Northwestern | INSTITUTE FOR POLICY RESEARCH

RESEARCH EXCELLENCE • POLICY IMPACT

IPR Working Paper Series

WP-24-23

Moralizing Partisanship When Surrounded by Co-Partisans Versus in Mixed Company

Michalis Mamakos
Northwestern University

Tessa Charlesworth
Northwestern University and IPR

Eli Finkel
Northwestern University and IPR

Version: September 16, 2024

DRAFT

Please do not quote or distribute without permission.

Abstract

Partisans tend to view their ingroup as moral and their outgroup as immoral. Here, the researchers examine whether left-wing and right-wing Reddit users (N > 1,000,000) express these partisan moralization views. Critically, they compare the rates of partisan moralization not only when users are in contexts (subreddits) of their ingroup (e.g., r/democrats, r/vegetarian, r/Conservative, r/Hunting), but also when in mixed-company contexts populated mostly by users without partisan engagement (e.g., r/Music, r/Parenting). First, the researchers developed four word embedding models—two for the users of each political side, one based on their comments in their ingroup contexts and one based on their comments in mixed-company contexts. Then, they evaluated the words of each model on two semantic dimensions, partisanship and morality, and they examined their correlation as an indicator of the expressed partisan moralization. The first analysis demonstrated that left-wing users express moralized partisanship to a similar degree when surrounded by co-partisans and when in mixed company. However, the moralized partisanship expressed by right-wing users in mixed company is weaker than that they express among co-partisans, as well as that expressed by left-wing users in mixed company. In a second analysis, the researchers divided partisan contexts based on whether they are inherently political (e.g., r/democrats) or not (e.g., r/vegetarian). This second analysis revealed that right-wing users express moralized partisanship more strongly than left-wing users in inherently political contexts, but right- and left-wingers are similar in nonpolitical partisan contexts. These asymmetries can potentially be attributed to the self-censoring of right-wingers due to fear of social sanction.

Introduction

Partisan moralization constitutes a core component of today's surging polarization [1]. People with higher moral convictions display greater partisan bias [2] and are more likely to endorse undemocratic processes to achieve their partisan goals [3]. Moreover, partisans disseminate moral content within their ingroup networks [4], especially when that content aligns with their convictions [5]. Much of the dissemination of partisan moral content today happens on social media [6], where users can join both political and nonpolitical communities (although the latter often host political discussions nonetheless [7]). Here we ask whether the social context alters partisan moralization: Do people express partisan moralization to a different degree when among their co-partisans versus in mixed company?

The present report examines the comments of left-wing (LW) and right-wing (RW) Reddit users in partisan subreddits (those that are disproportionately populated by users of one side) and in mixed-company subreddits (those that are nonpolitical in content, and populated mostly by users without partisan engagement and similarly by users of each side). For each of the two groups of users, we developed two word embedding models [8], one based on their comments in their respective ingroup subreddits (e.g., r/democrats, r/vegetarian, r/Conservative, r/Hunting), and one based on their comments in mixed-company subreddits (e.g., r/Music, r/Parenting). For each of the four models separately, we constructed a semantic dimension of partisanship by projecting [9] each word onto the average of vector representation differences of pairs of words (e.g., $\overrightarrow{republican} - \overrightarrow{democrat}$) from a list we devised. We constructed a semantic dimension of morality similarly, with pairs of words (e.g., $\overrightarrow{moral} - \overrightarrow{immoral}$). Both partisanship and morality dimensions were extensively validated (*SI Appendix*).

Results

Figure 1a depicts the z-scores in the semantic dimensions of partisanship (x-axis) and morality (y-axis) of the words in the embedding model developed with the comments of LW users in LW subreddits. These two dimensions were correlated ($r = -.24$, $P_{\mathrm{bootstrap}} < .001$). Results in Figure 1b reveal that LW users expressed comparable moralized partisanship in mixed-company subreddits ($r = -.25$, $P_{\mathrm{bootstrap}} < .001$). A bootstrapped distribution of the difference between these correlations suggests no statistically significant difference ($P_{\mathrm{bootstrap}} > .39$; Fig. 1c). Thus, LW users expressed moralized partisanship as strongly in mixed company as when surrounded by their ingroup left-wingers. In contrast, RW users expressed moralized partisanship more strongly in their ingroup partisan subreddits ($r = .40$, $P_{\mathrm{bootstrap}} < .001$; Fig. 1d) than in mixed-company subreddits ($r = .12$, $P_{\mathrm{bootstrap}} < .05$; Fig. 1e). The results in Figure 1f demonstrate that this difference was significant ($P_{\mathrm{bootstrap}} < .001$). Cross-group comparisons revealed that the correlation between partisanship and morality was stronger for RW users in RW subreddits than for LW users in LW subreddits ($P_{\mathrm{bootstrap}} < .001$), whereas this correlation

