PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Predicting implicit attitudes with natural language data

Sudeep Bhatia[a,1] and Lukasz Walasek[b]

Edited by Mahzarin Banaji, Harvard University, Cambridge, MA; received December 6, 2022; accepted May 8, 2023

Large-scale language datasets and advances in natural language processing offer opportunities for studying people's cognitions and behaviors. We show how representations derived from language can be combined with laboratory-based word norms to predict implicit attitudes for diverse concepts. Our approach achieves substantially higher correlations than existing methods. We also show that our approach is more predictive of implicit attitudes than are explicit attitudes, and that it captures variance in implicit attitudes that is largely unexplained by explicit attitudes. Overall, our results shed light on how implicit attitudes can be measured by combining standard psychological data with large-scale language data. In doing so, we pave the way for highly accurate computational modeling of what people think and feel about the world around them.

implicit attitudes | implicit association test | natural language processing | word embeddings | computational modeling

Significance

For decades, psychologists have studied attitudes with the goal of developing better techniques for measuring and predicting them. Here, we examine how large natural language datasets can be used to predict people's implicit attitudes, that is, automatic associations between a concept and an evaluation. We apply computational models that use the strength of association of words in language to proxy the associations in people's minds. We show that these models predict biases in implicit attitudes that cannot be explained using participant self-reports. We also develop a computational method that combines linguistic associations with existing psychological data, and show that it generates especially good predictions, thereby facilitating several practical applications.

Large-scale digital data have greatly improved our ability to model human cognition and behavior. One powerful application involves knowledge representations derived from natural language. By observing the distribution of words in language data, it is now possible to specify, for arbitrary linguistically encodable concepts, quantitative representations that capture what people know and associate with those concepts (1–3) (see ref. 4–7 for reviews). These distributed semantic representations (DSRs) typically take the form of vectors (also known as word embeddings) in high-dimensional semantic spaces. Since vectors in DSR models capture semantic representations, the similarity of vectors proxies the strength of association between any two concepts. This makes them useful for studying associative processes in memory (8–10) and judgment (11, 12), and subsequently implicit attitudes, which are the associations that people have between concepts and evaluations (13, 14).

With respect to implicit attitudes, researchers have shown that DSR-based measures of association generate the racial and gender biases observed in implicit association tests (IATs) (15). This finding has led to considerable further research, both in psychology and in fields like natural language processing and artificial intelligence, as it implies that everyday language contains the types of biases observed in experimental data, and that computational models trained on this language are biased in a human-like manner (16–21) (see ref. 22 for a review). Although biased computational models have far-reaching and harmful consequences for society, they provide both practical and theoretical utility for behavioral scientists. From a practical perspective, DSRs trained on different types of language corpora can be used to measure differences in bias across different cultural and linguistic groups. Additionally, by quantitatively predicting people's attitudes toward thousands of common concepts, DSR-based models can be used to develop interventions for mitigating harmful biases. Theoretically, the fact that DSR models possess human-like implicit attitudes implies that these attitudes are the product of the same mechanisms, and can be studied using the same theoretical tools, as cognitions and behaviors in other areas of psychology where DSRs have been shown to be useful. Overall, natural language provides a unique window into the elusive mental representations that underpin peoples' attitudes, beliefs, and preferences.

Despite these important implications, there is little research on how language biases can be used to accurately predict implicit attitudes. The dominant approach, known as the *word embedding association test* (WEAT), predicts implicit attitudes by measuring DSR-based associations between target concepts and evaluative words in the stimuli used in IATs (14) (see also ref. 13 for a related approach). Thus, for example, the implicit attitude toward target concepts like Flower and Insect in an IAT with evaluative attribute words like "lucky" and "disaster" (as in ref. 15) is predicted by 1) measuring the relative similarity of the word vector for "flower" with vectors for "lucky" vs. "disaster," 2) measuring the relative similarity of the vector for "insect" with vectors for "lucky" vs. "disaster," and 3) subtracting the first similarity from the second. DSR-based attitudes from this approach are moderately correlated with observed implicit attitudes in existing IATs (for example, there is a correlation of 0.26 between predicted and observed implicit attitudes for the four

Author affiliations: [a]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104-6018; and [b]Department of Psychology, University of Warwick CV4 7AL, Coventry, United Kingdom

Author contributions: S.B. and L.W. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: bhatiasu@sas.upenn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2220726120/-/DCSupplemental.

Published June 12, 2023.

evaluative IATs in table 1 of ref. 14). Although these moderate correlations are intriguing, it is possible that better predictions could be obtained by using other measurements of implicit attitudes in DSRs.

In this paper, we examine an alternative approach that relies on laboratory-based word valence norms collected by Warriner et al. (23). Valence norms—human ratings of emotional responses to common words—are widely used in psychological research, both to study the nature of emotions (e.g. ref. 24) and to analyze the effects of emotions on other psychological processes (e.g. ref. 25). We propose that by combining DSRs with valence norms, we can better proxy the associative evaluations that generate implicit biases than with DSRs alone. We explore two ways of using valence norms for implicit attitude prediction. The first reproduces the core calculations of WEAT but replaces IAT evaluative stimuli words with larger sets of positive and negative words obtained from valence norms. The second fits ratings for valence norm words on DSR vectors for those words, and then, with the best-fitting model, estimates the valence of concepts in IAT tasks using those concepts' DSR vectors. According to this *Valence Estimation Model* (VEM), the implicit attitude for a concept is simply a valence judgment made using the concept's distributed semantic representation. This valence judgment can be predicted using a DSR-based model that has been fit on valence norms data.

In addition to considering several ways of extracting attitudes from language, we also examine whether these attitudes are better predictors of IAT scores or explicitly held attitudes, and additionally if they capture variance in IAT scores that is separate from that captured by explicit attitude measures. We do so with a large dataset of both implicit and explicit attitude measures released by Project Implicit (26). Prior work has not compared language biases on implicit vs. explicit measures, and it is not known whether DSRs are able to add any additional predictive power to the modeling of implicit attitudes on top of explicit attitudes. Likewise, prior work has not explored the types of concepts for which DSRs make good implicit attitude predictions. Thus, our tests are necessary to determine not only the best ways of predicting attitudes with language, but also the psychological properties of language-based attitudes, and their relationship with established behavioral measures of attitudes.

