C, with the exception that they referenced the two targets encountered in the learning phase rather than the American and non-American labels in the abstract.

Study 2B

1075 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=26) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=9), were excluded from all analyses. Participant exclusions resulted in a final sample of 1040 participants.

Study 2B was procedurally identical to Study 2A, with the following exception: Participants were assigned to one of two between-participant conditions. The male target condition was identical to Study 2A; in the procedurally matched female target condition, participants were introduced to two female (rather than male) targets randomly selected from a set of four individuals drawn from the Chicago Face Database⁶⁴. This condition served as a test of robustness and generalizability²⁴.

Study 2C

1011 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=20) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=4), were excluded from all analyses. Participant exclusions resulted in a final sample of 987 participants.

Study 2C was procedurally similar to Study 2A–2B, but participants were assigned to one of four betweenparticipant conditions, which varied the race of the two targets to which participants were exposed (Asian, Black, multiracial, or White). Given that the Chicago Face Database does not contain multiple unique images of non-White individuals, for this study the faces were obtained from a different stimulus set⁶⁵. Four male individuals were selected for inclusion from each racial category. Each individual was depicted from five different angles (45° left, 30° left, frontal, 30° right, and 45° right). In addition, at the end of the learning phase, participants responded to four manipulation check items to ensure that they encoded the targets' racial category membership as intended and they accurately remembered which target was paired with which label(s) in the attribute conditioning procedure.

Study 2D

1213 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete the IAT (n=15) and participants with response latencies of 300 ms or below on at least 10% of IAT trials, indicating inattention (n=7), were excluded from all analyses. Participant exclusions resulted in a final sample of 1191 participants.

Study 2D was procedurally identical to Study 2C, with the following exception: Given participants' poor performance on the race manipulation check item in Study 2C, Study 2D featured a newly added racial categorization task designed to explicitly teach participants the two main targets' racial group membership. This task was administered at the beginning of the study, prior to the attribute conditioning manipulation. Beyond two main targets, six additional individuals (three from each non-focal racial category) were included. On each trial, participants were presented with a face, which they were asked to categorize as Asian, Black, multiracial, or White. Incorrect responses triggered a red X; participants were required to enter the correct response before they were allowed to proceed to the next trial. The categorization task included 25 trials for each racial group, resulting in a total of 100 trials. Trials were presented in an individually randomized order.

Study 3A

1068 participants with U.S. citizenship and U.S. residence were recruited from Project Implicit. Participants who did not complete both IATs (n=65) and participants with response latencies of 300 ms or below on at least 10% of trials on either IAT, indicating inattention (n=37), were excluded from all analyses. Participant exclusions resulted in a final sample of 966 participants.

The goals of Study 3A were to establish the individual-level predictive validity of the American/foreigngood/bad evaluative IAT used in Studies 1A–2D and to probe its relationship with the White American–Asian American/American–foreign stereotype IAT used in the archival analysis in Study 3B below. The former IAT used the same procedure and stimuli as the IATs in Studies 1A–2D; the latter IAT used the same procedure as the IATs in previous studies and the same category and attribute stimuli as the archival IAT used in Study 3B. Specifically, the category labels were "European American" and "Asian American." For each category, the category stimuli were three female and three male faces, randomly selected from a larger set of 12 faces. The attribute labels were "American" and "Foreign," and the attribute stimuli included images of a map of the United States, a 20-dollar bill, the Great Seal of the United States, the U.S. flag, a five-cent coin, and a U.S. passport, and images of a map of Belgium, a 20-hryvnia bill, the coat of arms of Flanders, the flag of Djibouti, a five-baht coin, and a Canadian passport, respectively.

Participants completed three sets of measures in randomized order: (*a*) the two IATs described above (in randomized order), (*b*) a 12-item scale measuring participants' immigration policy preferences (Cronbach's α =0.89), with items drawn from the ballot initiatives used in Study 3B and rewritten such that 6 items were formulated in a pro-immigrant and 6 items in an anti-immigrant direction; (*c*) a set of six self-report items, including three evaluation items (feeling thermometers toward Americans and non-Americans and a relative preference item comparing Americans and non-Americans) and three stereotype items (measures of explicit association of Asian Americans and White Americans with Americans set of what extent the participant believes Asian Americans and White Americans to be American).