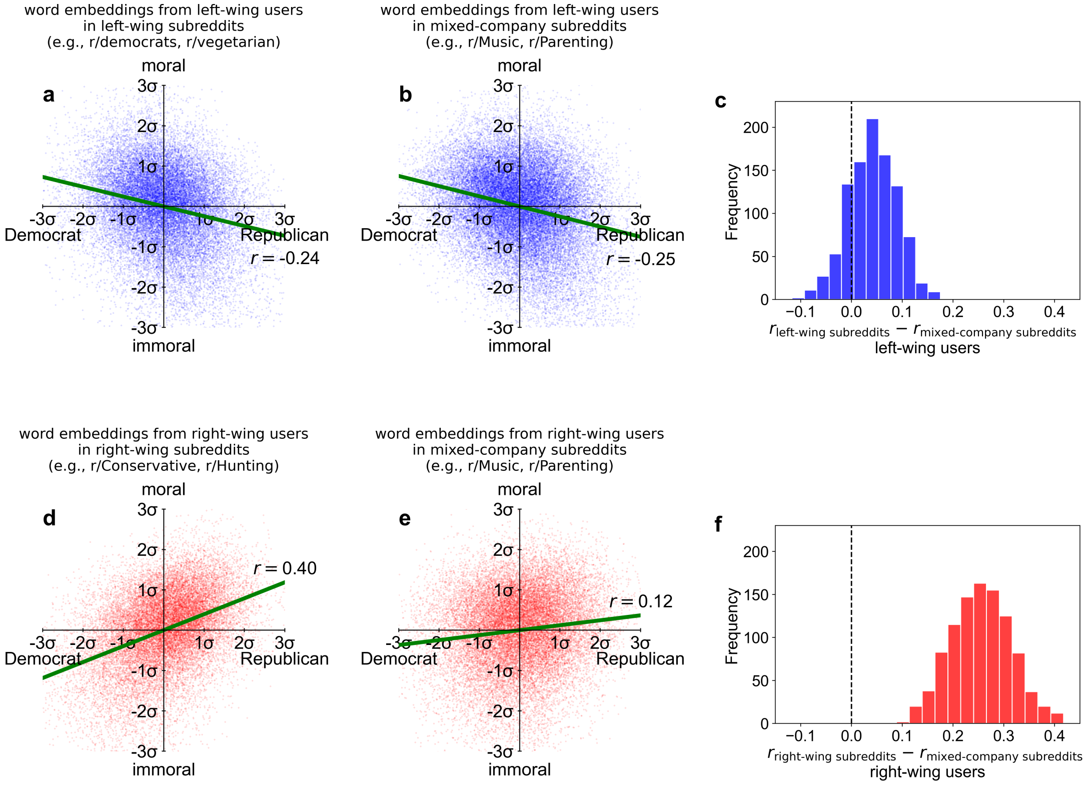

Fig. 1. Partisanship and morality of words from the embedding models of (a) left-wing users in left-wing subreddits, (b) left-wing users in mixed-company subreddits, (d) right-wing users in right-wing subreddits, and (e) right-wing users in mixed-company subreddits; (c) and (f) depict bootstrapped distributions of differences of correlations between partisanship and morality across partisan and mixed-company subreddits for left-wing users and right-wing users, respectively.

was stronger for LW users than for RW users in mixed-company subreddits ($P_{\text{bootstrap}} < .02$).