Results

Overview of Dataset and Models. The data used in the present paper were collected as part of the attitudes, identities, and individual differences study (26), released by Project Implicit. The dataset includes approximately 200,000 complete responses on 95 unique evaluative IATs along with several measures of explicit attitudes corresponding to each IAT. Each IAT has two concepts, and measures the relative implicit attitude for one concept compared to the other. Examples of IATs include astrology – science, Denzel Washington – Tom Cruise, dogs – cats, gay people – straight people, Redsox – Yankees, television – books, and wrinkles – plastic surgery. The primary dependent variable in our analysis was the aggregate D score. This is the average participant response time difference in categorizing one concept (e.g., astrology) as positive vs. negative relative to the other concept (e.g., science), and allows us to capture population-level implicit attitudes toward the two concepts in each IAT task. We had 95 aggregate D scores, one for each of the 95 IATs. We also considered the 11 different self-reported explicit attitude measures in the Project Implicit dataset. These include explicit *personal* evaluative judgments for the concepts in the IATs (e.g., how much the participant likes or dislikes astrology vs. science). They also include explicit assessments of *social and cultural* attitudes for the concepts (e.g., how much the participant thinks that others like or dislike astrology vs. science). The Project Implicit dataset also contains self-reported *meta-attitudinal* assessments of the concepts (e.g., the stability of the participant's feelings for astrology and science, the cultural pressure to think certain ways about astrology and science etc.), which we used to examine the types of words for which DSR-based models are particularly accurate. The self-reported explicit attitude and meta-attitudinal variables are summarized in *SI Appendix*, Table S1. Finally, in addition to these measures, the dataset contains a variety of demographic variables, which we used to assess the robustness of our results for different populations of participants.

In our primary analysis, we attempted to predict aggregate D scores for the IATs using DSRs, which are quantitative representations for words and concepts that are recovered based on word cooccurrence statistics in natural language. Although there are many approaches to extracting DSRs from language data, the two most prominent models are Word2Vec (2) and GloVe (3). We performed our primary analysis with a pretrained version of the former model and replicated it in robustness tests with a pretrained version of the latter model. We applied DSRs to the Project Implicit dataset using both the WEAT and the VEM, which are described in detail in the *Materials and Methods* section. The WEAT approach was implemented using the original IAT evaluative attribute words (15). By contrast, VEM was implemented by training a ridge regression model to predict the valence rating of words in Warriner et al.'s valence norms data (23) using their DSR vectors. The best-fit regression model was then applied to the DSR vectors for the individual target concepts in the Project Implicit IATs to predict their individual valence ratings. The difference in the predicted ratings for the two target concepts in an IAT (e.g. the difference in the predicted valence rating for astrology vs. science) was taken as a measure of the relative implicit attitude for one concept over the other.

We also attempted several variants of the WEAT model that replaced the IAT evaluative stimuli words with positively or negatively valenced words from our valence norms data. For our WEAT-10 model, we obtained the ten most positively and ten most negatively valenced words in Warriner et al.'s data (23), and used them to calculate the WEAT scores. Our WEAT-100 and WEAT-1000 models used larger numbers of positively and negatively valenced words in the same way. Finally, the WEAT-Full model used the full dataset of valence ratings; that is, it split the valence norms dataset into to two equally sized sets corresponding to positive and negative words, and used relative vector similarities of these two sets to calculate the WEAT score.

We additionally tested several variants of the estimation approach, which replaced the valence norms dataset with other emotion norms datasets. For example, our arousal estimation model replaced valence ratings with arousal ratings in Warriner et al.'s data (23). Likewise, our anger, disgust, fear, happiness, sadness, and surprise estimation models used ratings for words on these six basic emotions in Mohammad and Turney's data (27). Our reason for applying this approach with other types of emotions was to test its specificity to the valence dimension. If implicit attitudes correspond to word valence (and not other emotional qualities of words), then we should expect our main VEM to outperform the estimation approach applied to other emotional qualities.

Predicting Implicit Attitudes. We began by examining the performance of our Word2Vec DSR metrics in predicting IAT scores in the Project Implicit dataset. The aggregate D score for

the IATs was our dependent variable. As predictors, we considered 13 (5 WEAT-based and 8 estimation-based) DSR metrics, each of which served as a predictor variable in a separate regression. For example, our first regression specified $D_i = \beta_0 + \beta_1 \cdot WEAT_i + \varepsilon$, where D_i is the implicit attitude score in the i^{th} IAT test and $WEAT_i$ is the original WEAT model's prediction for this test. The R^2 values from these 13 regressions are displayed in Fig. 1*A*, and the outputs of the full regressions are shown in *SI Appendix*, Table S2. Here, we see that the original WEAT approach performed well, generating $R^2 = 0.13$, corresponding to a positive correlation that surpasses the Bonferroni corrected threshold for significance ($P < 0.05/13 = 0.004$). The WEAT approach applied to only the ten most positive and negative words in ref. 13 performed poorly, though increasing the number of positive and negative words used in WEAT substantially improved its performance. The best performing WEAT model, WEAT-Full, divided the entire valence norms dataset into equal-sized positive and negative sets of words and achieved an $R^2 = 0.20$. Even better correlations were obtained using VEM, which achieved an $R^2 = 0.26$, doubling the R^2 from the original WEAT method. It is interesting to note that the improvement of VEM over the original WEAT method is also apparent when examining the four evaluative IAT tests in ref. 14 (insects–flowers, weapons–instruments, African American–European American names, and old people–young people names). Here, even though WEAT predicted the direction of all four effects successfully, it achieved only a moderate correlation of $r = 0.26$ (equivalent to $R^2 = 0.07$) in predicting the magnitude of the four effects. Repeating the analysis in ref. 14 with VEM more than doubles the correlation, to $r = 0.58$ (equivalent to $R^2 = 0.34$). Fig. 1*A* also shows that the estimation method applied to other emotional qualities (like arousal or distinct emotions) performed relatively poorly. However, as indicated in *SI Appendix*, Table S2, the effects are in the expected direction, with negative emotions (e.g., anger, fear, sadness) having a negative relationship with implicit attitudes.