Given our exclusive interest in the correlation between the scores on the two IATs and their relationships with other variables, unlike in prior studies, we did not randomize the order of critical blocks. Rather, all participants completed the congruent (American–good/foreign–bad; White American–American/Asian American–foreign)

blocks first on both IATs. However, similar to Study 1A, we randomized the category labels (but not the category stimuli) for the non-American category between "alien," "foreigner," and "noncitizen" on the evaluative IAT.

Study 3B

For Study 3B, we identified a comprehensive set of 18 relevant ballot initiatives from an online database⁶⁶. All ballot initiatives occurred in the United States between 1994 and 2022 and concerned questions related to attitudes toward non-Americans, such as whether driver's license applications should require proof of citizenship or immigration status, and whether foreigners not eligible for U.S. citizenship should be prohibited from owning property. The county-level results of each ballot initiative were then obtained from the websites of the respective secretaries of state and standardized in format. The dependent measure was the share of anti-immigrant votes at the county level. Data from a total of 40,488,991 voters were included.

We then used archival data collected from the Project Implicit website between 2006 and 2024 to obtain county-level estimates of bias using a subset of the measures from Study 3A. The White American/Asian American/American-foreign stereotype IAT (which used the same stimuli and followed the same procedure as described under Study 3A above but consisted of seven, rather than five, blocks) served as a measure of implicit bias, and the difference between the two self-reported association items served as a measure of explicit bias. Both measures were scored in such a way that higher scores indicate stronger White-American/Asian-foreign bias.

We removed participants from outside the United States and those without location data and cleaned the data using standard cleaning procedures. We geolocated participants based on county FIPS and calculated the mean implicit and explicit bias for each relevant county. No cutoffs for county-level sample size were applied; rather, we explicitly probed the effects of county-level sample size in moderator analyses. Given concerns about reidentifiability⁶⁷, individual-level data are available from the first author only after showing proof of IRB approval; county-level data are openly shared on the Open Science Framework. Data from a final sample of 193,649 participants were included.

Finally, we fit three regression models to the data of each ballot initiative separately, one predicting countylevel anti-immigrant vote share from implicit bias only, one from explicit bias only, and one from both measures simultaneously. We then used the standardized regression coefficients with their associated standard errors as effect sizes to be meta-analyzed in three separate fixed-effects meta-analytic models. The results of these metaanalytic models, along with further models including methodological moderators of the meta-analytic effect of implicit bias on anti-immigrant vote share, are reported above.

Data availability

All materials, data, and analysis code are available from the Open Science Framework (https://osf.io/tqhn7/). The design and statistical analyses of Studies 1A, 1B, 1C, 2B, 2C, 2D, and 3A were preregistered, whereas Studies 2A and 3B were not. The links to each preregistration document are listed on the Open Science Framework.

Received: 16 December 2024; Accepted: 11 April 2025 Published online: 30 April 2025