Next, we examined whether the degree of partisan moralization expressed in partisan subreddits (see Figures 1a and 1d) differs based on whether these subreddits are inherently political (e.g., r/democrats, r/Conservative) or not (e.g., r/vegetarian, r/Hunting). To this end, we developed four new word embedding models (LW users in political LW subreddits, LW users in nonpolitical LW subreddits, RW users in political RW subreddits, and RW users in nonpolitical RW subreddits). Again, all semantic dimensions were validated (*SI Appendix*).

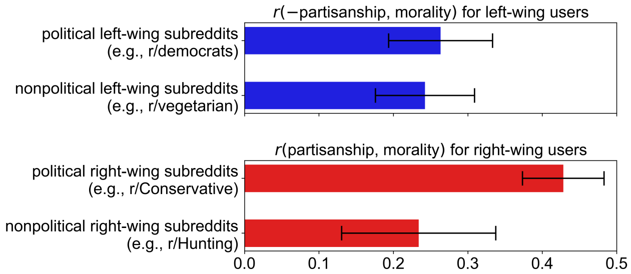

Fig. 2. Correlations between the partisanship and the morality of words in political and nonpolitical partisan subreddits, with 95% confidence intervals.

In Figure 2, the correlation between partisanship and morality was similar for (i) LW users in political LW subreddits ($r = .26$; dimension of partisanship is reversed), (ii) LW users in nonpolitical LW subreddits ($r = .24$), and (iii) RW users in nonpolitical RW subreddits ($r = .23$). However, the moralized partisanship expressed by RW users in political RW subreddits was significantly stronger ($r = .43$, $P_{\text{bootstrap-Bonferroni}} < .01$). These results show that our previous finding about the stronger partisan moralization of RW users compared to LW users in their respective ingroup contexts is driven by differences in political, rather than nonpolitical, contexts.

Our final analysis examined the words that each political side associated with the outgroup in mixed-company subreddits (which are all nonpolitical in content). The left column of Table 1 shows the top-10 words that LW users associated with RW targets in mixed company. These words centered on religion and oppression, and were, on average, immoral ($M = -1.35$, $P < .01$). In contrast, the right column of Table 1 shows the top-10 words that RW users associated with LW targets in mixed company. Here, the words lacked a specific theme and were, on average, not consistently moralized ($M = 0.21$, $P > .38$). This qualitative analysis reinforces the earlier quantitative results showing that LW users expressed stronger partisan moralization than RW users in mixed company.

Discussion

In this work, we examined the semantic association between partisanship and morality as revealed by the comments of left-wing and right-wing users both in their ingroup and in mixed-company subreddits. Our findings suggest that left-wing users express moralized partisanship to the same degree whether they are among their co-partisans or in mixed company. In contrast, for right-wing users the audience of a discussion is an important factor in whether or not they express partisan moralization. Right-wing users express moralized partisanship more strongly when among their co-partisans than in mixed company—where they expressed moralized partisanship more weakly than left-wing users. Furthermore, our results show that the moralizing tendency of right-wingers is particularly strong in inherently political spaces of their ingroup.

Table 1. Top-10 words associated with the outgroup, and their morality z-scores in parentheses. Seed words have been excluded.

Words that left-wingers associated with right-wingers in mixed-company subreddits (e.g., r/Music, r/Parenting)	Words that right-wingers associated with left-wingers in mixed-company subreddits (e.g., r/Music, r/Parenting)
evangelical (−0.52)	arts (−0.15)
fundamentalist (−1.62)	zoo (−0.32)
mormon (0.16)	tai (0.34)
gop (−0.87)	utopia (0.06)
batshit (−1.49)	art (0.57)
anti-science (−2.09)	grav (0.62)
abusive (−3.48)	paradise (−0.07)
homophobic (−2.13)	hostel (1.02)
evangelicals (−1.24)	waterloo (1.36)
church (−0.25)	spill (−1.28)
Average morality scores (positive = moral, negative = immoral)	
−1.35	0.21

Future work could fruitfully investigate the underlying causes of the observed asymmetries in the expression of partisan moralization. While our findings do not afford a causal attribution, it seems likely that right-wingers engage in self-censorship when in mixed company. This explanation aligns with recent findings that conservatives (and moderates) feel less free to speak their minds than liberals, for fear of facing backlash from proponents of “cancel culture” [10]. This case could have consequences for the democratic functioning of social media, especially given that left-wingers express moralized partisanship to the same degree regardless of the surrounding company. Under this explanation, another question that future research can test is whether the heightened moralization by right-wingers in political contexts is a form of reactance to the censorship they impose on themselves in other public spaces.

