AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2022, Vol. 151, No. 12, 3292–3303
https://doi.org/10.1037/xge0001247

Re-Examining the Spread of Moralized Rhetoric From Political Elites: Effects of Valence and Ideology

Sze-Yuh Nina Wang and Yoel Inbar
Department of Psychology, University of Toronto

We examine the robustness of previous research finding increased diffusion of Twitter messages ("tweets") containing moral rhetoric. We use a distributed language model to examine the moral language used by U.S. political elites in two corpora of tweets: one from 2016 presidential candidates Hillary Clinton and Donald Trump, and one from U.S. Members of Congress. Consistent with previous research, we find greater diffusion for tweets containing moral rhetoric, but this is qualified by moral language valence and elite ideology. For both presidential candidates and Members of Congress, negative moral language is associated with increased message diffusion. Positive moral language is not associated with diffusion for presidential candidates and is *negatively* associated with diffusion for Members of Congress. In both data sets, the relationship between negative moral language and message diffusion is stronger for liberals than conservatives.

Keywords: moral language, natural language analysis, political ideology, political psychology

Supplemental materials: https://doi.org/10.1037/xge0001247.supp

Morality is essential to a functioning society but also one of the greatest risks it faces. Membership in a moral community fosters trust, altruism, and cooperation (Haidt, 2001, 2007). But when people start to see social conflicts as connected to their moral values, the result can be distrust, hostility, and even violence toward perceived enemies (Fiske & Rai, 2014; Mooijman et al., 2018; Skitka, 2010; Skitka & Mullen, 2002). The moralization of conflict is particularly consequential in politics, where morality motivates groups to achieve political goals but also increases partisan hostility and reduces willingness to compromise (Marietta, 2008; Ryan, 2017).

Political elites (e.g., legislators and party leaders) use language to shift the public debate and persuade and motivate voters (Lakoff, 2004). Moral language is an especially powerful way to do this, as voters respond especially strongly to moral language from elites (Clifford & Jerit, 2013). As more political discussion shifts to online social media platforms (Aral, 2020; Marantz, 2019; Tufekci, 2017), moral language from elites may become even more influential, because sites such as Facebook and Twitter typically allow users to reshare messages from others. This allows messages from elites to reach both users who follow them directly and—if those users decide to reshare the messages—users who do not. Prominent political figures often have large direct followings on social media (e.g., Senator Bernie Sanders had more than 12 million Twitter followers as of July 2021 and former President Donald Trump had more than 88 million before his account was suspended), so if even a small percentage of people reshare a post, its reach can be substantial. Widely shared ("viral") social-media posts from elites can also become news in their own right and are then amplified across other media.

There is some evidence that moral language is associated with greater message spread in online social networks. Brady et al. (2017) found that in discussions of political topics on Twitter, tweets containing more "moral-emotional" words (e.g., "hate," "murder," "shame") were retweeted more (i.e., they were reshared by users more often). The presence of a single moral-emotional word in a tweet increased its expected retweet count by nearly 20%. In subsequent studies, Brady et al. (2019) found that U.S. politicians' use of "moral-emotional" language in tweets was likewise associated with more retweets from users, and that this relationship was stronger for conservative politicians.

However, the validity of these findings has recently been questioned. Burton et al. (2021) found that the word-counting technique used by Brady et al. (2017, 2019) has poor correspondence with human coders in many data sets; that associations between use of moral-emotional words and retweets do not consistently emerge

This article was published Online First June 20, 2022.

Sze-Yuh Nina Wang https://orcid.org/0000-0002-1988-0023

This research was supported by an Insight Development Grant (4301700836) from the Social Sciences and Humanities Research Council of Canada to Yoel Inbar and a Doctoral Fellowship from the Social Sciences and Humanities Research Council of Canada to Sze-Yuh Nina Wang. This research uses openly-available data analyzed in Brady et al. (2019; available at https://osf.io/reqx9/) and data from Wang and Inbar (2021; available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FQ8MIL). Code is available at https://github.com/szeyuhninawang/moral-diffusion. Versions of this work were presented at the Society for Personality and Social Personality conference, the Computational Social Sciences Laboratory at the University of Southern California, the Behavioural Science Lab at the University of Regina, and the Human Cooperation Lab at MIT.