References

- 1. Zangwill, I. The Melting-Pot (CreateSpace Independent Publishing Platform, 2016).
- Pew Research Center. Key findings about U.S. immigrants. https://www.pewresearch.org/short-reads/2020/08/20/key-findings-ab out-u-s-immigrants/ (2020).
- 3. Painter, M. A. & Qian, Z. Wealth inequality among immigrants: Consistent racial/ethnic inequality in the United States. *Popul. Res. Polic. Rev.* 35, 147–175 (2016).
- Iceland, J. & Scopilliti, M. Immigrant residential segregation in U.S. metropolitan areas, 1990–2000. *Demography* 45, 79–94 (2008).
 McCann, W. S. & Boateng, F. D. An analysis of hate crime victimization amongst immigrants. *Am. J. Crim Justice*. 47, 795–817 (2022).
- Skinner-Dorkenoo, A. L., George, M., Wages, J. E., Sánchez, S. & Perry, S. P. A systemic approach to the psychology of racial bias within individuals and society. Nat. Rev. Psychol. 1–15 https://doi.org/10.1038/s44159-023-00190-z (2023).
- 7. Banaji, M. R., Fiske, S. T. & Massey, D. S. Systemic racism: individuals and interactions, institutions and society. Cogn. Res. : Princ Implic. 6, 82 (2021).
- 8. Kurdi, B. & Charlesworth, T. E. S. A 3D framework of implicit attitude change. Trends Cogn. Sci. 27, 745–758 (2023).
- 9. Gawronski, B. & Bodenhausen, G. V. Beyond persons and situations: An interactionist approach to understanding implicit bias. *Psychol. Ing.* **28**, 268–272 (2017).
- Rose, J. Immigration agencies ordered not to use term illegal alien under new Biden policy. Natl. Public. Radio https://www.npr.or g/2021/04/19/988789487/immigration-agencies-ordered-not-to-use-term-illegal-alien-under-new-biden-polic (2021).
- 11. Oladipo, G. Dehumanizing, inaccurate and outdated: why did Biden say 'illegals' in his State of the Union address? *The Guardian* (2024). https://www.theguardian.com/us-news/2024/mar/08/biden-illegals-state-of-the-union-speech
- Morrison, K. R. & Chung, A. H. White or European American? Self-identifying labels influence majority group members' interethnic attitudes. J. Exp. Soc. Psychol. 47, 165–170 (2011).
- 13. Rios, K. Right-wing authoritarianism predicts prejudice against homosexuals but not gay men and lesbians. J. Exp. Soc. Psychol. 49, 1177–1183 (2013).
- 14. Rios, K. & Mischkowski, D. Shaping responses to torture: What you call it matters. Pers. Soc. Psychol. Bull. 45, 934-946 (2018).
- 15. Rios, K. & Ingraffia, Z. A. Judging the actions of whistle-blowers versus leakers: Labels influence perceptions of dissenters who
- expose group misconduct. *Group. Process. Intergroup Relat.* **19**, 553–569 (2016). 16. Rucker, J. M., Murphy, M. C. & Quintanilla, V. D. The immigrant labeling effect: The role of immigrant group labels in prejudice
- against noncitizens. *Group. Processes Intergroup Relations*. 22, 1139–1160 (2019).
 17. Knoll, B. R., Redlawsk, D. P. & Sanborn, H. Framing labels and immigration policy attitudes in the Iowa caucuses: Trying to Out-Tancredo Tancredo. *Polit. Behav.* 33, 433–454 (2010).
- Ommundsen, R., Larsen, K. S., van der Veer, K. & Eilertsen, D. E. Framing unauthorized immigrants: The effects of labels on evaluations. *Psychol. Rep.* 114, 461–478 (2014).
- 19. Pearson, M. R. How undocumented workers and illegal aliens affect prejudice toward Mexican immigrants. Social Influence. 5, 118-132 (2010).