Another possible explanation for our findings is that right-wingers generally do not moralize partisanship as much and they are inclined to do so more when among ingroup members. However, results showing that conservatives have stronger tendencies to moralize than liberals, even in nonpolitical contexts, render this explanation unlikely [11]. Ultimately, this research sheds light on the social media contexts that might be the most hostile to democratic communication. Targeted interventions can be informed by understanding which users in which contexts are especially likely to moralize partisanship.

Methods

Using the partisan segregation measure of Waller and Anderson [12], we classified subreddits as left-wing if they had segregation at least 2 *SD*s below the neutral point of 0 ($N_{\text{left-wing subreddits}} = 291$), and as right-wing if they had segregation at least 2 *SD*s above that neutral point ($N_{\text{right-wing subreddits}} = 176$). We classified subreddits as mixed-company if they met both criteria of (i) having segregation at most 0.25 *SD*s away from the neutral point of 0 and (ii) being of nonpolitical content ($N_{\text{mixed-company subreddits}} = 2,078$). Mixed-company subreddits are populated mostly by users without engagement in partisan subreddits [13]. Subreddits were classified as of political content if they appeared in the list of political subreddits devised by Hofmann et al. [14].

We classified as left-wing the users with at least 10 comments in left-wing subreddits, 10 comments in mixed-company subreddits, and exactly 0 comments in right-wing subreddits ($N_{\text{left-wing users}} = 671,979$) in the period 2006-2022. Similarly, we classified as right-wing the users with at least 10 comments in right-wing subreddits, 10 comments in mixed-company subreddits, and exactly 0 comments in left-wing subreddits ($N_{\text{right-wing users}} = 430,148$). Further details are reported in *SI Appendix*.

Author contributions statement

MM, TESC, EJF designed research; MM performed research; MM analyzed data; and MM, TESC, EJF wrote the paper.

References

1. Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. Political sectarianism in america. *Science*, 370(6516):533–536, 2020.
2. Kristin N Garrett and Alexa Bankert. The moral roots of partisan division: How moral conviction heightens affective polarization. *British Journal of Political Science*, 50(2):621–640, 2020.
3. Linda J Skitka and G Scott Morgan. The social and political implications of moral conviction. *Political Psychology*, 35:95–110, 2014.
4. William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318, 2017.
5. Antoine Marie, Sacha Altay, and Brent Strickland. Moralization and extremism robustly amplify myside sharing. *PNAS Nexus*, 2(4):pgad078, 2023.
6. Jay J Van Bavel, Claire E Robertson, Kareena Del Rosario, Jesper Rasmussen, and Steve Rathje. Social media and morality. *Annual Review of Psychology*, 75:311–340, 2024.
7. Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 525–536, 2021.
8. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.
9. Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.
10. James L Gibson and Joseph L Sutherland. Keeping your mouth shut: Spiraling self-censorship in the united states. *Political Science Quarterly*, 138(3):361–376, 2023.
11. Jim Albert Charlton Everett, Cory J Clark, Peter Meindl, Jamie B Luguri, Brian D Earp, Jesse Graham, Peter H Ditto, and Azim F Shariff. Political differences in free will belief are associated with differences in moralization. *Journal of Personality and Social Psychology*, 120(2):461, 2021.
12. Isaac Waller and Ashton Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268, 2021.
13. Michalis Mamakos and Eli J Finkel. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus*, 2(10):pgad325, 2023.
14. Valentin Hofmann, Hinrich Schütze, and Janet B Pierrehumbert. The reddit politosphere: A large-scale text and network resource of online political discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1259–1267, 2022.