In Fig. 2*A*, we present a scatter plot of aggregate D scores against VEM predictions for each of the IATs. Here, positive values on the X and Y axes correspond to positive VEM and implicit biases favoring concept B in IATs titled A – B (e.g. the skeptical – trusting IAT shows a bias favoring trusting). Fig. 2*A* reveals a strong positive correlation between the IAT D scores and our model's predictions [$r(91) = 0.51$; $P < 0.001$; 95% CI = [0.34, 0.65]]. That said, there are some outliers in Fig. 2*A*, such as the Pepsi—Coke IAT and the old people—young people IAT, for which model predictions are in the opposite direction to human data. Both of these IAT tests used visual, rather than lexical stimuli, and our attempt to model these IAT results using DSRs involved simply querying our model vocabularies for the phrases "pepsi," "coke," "old_people" and "young_people," along with their upper and title case variants (see *Materials and Methods* and *SI Appendix* for more details). This resulted in a prediction of more positive attitude for Pepsi over Coke and for old people over young people, which is the opposite of what was found in the Project Implicit data. We suspect that our Pepsi – Coke error is caused by disambiguation issues ("coke" refers to the beverage but also to the illegal drug and the coal-based fuel, both of which are negatively valenced), whereas our old people – young people error is likely a spurious prediction caused by using single low-frequency phrases to elicit model predictions. Indeed, when tested separately, we find that VEM achieves an $R^2 = 0.55$ for lexical stimuli but only $R^2 = 0.12$ for image stimuli. In Supplemental Information, we consider alternate methods for obtaining DSR vectors for concepts with visual IAT stimuli. However, since these methods were tested post hoc, we do not use these methods for the main analysis in this paper.

In Fig. 2*B* we present a scatter plot of VEM's predicted valence for the individual target concepts in each IAT. This scatter plot shows that there is a general positive relationship in the predicted implicit attitude for the concepts that make up each topic (e.g., concept pairs like friends—family are both given high predictions, whereas those like capital punishment—imprisonment are both given low predictions) [$r(91) = 0.53$; $P < 0.001$; 95% CI = [0.37, 0.66]]. That said, this figure also points out several concept pairs for where there are strong valence differences. These are pairs like skeptical—trusting and gay people—straight people, pairs for which we also observe a corresponding bias in the Project Implicit data.

Predicting Explicit Attitudes. Next, we replicated the above analysis for the explicit attitude measures in the Project Implicit dataset; that is, we attempted to predict explicit attitudes using DSRs. For simplicity, we performed this analysis only for VEM (the best-performing DSR metric). There are 11 explicit attitude measures, and as in the previous analysis, we used each of these measures as dependent variables in 11 separate regressions. The R^2 values from these 11 regressions are displayed in Fig. 1*B*, and the outputs of the full regressions are shown in *SI Appendix*, Table S3

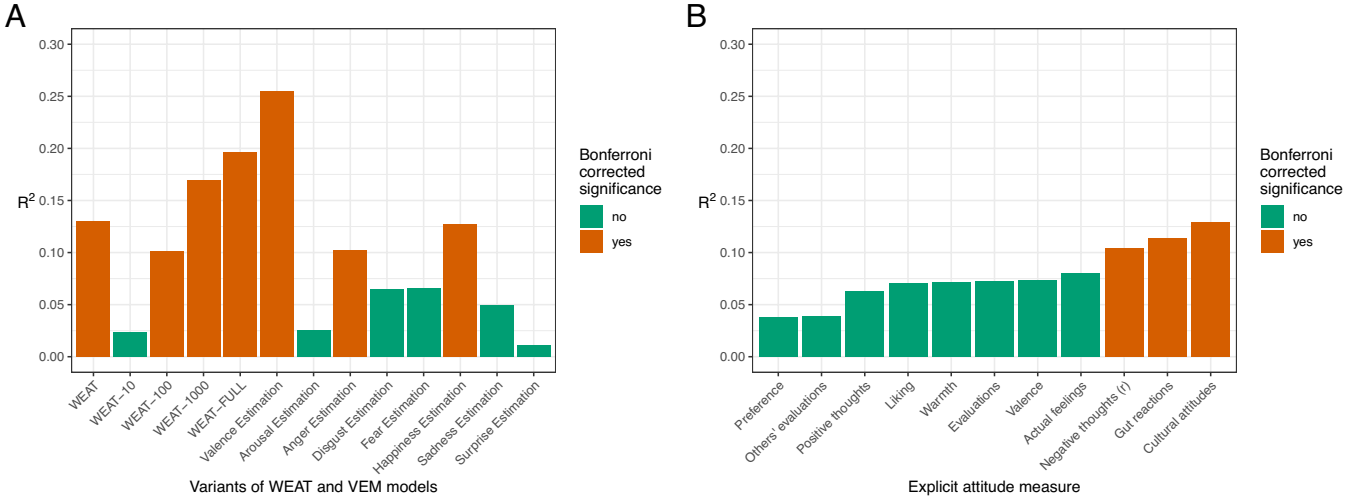

Fig. 1. (*A*) R^2 values for different variants of WEAT and estimation models used to predict aggregate IAT D scores. (*B*) R^2 values for the VEM used to predict explicit attitude measures associated with IATs. Tests that survive Bonferroni correction are indicated in orange. Note that the negative thoughts variable is reverse coded in panel *B*.