Correspondence concerning this article should be addressed to Sze-Yuh Nina Wang, Department of Psychology, University of Toronto, 1265 Military Trail, Toronto, ON M1C 1A4, Canada. Email: szeyuhnina.wang@mail.utoronto.ca

across different topics of discussion; and that irrelevant features of messages (the number of times they contain the characters x, y, and z) are associated with retweets just as strongly (or more strongly) as the number of moral-emotional words they contain. It is therefore an open empirical question whether moral language use is associated with greater message spread on Twitter. Here, we address this question using a validated natural language analysis technique called "distributed dictionary representations" (DDR; Garten et al., 2018) that avoids some of the limitations of word counting.

Natural Language Analysis With Distributed Dictionary Representations

Word counting (as implemented by, e.g., the Linguistic Inquiry and Word Count [LIWC] software; Tausczik & Pennebaker, 2010) has been the standard technique for automated text coding in psychological research. It is an attractive alternative to human coding for large amounts of text, but it is also limited in several ways. Most problematically, to measure a concept, researchers must specify the entire dictionary of words or word stems that represent it. If the dictionary coverage is too narrow, important parts of the concept will be omitted and text will not match when it should. But if the coverage is too broad, text will match when it should not. Word-counting software does not know what words mean, only what they look like, so including more terms can easily lead to false-positives (e.g., "happ*" matches "happy"and happiness, "but also" happened "and happenstance"). This problem is exacerbated for short texts such as tweets (which were limited to 140 characters until 2017 and are currently limited to 280 in English). Particularly for less common concepts, most tweets might contain no dictionary words at all. This may be why word-counting measures of moral sentiment on Twitter often do not show good agreement with human coders (Garten et al., 2018).

We avoid the limitations of word counting by using "distributed dictionary representations" (DDR; Garten et al., 2018; Hoover et al., 2020), a text analysis technique that combines researcher-specified dictionaries with a distributed language model. Distributed language models encode words as points in a many-dimensional space in which semantically similar words (e.g., "car" and "automobile") are close to each other (Landauer & Dumais, 1997; Osgood et al., 1957; Rogers & McClelland, 2004; Salton et al., 1975). Because distributed language models encode meaning (i.e., semantics) rather than spelling (i.e., orthography), they can uncover the similarity in meaning between texts that do not share any words. For example, a text containing "car" might be considered highly similar to a dictionary consisting of "automobile," "auto," and "vehicle." This means that dictionaries can be short but still cover a concept well, and texts can be short but still be considered similar to a dictionary. This eliminates the two major shortcomings of word-counting approaches. Indeed, validation studies have shown that using DDR to measure moral sentiment in tweets shows high agreement with human coders (Wang & Inbar, 2021), substantially more so than word-counting does (Garten et al., 2018; Hoover et al., 2020).

The Current Research

We use DDR to measure the moral sentiment expressed in two corpora of tweets from political elites. The first, which was collected by Brady et al. (2019), contains 9,505 tweets posted by Donald Trump and Hillary Clinton in the year leading up to the 2016 U.S. presidential election. The second, which we collected, contains all original tweets from members of the U.S. Congress from January 1, 2016 to January 31, 2018 (687,360 tweets; this dataset was first described in Wang & Inbar, 2021). Our primary research question is whether tweets expressing greater moral sentiment are retweeted more. We also have two secondary research questions: (a) whether positive or negative moral sentiment is more strongly associated with message spread, and (b) whether the relationship between moral language and message spread varies by elites' ideology (i.e., liberal or conservative).

Is Positive or Negative Language More Associated With Message Spread?