Supporting Information for
Moralizing partisanship when surrounded by co-partisans versus in mixed company

Michalis Mamakos[a,1], Tessa E. S. Charlesworth[a], Eli J. Finkel[a,b]

[a] Kellogg School of Management, Northwestern University
[b] Department of Psychology, Northwestern University

[1] Corresponding author
Email: mamakos@u.northwestern.edu

Extended Methods

Developing the word embedding models. The word embedding models were trained with the Word2vec algorithm. We selected the number of dimensions of the embeddings and the exponent that determines the negative sampling distribution based on 7 intrinsic evaluation tasks of word similarity (1-7): SimVerb-3500, WordSim-353, RareWord, SimLex-999, MEN, MTurk-287, and MTurk-771. For our four corpora corresponding to the results in Figure 1 of the main text, the results from the intrinsic evaluation tasks suggested that performance increased with embedding dimensionality, but only marginally after 300 dimensions, and so we selected this number of dimensions. Regarding the exponent for the negative sampling, the same results favored the value −0.5 suggested by Caselles-Dupré et al. (8) over the value 0.75 typically employed in models trained with the Word2vec algorithm. We set the parameter for the minimum frequency of words to 500 (discarding words with a smaller frequency than this in a corpus) because we found this beneficial for the validation of the semantic dimension of partisanship, which is described below. For all the other parameters of Word2vec, we used the default values of Gensim.[1]

Constructing and validating the semantic dimension of partisanship. To construct the semantic dimension of partisanship, we devised the following list of 6 pairs of words: [republican, democrat], [republicans, democrats], [conservative, liberal], [conservatives, liberals], [conservatism, liberalism], and [right-wing, left-wing].

Initially, we had also considered the pairs [righty, lefty], [righties, lefties], [rightist, leftist], [rightists, leftists], and [repubs, dems], but at least one of the words of each pair did not meet the minimum frequency in at least one of the four corpora corresponding to the models in Figure 1 of the main text, and thus we left out of the list these five pairs. The partisanship score of a word was computed as the cosine similarity of its vector representation to the average vector difference of the pairs of words in the list, where the differences were taken after having normalized the words in the list to have unit length. That is, considering the vector *rightleft* defined as,

$$\begin{aligned}\overrightarrow{rightleft} = &\left(\overrightarrow{republican} - \overrightarrow{democrat}\right) + \left(\overrightarrow{republicans} - \overrightarrow{democrats}\right)\\ &+ \left(\overrightarrow{conservative} - \overrightarrow{liberal}\right) + \left(\overrightarrow{conservatives} - \overrightarrow{liberals}\right)\\ &+ \left(\overrightarrow{conservatism} - \overrightarrow{liberalism}\right) + \left(\overrightarrow{right-wing} - \overrightarrow{left-wing}\right)\end{aligned}$$

then the partisanship score of a word is defined as,

$$partisanship(word) = \cos\left(\overrightarrow{word}, \frac{1}{6}\overrightarrow{rightleft}\right)$$

Notice that the division of *rightleft* by 6 is not necessary, as cosine similarity is scale-invariant.

[1] https://radimrehurek.com/gensim/models/word2vec.html

The partisanship scores were then z-scored separately for each model. Therefore, every word in an embedding model is assigned a partisanship score so that a lower score indicates that a word is more semantically similar to the political left, and a higher score that a word is more semantically similar to the political right. This approach was originally proposed by Kozlowski et al. (reference in the main text) who applied it to generate other semantic dimensions of social interest, such as affluence (rich vs. poor) and gender (feminine vs. masculine). Note also that prior work has shown that semantic dimensions can be constructed even with fewer words than those used here (12 words; 6 pairs) to construct the semantic dimension of partisanship (9).

To validate the semantic dimension of partisanship, we needed an externally provided vocabulary with annotated left-wing and right-wing words. To the best of our knowledge, such a vocabulary was not publicly available, and thus we developed one. We first extracted a list of 800 (non-unique) words from two popular pre-trained word embedding models, as follows. Using the model glove-twitter-200, we derived the 100 most similar words to the word *democrat*. We then did the same for the words *republican*, *democrats*, and *republicans*. We repeated this step using the model glove-wiki-gigaword-300. There were 327 unique words among these 800 words. We then recruited 7 undergraduate research assistants from our university to assess "*With which political party are the following words associated?*" with 3 options available for each word: *Democrats*, *Neither*, and *Republicans*. For 303 of the 327 words (93% of the words), at least 5 of the 7 research assistants provided the same assessment. Among these 303 words, 54 were assessed as being associated with Democrats and 57 with Republicans (for the rest 192 words, at least five research assistants chose *Neither*).