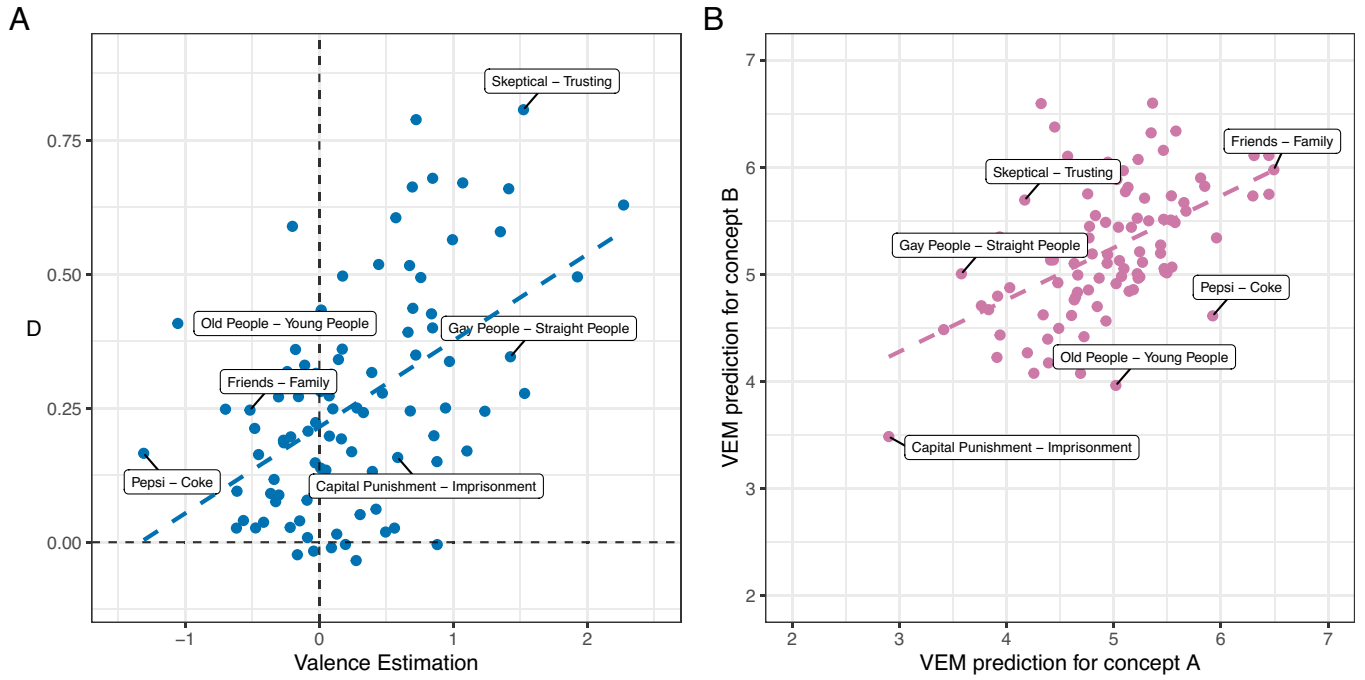

Fig. 2. (*A*) Predictions of the VEM and aggregate D scores for each of the IATs examined in this paper. (*B*) Predictions of the VEM for each of the individual target concepts in the IATs. Here, the *x* axis represents predictions for concept A and the *y* axis represents predictions for concept B, in IATs title A – B. The dotted lines indicate the best linear fit to the data.

(note that the labels for these metrics correspond to those used in the Project Implicit dataset, and further details about these variables are presented in *SI Appendix*, Table S1). Here, we can see that although VEM predictions were positively correlated with some explicit attitude measures, R^2*s* for these explicit attitude measures were much lower than those for D. In fact, only 3 of the 11 regressions crossed the Bonferroni corrected significance threshold ($P < 0.05/11 = 0.004$). These were for the gut feelings, negative thoughts, and cultural attitudes variables. The participants' beliefs about the relative cultural attitudes toward the two IAT concepts was the best predicted explicit attitude variable with an $R^2 = 0.13$ (roughly half of the R^2 achieved when predicting D).

Comparing DSRs with Explicit Attitude Measures. The above tests found that VEM captures about twice as much variance in implicit attitudes as existing WEAT metrics, and additionally, that it captures about twice as much variance in implicit attitudes as it does in explicit attitudes. We next examined whether VEM captured variance in implicit attitudes that was not predictable using explicit attitudes. For this, we ran two sets of regressions for each of the 11 explicit attitude variables. The first of these regressions attempted to predict D using only the explicit attitude variable as a predictor (e.g., *D* ~ gut feeling), whereas the second attempted to predict D using both the explicit attitude variable and the VEM score (e.g., *D* ~ gut feeling + VEM). The results of these two sets of regressions are shown in Fig. 3*A*, which reveals that the addition of VEM scores to each of the explicit attitude variables substantially improves predictions of D. The best performing model combines self-reported preference with VEM and achieves an $R^2 = 0.41$. Using self-reported preference alone achieves only $R^2 = 0.24$. *SI Appendix*, Table S4 provides the results of each of these regressions, and demonstrates that VEM was a significant predictor (with a Bonferroni corrected $P < 0.05/22 = 0.002$) in every regression.

We also repeated the above analysis in two additional regressions, which used all 11 explicit attitude measures as predictors at the same time. Here, we found that the regression model without VEM achieved an adjusted $R^2 = 0.26$, whereas the regression model with VEM achieved an adjusted $R^2 = 0.37$. In the latter test, VEM was the only variable that surpassed the Bonferroni-corrected threshold for significance ($P < 0.05/22 = 0.002$).

Factor Structure of Attitudes. The results of the previous section demonstrate that VEM captures variance in implicit attitudes that cannot be captured with explicit attitudes. Although this does suggest that associative biases in language uniquely track implicit attitudes, our previous analysis does not consider the full set of interrelationships between the variables. Thus, we attempted an exploratory factor analysis in which we decomposed the covariance of our VEM predictions, the implicit attitude (D) score, and the 11 explicit attitude variables, over the 95 different IATs in the Project Implicit dataset. This factor analysis was performed with varimax rotation and Kaiser normalization, and revealed a clean three-factor solution which captures 90.53% of the variance in the data. All variables had loadings above 0.7 and no variable loaded onto more than one factor. The factor loadings for the 13 variables in this analysis are shown in Fig. 3*B*. This figure shows that the first factor is uniquely composed of the nine personal explicit attitude variables, the second factor is uniquely composed of the two social/cultural explicit attitude variables, and the third factor is uniquely composed of VEM and D variables. This provides unequivocal evidence that associations in language obtained though VEM uniquely track implicit attitudes, and moreover that explicit attitudes can be separated into two factors: One corresponding to personally held attitudes, and the other corresponding to attitudes that are believed to be held by others.

Additional Tests. In our *SI Appendix* section, we report additional analyses designed to explore the determinants of VEM's performance, and to evaluate the robustness of its implicit attitude predictions. In summary, we show that VEM performs better at