- Augoustinos, M. & Quinn, C. Social categorization and attitudinal evaluations: Illegal immigrants, refugees or asylum seekers? New. Rev. Social Psychol. 2, 29–37 (2003).
- Boucher, K. L. & Rydell, R. J. Impact of negation salience and cognitive resources on negation during attitude formation. Pers. Soc. Psychol. Bull. 38, 1329–1342 (2012).
- 22. Nosek, B. A. Moderators of the relationship between implicit and explicit evaluation. J. Exp. Psychol. Gen. 134, 565-584 (2005).
- Greenwald, A. G. & Banaji, M. R. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychol. Rev.* 102, 4–27 (1995).
 Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* 45, e1 (2020).
- Hukoni, J. Hie generation of rates being being
- intersectional theory primer. Soc. Pers. Psychol. Compass 18, (2024).
 26. Yogeeswaran, K. & Dasgupta, N. Will the real American please stand up? The effect of implicit national prototypes on discriminatory behavior and judgments. Pers. Soc. Psychol. Bull. 36, 1332–1345 (2010).
- Yogeeswaran, K., Gale, J., Hoffman, A. & Devos, T. Does county-level implicit national exclusion predict political participation among Asian Americans? *Cult. Divers. Ethn. Minor. Psychol.* https://doi.org/10.1037/cdp0000595 (2023).
- Conitzer, V. et al. American = White in multimodal language-and-image AI. Proc. 2022 AAAIACM Conf. AI Ethics Soc. 800–812 https://doi.org/10.1145/3514094.3534136 (2022).
- 29. Devos, T. & Banaji, M. R. American = White? J. Personal. Soc. Psychol. 88, 447-466 (2005).
- Devos, T. & Mohamed, H. Shades of American identity: Implicit relations between ethnic and national identities. Soc. Pers. Psychol. Compass. 8, 739–754 (2014).
- Lee, M. H. J., Montgomery, J. M. & Lai, C. K. America's racial framework of superiority and Americanness embedded in natural language. PNAS Nexus pgad485 (2024). https://doi.org/10.1093/pnasnexus/pgad485
- 32. Zou, L. X. & Cheryan, S. Two axes of subordination: A new model of racial position. J. Pers. Soc. Psychol. 112, 696–717 (2017).
- 33. Tessler, H., Choi, M. & Kao, G. The anxiety of being Asian American: Hate crimes and negative biases during the COVID-19 pandemic. Am. J. Crim Justice. 45, 636–646 (2020).
- 34. Kurdi, B. & Banaji, M. R. Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated: A commentary on Payne, Vuletich, and Lundberg. *Psychol. Inq.* **28**, 281–287 (2017).
- Payne, B. K., Vuletich, H. A. & Lundberg, K. B. The bias of crowds: How implicit bias bridges personal and systemic prejudice. Psychol. Ing. 28, 233–248 (2017).
- 36. Payne, B. K. & Hannay, J. W. Implicit bias reflects systemic racism. Trends Cogn. Sci. 25, 927-936 (2021).
- Hehman, E., Calanchini, J., Flake, J. K. & Leitner, J. B. Establishing construct validity evidence for regional measures of explicit and implicit racial bias. J. Exp. Psychol. Gen. 148, 1022–1040 (2019).
- Calanchini, J. et al. Regional intergroup bias. Adv. Exp. Soc. Psychol. 281–337. https://doi.org/10.1016/bs.aesp.2022.04.003 (2022).
 Piantadosi, S., Byar, D. P. & Green, S. B. The ecological fallacy. Am. J. Epidemiol. 127, 893–904 (1988).
- Partiadosi, S. Byar, D. F. & Green, S. D. In coordination, *P. phaemin.* 127, 055–064 (1966).
 Ratliff, K. A. et al. Documenting bias from 2007–2015: Pervasiveness and correlates of implicit attitudes and stereotypes II. 1–48 (2020).
- 41. Nosek, B. A. et al. Pervasiveness and correlates of implicit attitudes and stereotypes. Eur. Rev. Social Psychol. 18, 36-88 (2007).
- 42. Plant, E. A. & Devine, P. G. Internal and external motivation to respond without prejudice. J. Personal. Soc. Psychol. 75, 811-832 (1998).
- 43. Unkelbach, C. & Förderer, S. A model of attribute conditioning. Social Psychol. Bull. 13, 479-423 (2018).
- 44. Unkelbach, C. & Högden, F. Why does George Clooney make coffee sexy? The case for attribute conditioning. *Curr. Dir. Psychol. Sci.* 28, 540–546 (2019).
- McDonald, M. M., Navarrete, C. D. & Sidanius, J. Developing a theory of gendered prejudice: An evolutionary and social dominance perspective. in Social Cognition, Social Identity, and Intergroup Relations 189–220 (Psychology, doi:https://doi.org/1 0.4324/9780203816790-12. (2011).
- 46. Martinez, J. E., Oh, D. & Todorov, A. Immigrant faceism: Documentation statuses evoke racialized mental representations. (2021).
- 47. Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105 (1959).