The following is the list of the 54 words associated with Democrats:
atheist, atheists, barack, biden, clinton, corzine, crist, cuomo, daschle, democrat, democratic, democrats, dems, dianne, dnc, dodd, feinstein, feminism, feminist, feminists, gephardt, gillibrand, gore, hillary, hipster, kennedy, kerry, ldp, left-wing, leftist, lib, liberal, liberalism, liberals, libs, lieberman, markey, massachusetts, mcgovern, murtha, naacp, ndp, obama, obamacare, pelosi, progressive, progressives, rangel, rodham, schumer, socialism, socialist, socialists, spd.

The following is the list of the 57 words associated with Republicans:
bjp, boehner, bush, catholic, catholics, centre-right, christian, christians, christie, conservador, conservatism, conservative, conservatives, cornyn, cpac, damato, dole, evangelical, evangelicals, fundamentalist, gingrich, giuliani, gop, gramm, hastert, huckabee, iowa, likud, lott, mccain, mcconnell, mitt, mormon, nationalist, newt, nra, palin, pataki, pro-life, reagan, repub, republican, republicans, repubs, right, right-wing, rightist, rnc, romney, santorum, tcot, teaparty, tories, tory, traditionalist, ukip, whig.

Then, for each of our models, we assessed the partisanship score of the words associated with Democrats and of the words associated with Republicans. The model based on the comments of left-wing users in left-wing subreddits (Figure 1a, main text) achieved *d* = 3.08. The model based on the comments of left-wing users in mixed-company subreddits (Figure 1b, main text) achieved *d* = 2.23. The model based on the comments of right-wing users in right-wing subreddits (Figure 1d, main text) achieved *d* = 3.14. The model based on the comments of right-wing users in mixed-company subreddits (Figure 1e, main text) achieved *d* = 2.08. Therefore, we observe that the distance between the words about Democrats and the words about Republicans was at least 2 SDs in each of the four models, suggesting successful validation of the dimension of partisanship in all these models. Similar conclusions were derived for the validation of partisanship in the four models in Figure 2 of the main text: for left-wing users in political left-wing subreddits, *d* = 2.85; for left-wing users in nonpolitical left-wing subreddits, *d* = 2.90; for right-wing users in political right-wing subreddits, *d* = 3.32; for right-wing users in nonpolitical right-wing subreddits, *d* = 2.64.

Constructing and validating the semantic dimension of morality. To construct the semantic dimension of morality, we devised a list of 18 pairs of words, with 3 pairs tapping general morality

and 3 pairs for each of the five moral foundations of the Moral Foundations Theory (care, fairness, loyalty, authority, sanctity):
[moral, immoral], [ethical, unethical], [righteous, wicked], [care, neglect], [peaceful, violent], [empathetic, apathetic], [fair, unfair], [unbiased, biased], [justice, injustice], [loyalty, treason], [fidelity, infidelity], [ally, enemy], [comply, defy], [respectful, disrespectful], [lawful, unlawful], [holy, unholy], [clean, filthy], [integrity, corruption]

The derivation of the morality scores for the words of each model was then similar to that for the dimension of partisanship.

We first validated the semantic dimension of morality for each model with the Moral Foundations Dictionary (10), which consists of words about virtues and vices. The model based on the comments of left-wing users in left-wing subreddits (Figure 1a, main text) achieved *d* = 2.20. The model based on the comments of left-wing users in mixed-company subreddits (Figure 1b, main text) achieved *d* = 2.32. The model based on the comments of right-wing users in right-wing subreddits (Figure 1d, main text) achieved *d* = 2.38. The model based on the comments of right-wing users in mixed-company subreddits (Figure 1e, main text) achieved *d* = 2.41. The models corresponding to Figure 2 achieved *d* between 2.09 and 2.34. These outcomes suggest successful validation of the semantic dimension of morality for all our 8 models.