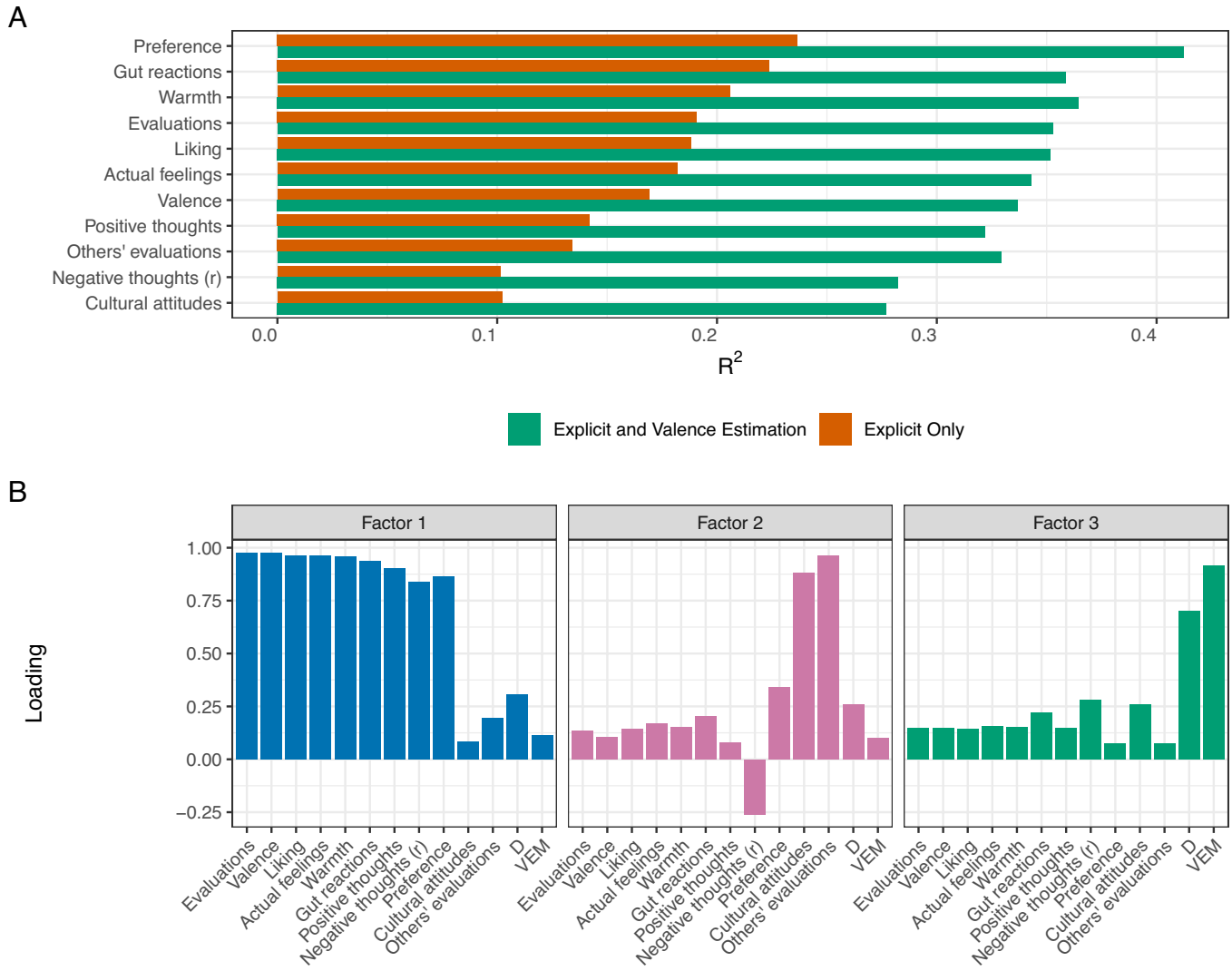

Fig. 3. (*A*) R^2 values from models predicting aggregate D scores using explicit attitude measures with and without the addition of VEM scores. (*B*) Factor loadings for VEM scores, along with implicit and explicit attitudes in the Project Implicit dataset. Note that the negative thoughts variable was reverse coded in both analyses.

predicting D for IATs for which there is high cultural pressure to think positively or negatively about the concepts (*SI Appendix*, Fig. S1*A*). We return to these results in our discussion section. We also show that our results replicated when we separately used VEM to predict D of participants who scored above or below the population median on ten demographic variables included in the Project Implicit dataset (*SI Appendix*, Fig. S1*B*). We also replicated the general pattern of results concerning VEM's performance with GloVe vectors (*SI Appendix*, Fig. S2*A*). Finally, we show that our predictive performance increased to R^2 = 0.32, when we supplemented our list of concepts and phrases for the visual IATs with participant-generated words (*SI Appendix*, Fig. S2*B*). This indicates that substantial improvements to our approach are possible with carefully curated word stimuli.

Discussion

Our results show that semantic representations derived from large-scale language data can accurately predict implicit attitudes for a large and diverse set of concepts, including trait words, people, places, social groups, brands, and abstract ideas. These results also illustrate the value of standard psychological data—in particular, valence norms—for modeling implicit attitudes. We show that these data significantly improve the correlations achieved by existing methods, like the WEAT. However, even better predictions are possible using a different approach, the VEM, which predicts the valence of words using a statistical model applied to their word vectors. By combing DRSs with valence norms, VEM doubles the amount of variance in implicit attitudes that can be explained by language biases, relative to previous work. These results persisted when we fit separate models to different subsets of participants split on key demographics, as well as when we used alternative DSR models.

We also tested the interrelationships between language biases, implicit attitudes, and explicit attitudes. First, we show that language biases predict implicit attitudes better than any individual explicit attitude measure. More rigorous analysis shows that language biases explain variance in implicit attitudes that cannot be accounted for by explicit attitude measures. We also found that language biases are much more correlated with implicit attitudes than they are with explicit attitudes. Finally, and perhaps most importantly, a factor analysis revealed that language biases and implicit attitudes load onto the same factor, which is different to the factors corresponding to personally held or culturally held explicit attitudes. This provides strong evidence that language biases uniquely track implicit (and not explicit) attitudes.

Our results also have implications for our understanding of the psychological underpinnings of implicit attitudes and the way they relate to biases in language. For example, we found that DSR models trained on valence norms provide a much better prediction of implicit attitudes than similar models trained on other types of emotion norms (such as arousal, anger, etc.). This is consistent with theories that propose that implicit attitudes are the outcomes of associative processes that drive the accessibility of valence-based content (28–31), as well as with recent empirical work that shows that IAT results are largely invariant to the specific valence words used as stimuli (32). However, this result does raise an interesting question: Why do valence norms, which are explicit participant ratings of word valence, predict implicit attitudes for IAT stimuli much better than they predict explicit valence judgments for IAT stimuli? The answer to that is that the valence norms stimuli, on which VEM is trained, involve common (noncontroversial) words, for which both implicit and explicit valence judgments are aligned. In the case of IAT stimuli, for which implicit and explicit valence assessments diverge (28–31), VEM continues to track implicit valence judgments, which like DSRs, reflect the associations inherent in language.