- 48. Gawronski, B. Back to the future of dissonance theory: Cognitive consistency as a core motive. Soc. Cognition. 30, 652-668 (2012).
- 49. Spearman, C. The proof and measurement of association between two things. Am. J. Psychol. 15, 72-30 (1904).
- 50. Tajfel, H. & Turner, J. in An Integrative Theory of Intergroup Conflict. 33-47 (eds Austin, W. G. & Worchel, S.) (Brooks/Cole Publishing Company, 1979).
- Greenwald, A. G. & Pettigrew, T. F. With malice toward none and charity for some: Ingroup favoritism enables discrimination. Am. Psychol. 69, 669–684 (2014).
- 52. Shafer, B. E. American exceptionalism. Annu. Rev. Political Sci. 2, 445-463 (1999).
- Kurdi, B., Ratliff, K. A. & Cunningham, W. A. Can the implicit association test serve as a valid measure of automatic cognition? A response to Schimmack (2021). Perspect. Psychol. Sci. 16, 422–434 (2021).
- 54. Nisbett, R. E., Zukier, H. & Lemley, R. E. The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cogn. Psychol.* 13, 248–277 (1981).
- Kurdi, B., Mann, T. C., Charlesworth, T. E. S. & Banaji, M. R. The relationship between implicit intergroup attitudes and beliefs. Proc. Natl. Acad. Sci. 116, 5862–5871 (2019).
- Phills, C. E., Hahn, A. & Gawronski, B. The bidirectional causal relation between implicit stereotypes and implicit prejudice. Pers. Soc. Psychol. B. 46, 1318–1330 (2020).
- 57. Irving, L. H. & Smith, C. T. Measure what you are trying to predict: Applying the correspondence principle to the implicit association test. J. Exp. Soc. Psychol. 86, 103898-103814 (2020).
- Amodio, D. M. & Devine, P. G. Stereotyping and evaluation in implicit race bias: Evidence for independent constructs and unique effects on behavior. J. Personal. Soc. Psychol. 91, 652–661 (2006).
- 59. Paluck, E. L., Porat, R., Clark, C. S. & Green, D. P. Prejudice reduction: Progress and challenges. Ann. Rev. Psychol. 72, 533-560 (2021).
- 60. Ferguson, M. J., Shen, X., Cone, J. & Mann, T. C. How do we reduce implicit bias toward outgroups? In The Cambridge Handbook of Implicit Bias and Racism (eds Krosnick, J. A., Stark, T. H. & Scott, A. L.) (Cambridge University Press, Cambridge, UK, (2023).
- Kurdi, B., Mann, T. C., Axt, J. & Ferguson, M. J. The fragility of implicit attitude updating: The role of cognitive and ecological constraints. (2023). https://doi.org/10.31234/osf.io/mwfah
 C. M. C. M.
- 62. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. Measuring individual differences in implicit cognition: The implicit association test. J. Personal. Soc. Psychol. 74, 1464–1480 (1998).
- Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the implicit association test: I. An improved scoring algorithm. J. Personal. Soc. Psychol. 85, 197–216 (2003).
- Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods.* 47, 1122–1135 (2015).
- 65. Righi, G., Peissig, J. J. & Tarr, M. J. Recognizing disguised faces. Visual Cognition. 20, 143–169 (2012).
- 66. Ballotpedia. Immigration on the ballot. https://ballotpedia.org/Immigration_on_the_ballot#By_year
- Morehouse, K. N., Kurdi, B. & Nosek, B. A. Responsible data sharing: Identifying and remedying possible re-Identification of human participants. Am. Psychol. https://doi.org/10.1037/amp0001346 (2024).

Author contributions

B.K., K.O., E.H., and M.J.F. conceived and designed the studies; B.K. performed the studies; B.K. analyzed data; B.K. drafted and K.O., E.H., and M.J.F. revised the paper.

Declarations

Competing interest

B.K. is a member of the Scientific Advisory Board of Project Implicit, a 501(c)(3) non-profit organization and international collaborative of researchers who are interested in implicit social cognition. K.O., E.H., and M.J.F. have no competing interests to disclose.

Ethical approval

The project received ethical approval from the Yale University Institutional Review Board ("Social learning and memory", #2000029368) and the University of Illinois Urbana–Champaign Office for the Protection of Research Subjects ("Regional and individual determinants of attitudes and stereotypes", IRB23-0325). All research was performed in accordance with relevant guidelines and regulations. All participants in Studies 1–3A provided informed consent; Study 3B relied entirely on archival data.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/1 0.1038/s41598-025-98384-3.

Correspondence and requests for materials should be addressed to B.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025