To further validate the semantic dimension of morality, we also considered correlations with the extended Moral Foundations Dictionary (11). Each word in this dictionary is rated according to how relevant the word is to the five moral foundations, ranging from 0 (not relevant) to 1 (completely relevant). For example, the word "tortured" is rated as highly relevant to care, while the word "employed" is rated as irrelevant to care. We computed the dictionary-based (unsigned) morality score of a word as the mean across these five values, such that higher scores indicate the word refers more to moral foundations, in general. Notice that these ratings of moral relevance alone do not specify whether a word is about a virtue or a vice. Thus, we also used the ratings of all words on sentiment scores (which are part of this dictionary), ranging from −1 (most negative sentiment) to 1 (most positive sentiment). For example, the words "friendly" and "hostile" are both equally relevant to morality, but "friendly" is positive in sentiment and "hostile" is negative in sentiment. We assigned a negative sign to the morality score of a word if its sentiment scores were negative and a positive sign if its sentiment scores were positive, excluding words with mixed sentiment signs across different foundations. For all 8 models, the correlation between the model-based morality and the dictionary-based morality of the words was between .58 and .62. These adequately high correlations, which show little variation across our models, are therefore in line with our conclusion about the successful validation of the dimension of morality.

The dimension of valence. To provide further intrinsic evaluation for our word embedding models, we also constructed a semantic dimension of valence based on the 25 pleasant and the 25 unpleasant words used in Caliskan et al. (12), and then we examined the correlation of this dimension with the human judgments derived in Warriner et al. (13). For the four models corresponding to Figure 1 of the main text, this correlation was between .59 and .63, and for the four models corresponding to Figure 2 of the main text, this correlation was between .47 (left-wing users in political left-wing subreddits) and .62 (left-wing users in nonpolitical left-wing subreddits). Thus, the correlation between the dimension of valence in our models and the human judgments of valence was overall moderate-to-high, providing confidence in the trained embedding models.

Avoiding overrepresentation of users. To avoid overrepresenting the highly active users, for users with more than 50 comments in a type of context (partisan or mixed-company), we randomly sampled 50 of their comments in that context type before developing the four word embedding models in Figure 1. The models corresponding to Figure 2 were developed based on data used for the models corresponding to Figure 1, as described below.

Paired bootstrap. Now, we describe the bootstrap method used to generate the results in Figures 1c and 1f of the main text. Notice that by construction of our dataset, all users had comments both in their ingroup partisan subreddits and in the mixed-company subreddits (between 10 and 50 comments in each type of subreddits). Thus, the models developed for these two types of subreddits are not independent, as individual users might have idiosyncrasies that are reflected in the semantics of the language in their comments, regardless of the context in which they are posted. To respect this dependency, in each bootstrap replication we first sampled users with replacement. That is, for the bootstrap results corresponding to Figure 1c, we sampled with replacement 671,979 (= $N_{\text{left-wing users}}$) users, and for the bootstrap results corresponding to Figure 1f, we sampled with replacement 430,148 (= $N_{\text{right-wing users}}$) users. Then, for a bootstrap replication corresponding to Figure 1c (1f), we formed a corpus in which the comments of a user in left-wing (right-wing) subreddits appeared as many times as that user appeared in the bootstrapped sample, and a corpus in which the comments of that user in mixed-company subreddits also appeared as many times as the user appeared in the bootstrapped sample. After the models were trained and the correlations between partisanship and morality were computed, the replication-specific difference between the correlations for partisan and mixed-company subreddits was taken. We performed 1,000 bootstrap replications, which implies that in Figures 1c and 1f combined, the total number of models trained was 4,000.

The reported bootstrapped p-values are based on the proportions of the bootstrap replications against the null hypothesis. For instance, in Figure 1c, in 198 of the 1,000 replications (19.8%) the difference between the two correlations was negative. Notice that this proportion corresponds to a one-sided hypothesis test. To report a p-value for a two-sided hypothesis test, we doubled this proportion (39.6%), and thus in the main text we are reporting $P_{\text{bootstrap}} > .39$. Moreover, all the 1,000 bootstrapped correlations for the left-wing users in left-wing subreddits and all the 1,000 bootstrapped correlations for these users in mixed-company subreddits were negative. Thus, $P_{\text{bootstrap}} < .001$ is reported for the results corresponding to Figures 1a and 1b of the main text. In Figure 1f, all the 1,000 bootstrapped differences were positive, with all the 1,000 bootstrapped correlations for right-wing users in right-wing subreddits being positive. For such users in mixed-company subreddits, 23 of the 1,000 bootstrapped replications were negative, and thus we are reporting $P_{\text{bootstrap}} < .05$ for a two-sided hypothesis test.