In *SI Appendix*, we show that language biases are much better predictors of implicit attitudes when there is cultural pressure to think or evaluate the concepts in a certain way. Consistent with this result, we note that the Project Implicit cultural attitude variable (which captures participants' explicit beliefs about cultural attitudes toward concepts) is the explicit attitude measure that is most correlated with language bias. These two results are consistent with theories that argue that implicit attitudes are reflective of biases and stigmas inherent in culture (30, 33, 34). Participants may not be able to explicitly describe these cultural biases and may be unaware that these biases even exist. But cultural biases are nonetheless reflected in language, which is why language (even more so than explicit beliefs about cultural attitudes) is such a good predictor of implicit attitudes.

The nuanced arguments advanced in the previous two paragraphs can be formalized and tested using more rigorous statistical tools like structural equation models. Although we have not done so in the present paper, due to power issues [95 observations, with several predictor and mediating variables, is not enough to test complex causal models (35)], we believe that doing so should be the focus of future research. This will require both implicit and explicit attitude data on a larger set of concepts, as well as potentially DSR models trained on different types of language data.

Future work should also explore ways to improve the predictive performance of our models. After all, even though we have greatly improved upon the correlations reported in prior work, a lot of the variation in people's IAT scores remains unexplained. In *SI Appendix* section, we have shown that predictive performance can be greatly increased if we ask participants to generate word sets for image IATs. A similar improvement could be possible if we ask participants to list the synonyms and hyponyms of lexical IAT concepts. In fact, there are already established psychological datasets of word associations and semantic relations (36, 37). As with valence norms, combining these datasets with DSRs derived from language has the potential to advance our ability to model people's attitudes.

In recent years, IAT has come under scrutiny by scientists who challenge its validity as a predictor of behavior, and indeed challenge the theoretical distinction between implicit and explicit attitudes (38, 39); see also refs. 40 and 41 for responses. Our analysis applies only to the aggregate magnitude of the bias obtained from IATs, and therefore our results do not speak to the validity of IAT as a measure of individual differences, which is the focus of much of this criticism. This does not undermine the significance of our findings, however, as group-level differences in the strength of the IAT bias for concepts are well-established and can have important societal consequences (40). Our findings that IAT scores load onto the same factor as language biases (a factor that is distinct from both personally held, and beliefs about culturally held explicit attitudes) also shows an important real-world correlate of implicit attitudes that cannot be predicted by explicit attitudes (addressing some of the challenges of refs. 38 and 39).

The success of our approach corroborates a growing body of research that shows how semantic representations based on large volumes of language data can be used to predict cognitions and behaviors. Earlier versions of this technique (1–3, 7, 9–12, 16–22, 42) (see ref. 4–6 for a review) use the relative similarities of DSR word vectors to proxy people's associations with various concepts. As with WEAT, these methods achieve a moderate level of performance in predicting participant responses. More recent work has refined DSR predictions with the use of laboratory-based human data. In this work, a small set of human responses are used to train DSR models to predict (out-of-sample) responses for arbitrary words and concepts (8, 43–46) (see ref. 47 for a review). Like the VEM model, this technique uses a *pretrained* language model to specify the underlying representations and associations for thousands of words, but *fine-tunes* these representations to accurately model a psychological variable of interest. By doing so, it is able to achieve far better predictions than earlier vector similarity methods. Indeed, it is this combination of pretraining and fine-tuning that is responsible for many of the recent successes of language models in natural language processing and artificial intelligence.

Ultimately, by developing and testing the VEM approach, we offer a predictive tool that greatly exceeds the performance of previous DSR methods and explicit attitude measures, for modeling implicit attitudes. In this way, we facilitate many practical applications. For example, our approach could be used to predict, in an a-priori manner, IAT scores for millions of concept pairs. These scores could be used as covariates in research in which it is cost prohibitive to obtain IAT scores for underlying concepts. These scores could also be used to identify and correct implicit biases in organizational and social settings. In a similar way, our approach could also be used to advance fair and nondiscriminatory AI technologies. AI can perpetuate discrimination when underlying models are trained on biased data. VEM allows us to quantify such a bias in a much more accurate manner than previous methods. Finally, unlike prior methods for extracting implicit attitudes from DSRs, VEM is not constrained by experimenter's decisions about the words to use when calculating association. As such, our approach circumvents any preexisting beliefs about the nature of the bias, which may be colored by the (implicit) biases held by the researchers.

Taken together, our paper reveals that human language uniquely tracks implicit biases, and that computational models applied to large-scale language data can be used to predict implicit biases. One may wonder, what is the added value of finding better models of implicit attitudes based on DSRs? Here, we echo the conclusions drawn by ref. 22, who point out that language that is filled with harmful prejudices and false stereotypes will contribute toward the propagation of these attitudes in society. It is therefore crucial that existing biases are better understood and more accurately measured if we are to develop better theories and methods for minimizing their impact. Our paper is one step toward this goal.

Materials and Methods

Variables Used in the Main Analysis. We obtained the aggregate D scores by simply averaging the individual-level D scores for each of the 95 IATs in the Project Implicit dataset (26). This gave us 95 aggregate D scores, which we attempted to predict using associations obtained from DSRs, using different variants of the WEAT and VEM methods (see below for details). There are also 11 explicit attitude measures in the Project Implicit data. For 9 of these (gut reactions, actual feelings, valence, warmth, liking, evaluations, positive thoughts, negative thoughts, and cultural attitudes), the Project Implicit dataset provides each participant's difference in ratings for the two concepts (e.g., difference in liking ratings). We averaged the difference variables across all participants. Two of the explicit attitude measures, (self) preference and other's evaluations, were comparative measures, without an associated difference variable. These were also averaged across participants. Finally, we reversed the negative thoughts variable prior to our analysis. We also used various meta-attitudinal and demographic variables for robustness tests, details of which are provided in *SI Appendix*.

Word Embedding Models. The Word2Vec model (2) used in this paper was pretrained on a very large corpus of news articles, and has 300 dimensional representations for over 3 million words and phrases. This model is especially useful for our tests as many of the concepts in the IAT stimuli are associated with multiword phrases (e.g., "Burger King," "Hillary Clinton," "free will," and "New York"). The model's vast vocabulary contains many of these concepts, and thus the model can be used to make predictions for most IAT tests in the Project Implicit dataset. We also replicated our main results with a GloVe model (3) in *SI Appendix*. Note that both Word2Vec and GloVe have been found to be useful for predicting cognitions and behaviors in human participants, and both are therefore suitable models for the study of implicit attitudes (4).