Political and nonpolitical partisan subreddits. The four models corresponding to Figure 2 (main text) were developed by splitting the comments used for the models corresponding to Figures 1a and 1d (main text). The 291 left-wing subreddits consisted of 45 political and 246 nonpolitical subreddits, and the 176 right-wing subreddits consisted of 34 political and 142 nonpolitical subreddits. There were 118,980 left-wing users with at least one comment in political left-wing subreddits and 660,743 left-wing users with at least one comment in nonpolitical left-wing subreddits, with only 11,227 left-wing users (1.7% of all the 671,979 left-wing users) having commented in political but not in nonpolitical left-wing subreddits. There were 125,604 right-wing users with at least one comment in political right-wing subreddits and 391,163 right-wing users with at least one comment in nonpolitical right-wing subreddits, with only 38,972 right-wing users (9.1% of all the 430,148 right-wing users) having commented in political but not in nonpolitical right-wing subreddits. Therefore, the vast majority of all the users had at least one comment in nonpolitical partisan subreddits (98.3% of the left-wing users and 90.9% of the right-wing users).

In Figure 2 (main text), the 95% confidence intervals were computed by multiplying bootstrapped standard errors (based on 1,000 replications) by ±1.96. Because not every user had comments in both political and nonpolitical partisan subreddits, we did not consider a paired bootstrap as in the results for Figure 1 (main text). However, we still sampled users with replacement, in each bootstrap replication. All the 4,000 bootstrapped correlations corresponding to Figure 2 (main text) were positive, and thus it holds that $P_{\text{bootstrap}} < .001$ for each group-context pair. In the comparison between left-wing and right-wing users in their respective political partisan subreddits, we compared 1,000 pairs of the (randomly paired) bootstrapped replications (14). The correlation of the left-wing users was higher than that of the right-wing users in only 1 of the 1,000 comparisons ($P_{\text{bootstrap}} < .002$). In similar comparisons of the right-wing users in political partisan subreddits against the left-wing users in nonpolitical partisan subreddits and the right-wing users in nonpolitical partisan subreddits, the correlation of the right-wing users in political partisan subreddits was always higher ($P_{\text{bootstrap}} < .001$). To take into account the three simultaneous comparisons, in the main text we have reported $P_{\text{bootstrap-Bonferroni}} < .01$.

Other supporting information. The results of the four Pearson correlations between partisanship and morality depicted in Figure 1 of the main text were replicated when Spearman correlations were considered instead (ρ = −.22 for the results in Figure 1a, ρ = −.23 for the results in Figure 1b, ρ = .38 for the results in Figure 1d, and ρ = .11 for the results in Figure 1e), and when the correlations were taken only for the common words (N = 16,902 common words) of the four models (r = −.22 for the results in Figure 1a, r = −.22 for the results in Figure 1b, r = .39 for the results in Figure 1d, and r = .12 for the results in Figure 1e).

The 10 words in the left column of Table 1 (main text) are the 10 right-most words in Figure 1b (main text), excluding the 12 words about partisanship in the list presented above. Similarly, the 10 words in the right column of Table 1 (main text) are the 10 left-most words in Figure 1e (main text), excluding the same 12 words about partisanship.

Moderators and usernames including the term “bot” were excluded from our sets of users.

Reddit comments were extracted from the Pushshift dataset (15).

SI References

1. Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
2. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001, April). Placing search in context: The concept revisited. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 406-414).
3. Luong, M. T., Socher, R., & Manning, C. D. (2013, August). Better word representations with recursive neural networks for morphology. In *Proceedings of the 17th Conference on Computational Natural Language Learning* (pp. 104-113).
4. Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.
5. Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.
6. Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011, March). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 337-346).
7. Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012, August). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1406-1414).
8. Caselles-Dupré, H., Lesaint, F., & Royo-Letelier, J. (2018, September). Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 352-356).
9. Charlesworth, T. E., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, *119*(28), e2121798119.
10. Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029.
11. Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior research methods*, *53*, 232-246.

12. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.
13. Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191-1207.
14. Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: a primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920911881.
15. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media* (Vol. 14, pp. 830-839).