Valence Norms. We used a dataset of valence norms collected by Warriner et al. (23). This dataset has ratings for 13,915 words on a scale of 1 to 9 (with higher ratings corresponding to more positive valence). The most positively valenced words in this dataset are "vacation" and "happiness," whereas the most negatively valenced words are "pedophile" and "rapist." This dataset also has arousal norms, with higher ratings corresponding to words with higher arousal. The most arousing words in this dataset are "insanity" and "gun," whereas the least arousing words in this dataset are "grain" and "calm." Finally, we used an emotion norms dataset collected by Mohammad and Turney (27), with ratings for 14,182 words on six emotions: anger, disgust, fear, happiness, sadness, and surprise. These ratings are binary, with words pertaining to a given emotion being given a rating of 1 (and other words rated as 0).

WEAT. Each word (or phrase) in the Word2Vec and GloVe models' vocabulary is a point in a 300 dimensional semantic space. Thus, word i can be written as a vector $\boldsymbol{w_i} = [w_{i1}, w_{i2}, \ldots, w_{i300}]$. The similarity between the vectors for words proxies word association. Although there are many different similarity metrics, the most prominent metric (and the one that we use in our paper) is cosine similarity, which is the cosine of the angle of the words' vectors. With this metric, we can write the association between words i and j as:

$$s(word_i, word_j) = cossim(\boldsymbol{w_i}, \boldsymbol{w_j}) = (\boldsymbol{w_i} \cdot \boldsymbol{w_j})/||\boldsymbol{w_i}|| \cdot ||\boldsymbol{w_j}|| \quad [\mathbf{1}]$$

Cosine similarity is bounded between −1 and +1. Thus, words with highly positive cosine similarities with each other are considered to be positively associated, whereas words with highly negative cosine similarities with each other are considered to be negatively associated.

The cosine similarity–based word association measure can be used for several applications, including the measurement of implicit attitudes. For example, ref. 14 have used cosine similarity in the WEAT, which measures the relative association of two concepts with positively and negatively valenced words. For concept words X and Y and positively and negatively valenced words A and B, WEAT specifies the implicit association as:

$$WEAT(X, Y, A, B) = AVE_{x \in X}[S(x, A, B)] - AVE_{y \in Y}[S(y, A, B)] \quad [\mathbf{2}]$$

with:

$$S(z, A, B)] = AVE_{a \in A}[s(z, a)] - AVE_{b \in B}[s(z, b)] \quad [\mathbf{3}]$$

Intuitively, $WEAT(X, Y, A, B)$ is the average similarity of words in X with words in A vs. B, minus the average similarity of words in Y with words in A vs. B. Since A consists of positively valenced words and B consists of negatively valenced words, a positive WEAT score implies that words in X have more positive valence associations than words in Y. As discussed above, the WEAT measurement of attitudes has been shown to predict the direction of attitudes observed in several IATs. A related measure, used in ref. 13, gives similar results. Note that we did not normalize the similarity differences in Eq. **3** by dividing them by the SD. This has been used in ref. 14 for some statistical tests, however, we found that doing so for our analysis obscures cross-IAT variability in the data and makes it hard to compare the implicit attitude in one IAT test to another.

In our main implementation of WEAT, we used the original IAT evaluative stimuli words from Appendix A of ref. 15. However, we also replicated our analysis with evaluative stimuli words taken from Warriner et al.'s valence norms data (23). For each of these alternate models, we used the top N and bottom N rated words to specify A and B in Eqs. **2** and **3**. There were four such models, WEAT-10, WEAT-100, WEAT-1000, and WEAT-Full, using $N = 10$, $N = 100$, $N = 1{,}000$ and $N = 6{,}957$ words, respectively. The last of these used the full dataset of words in Warriner et al.'s data (23) (with a median split on ratings to classify a word as positive or negative).

VEM. The VEM attempts to predict the valence of each of the words in an IAT's stimuli set, and subsequently uses these predictions to estimate the valence of the associated constructs. Formally, this is done by training a regression model on Warriner et al.'s valence norms data (23), with valence ratings as a dependent variable and DSR vector dimensions as predictor variables. We can write the 300-dimensional vector representation of a word i as $\boldsymbol{w_i}$ and its valence rating as V_i. VEM proposes $V_i = \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{w_i}$, and attempts to find the best fitting β_0 and $\boldsymbol{\beta}$ for the 13,915 words in the valence norms data. Due to the high dimensionality of the word vectors, we used a ridge regression instead of a standard linear regression. Prior work has found that such regression models do a good job at predicting valence ratings for new words (48, 49) (also see ref. 43–46). Indeed, we verified that the above approach was able to predict valence ratings accurately using a 10-fold cross validation exercise. This revealed an average out-of-sample $R^2 = 0.62$ on the valence norms dataset. This corresponds to a correlation of $r = 0.79$, which is higher than the correlation achieved using the WEAT metric on Warriner et al.'s dataset (50). The superiority of VEM over WEAT in predicting valence ratings indicates that VEM may also provide a better account of implicit attitudes.

After fitting β_0 and $\boldsymbol{\beta}$ on the full valence norms data, we applied VEM to the concepts in the Project Implicit IAT tasks. For an IAT with two sets of target words X and Y, the VEM approach specifies the implicit association as:

$$VEM(X, Y) = AVE_{x \in X}[V(x)] - AVE_{y \in Y}[V(y)] \quad [\mathbf{4}]$$

with:

$$V(word_i) = \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{w_i} \quad [\mathbf{5}]$$

We replicated VEM with the arousal ratings in Warriner et al.'s data (23) and emotion ratings in Mohammad and Turney's data (27). The ratings in ref. 27 are binary, so we used a logistic regression (with L2 regularization) to fit the model, and used the logarithm of predicted probability estimates to capture the (continuous) emotion prediction for a concept.

Data, Materials, and Software Availability. Raw Survey data have been deposited in OSF (https://osf.io/xwavp/). Previously published data were used for this work (https://osf.io/gvzwm/).

ACKNOWLEDGMENTS. Funding was received from the NSF grant SES-1847794 and from a grant from The Honesty Project at Wake Forest University and the John Templeton Foundation.

1. T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211 (1997).
2. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" in *Advances in Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, 2013).
3. J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2014).
4. S. Bhatia, R. Richie, W. Zou, Distributed semantic representations for modelling human judgment. *Curr. Opin. Behav. Sci.* **29**, 31–36 (2019).
5. F. Günther, L. Rinaldi, M. Marelli, Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspect. Psychol. Sci.* **14**, 1006–1033 (2019).
6. A. Lenci, Distributional models of word meaning. *Annu. Rev. Linguist.* **4**, 151–171 (2018).
7. P. Mandera, E. Keuleers, M. Brysbaert, Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *J. Mem. Lang.* **92**, 57–78 (2017).
8. R. Richie, A. Aka, S. Bhatia, Free association in a neural network. *Psychol. Rev.*, 10.1037/rev0000396 (2022).
9. T. T. Hills, M. N. Jones, P. M. Todd, Optimal foraging in semantic memory. *Psychol. Rev.* **119**, 431 (2012).
10. M. W. Howard, M. J. Kahana, When does semantic similarity help episodic retrieval? *J. Mem. Lang.* **46**, 85–98 (2002).
11. S. Bhatia, Associative judgment and vector space semantics. *Psychol. Rev.* **124**, 1 (2017).
12. S. Bhatia, L. Walasek, Association and response accuracy in the wild. *Mem. Cognit.* **47**, 292–298 (2019).
13. S. Bhatia, The semantic representation of prejudice and stereotypes. *Cognition* **164**, 46–60 (2017).
14. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
15. A. Greenwald, A. G. McGhee, D. E. Schwartz, L. K. Jordan, Measuring individual differences in implicit cognition: The implicit association test. *J. Pers. Soc. Psychol.* **74**, 1464–1480 (1998).
16. N. Bhatia, S. Bhatia, Changes in gender stereotypes over time: A computational analysis. *Psychol. Women Q.* **41**, 106–125 (2021).
17. T. E. Charlesworth, V. Yang, T. C. Mann, B. Kurdi, M. R. Banaji, 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 (2021).
18. D. DeFranza, H. Mishra, A. Mishra, How language shapes prejudice against women: An examination across 45 world languages. *J. Personality Soc. Psychol.* **119**, 7 (2020).
19. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
20. B. Kurdi, T. C. Mann, T. E. Charlesworth, M. R. Banaji, The relationship between implicit intergroup attitudes and beliefs. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5862–5871 (2019).
21. M. Lewis, G. Lupyan, Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat. Hum. Behav.* **4**, 1021–1028 (2020).
22. T. E. Charlesworth, M. R. Banaji, "Word embeddings reveal social group attitudes and stereotypes in large language corpora" in *Atlas of Language Analysis in Psychology* (2021).
23. A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **45**, 1191–1207 (2013).
24. E. Verona, J. Sprague, N. Sadeh, Inhibitory control and negative emotional processing in psychopathy and antisocial personality disorder. *J. Abnormal Psychol.* **121**, 498 (2012).
25. S. T. Kousta, D. P. Vinson, G. Vigliocco, Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition* **112**, 473–481 (2009).
26. I. Hussey, S. Hughes, B. A. Nosek, The implicit and explicit Attitudes, Identities, and Individual Differences (AIID) Dataset. https://osf.io/pcjwf/ (2018)
27. S. M. Mohammad, P. D. Turney, Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **29**, 436–465 (2013).
28. A. G. Greenwald, M. R. Banaji, Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychol. Rev.* **102**, 4 (1995).
29. T. D. Wilson, S. Lindsey, T. Y. Schooler, A model of dual attitudes. *Psychol. Rev.* **107**, 101 (2000).
30. L. A. Rudman, Sources of implicit attitudes. *Curr. Dir. Psychol. Sci.* **13**, 79–82 (2004).
31. B. Gawronski, G. V. Bodenhausen, Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychol. Bulletin* **132**, 692–731 (2006).
32. J. R. Axt, T. Y. Feng, Y. Bar-Anan, The good and the bad: Are some attribute words better than others in the Implicit Association Test? *Behav. Res. Methods* **53**, 2512–2527 (2021).
33. B. K. Payne, H. A. Vuletich, K. B. Lundberg, The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychol. Inquiry* **28**, 233–248 (2017).
34. H. R. Arkes, P. E. Tetlock, Attributions of implicit prejudice, or "would Jesse Jackson 'fail' the Implicit Association Test?" *Psychol. Inquiry* **15**, 257–278 (2004).
35. D. P. MacKinnon, C. M. Lockwood, J. M. Hoffman, S. G. West, V. Sheets, A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* **7**, 83 (2002).
36. G. A. Miller, WordNet: A lexical database for English. *Commun. ACM* **38**, 39–41 (1995).
37. S. De Deyne, D. J. Navarro, A. Perfors, M. Brysbaert, G. Storms, The "Small World of Words" English word association norms for over 12,000 cue words. *Behav. Res. Methods* **51**, 987–1006 (2019).
38. F. L. Oswald, G. Mitchell, H. Blanton, J. Jaccard, P. E. Tetlock, Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *J. Pers. Soc. Psychol.* **105**, 171–192 (2013).
39. U. Schimmack, The Implicit Association Test: A method in search of a construct. *Perspect. Psychol. Sci.* **16**, 396–414 (2021).
40. A. G. Greenwald, M. R. Banaji, B. A. Nosek, Statistically small effects of the Implicit Association Test can have societally large effects. *J. Personality Soc. Psychol.* **108**, 553–561 (2015), 10.1037/pspa0000016.
41. B. Kurdi, K. A. Ratliff, W. A. Cunningham, 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).
42. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Soc. Rev.* **84**, 905–949 (2019).
43. S. Bhatia, Predicting risk perception: New insights from data science. *Manage. Sci.* **65**, 3800–3823 (2019).
44. S. Bhatia, C. Olivola, N. Bhatia, A. Ameen, Predicting leadership perception with large-scale natural language data. *Leadersh. Q.* **33**, 101535 (2021).
45. N. Gandhi, W. Zou, C. Meyer, S. Bhatia, L. Walasek, Computational methods for predicting and understanding food judgment. *Psychol. Sci.* **33**, 579–594 (2022).
46. S. Bhatia, R. Richie, Transformer networks of human concept knowledge. *Psychol. Rev.*, 10.1037/rev0000319 (2022).
47. S. Bhatia, A. Aka, Cognitive modeling with representations from large-scale digital data. *Curr. Dir. Psychol. Sci.* **31**, 207–214 (2022).
48. G. Recchia, M. M. Louwerse, Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Q. J. Exp. Psychol.* **68**, 1584–1598 (2015).
49. J. Sedoc, D. Preoţiuc-Pietro, L. Ungar, "Predicting emotional word ratings using distributional representations and signed clustering" in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Association for Computational Linguistics, 2017)
50. A. Toney, A. Caliskan, "ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2021), pp. 7203–7218.