There are conflicting findings on the relationship between valence and message spread. In a sample of all public tweets from 2014; more positive sentiment was associated with greater message spread (Ferrera and Yang, 2015). Other research has found, however, that in tweets about news or politics, more negative sentiment is associated with greater message spread (Schöne et al., 2021). Brady et al. (2019) found that although both positive and negative moral-emotional words were associated with message spread, the relationship was stronger for negative words. We therefore separately coded positive and negative moral sentiment to test for effects of valence.

Does Elites' Ideology Matter?

There is not much research on the relationship between elite ideology and the use (or consequences) of moral rhetoric online. Brady et al. (2019) found that the relationship between moral-emotional language and retweets was stronger for conservative elites, but that this was because *positive* moral-emotional language was more strongly associated with retweets for conservatives than liberals (whereas negative moral-emotional language was associated equally strongly with retweets for conservatives and liberals). Conversely, Wang and Inbar (2021) found Democratic legislators (vs. Republicans) used more moral rhetoric overall on Twitter between 2016 and 2018. In analyses reported only in the online supplemental materials, they also tested the relationship between different types of moral rhetoric and retweets separately for Democrats and Republicans, finding that moral rhetoric generally predicted retweets directionally more strongly for Democrats, but that these differences often were not statistically significant. In the current data, we therefore test whether the effects of moral sentiment (positive and negative) on retweets varies by elites' ideology.

Method

Data and Code Availability

We used openly-available data analyzed in Brady et al. (2019; available at https://osf.io/reqx9/) and data from Wang and Inbar (2021; available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FQ8MIL). Our code is available at https://github.com/szeyuhninawang/moral-diffusion. This study did not require IRB approval as it used publicly available data from social media platforms.

Data Sets

We analyzed two data sets. The first, which was first described in Brady et al. (2019), consisted of 9,505 tweets posted by Donald Trump and Hillary Clinton in the year leading up to the 2016 U.S. presidential election. The second, which was first described in Wang and Inbar (2021), consisted of all tweets from members of the U.S. Senate and House between January 1, 2016 to January 31, 2018; inclusive (687,360 tweets).

Natural Language Model

We measured moral language using distributed dictionary representations (DDR; Garten et al., 2018). DDR combines a small researcher-specified dictionary of terms representing a concept (e.g., morality) with a distributed language model. Because the language model encodes words as n-dimensional vector of real numbers (where n = the number of dimensions in the space), averaging the normalized vectors for each word in the dictionary produces a single composite representation of the dictionary meaning.[1] Likewise, the words in a text (in this case, a tweet) can be combined into a composite representation. The similarity between texts and the dictionary can then be computed by cosine similarity (scores range from −1 to 1, with higher values reflecting greater similarity). To represent positive and negative moral language, we used seed words designed to measure "virtue" (positive) and "vice" (negative) aspects of the five "moral foundations" of *harm*, *fairness*, *loyalty*, *authority*, and *purity* (Graham et al., 2009; Haidt and Graham, 2007). These seed words were drawn from the larger Moral Foundations Dictionary (Graham et al., 2009) by Garten et al. (2018) and have been previously validated (Garten et al., 2018; Wang & Inbar, 2021), with past work suggesting that smaller sets of seed words can perform better than models trained on the full Moral Foundations Dictionary. To separately measure positive and negative moral language, we combined all the "virtue" (positive) terms (*authority, obey, respect, tradition, purity, sanctity, sacred, wholesome, loyal, solidarity, patriot, fidelity, kindness, compassion, nurture, empathy, fairness, equality, justice, rights*) into one dictionary. We combined all the "vice" (negative) terms (*impurity, depravity, degradation, unnatural, betray, treason, disloyal, traitor, cheat, fraud, unfair, injustice, subversion, disobey, disrespect, chaos, suffer, cruel, hurt, harm*) into another.

Measuring Ideology

We assess ideology differently in the two data sets. In the Clinton-Trump dataset, we assume Trump (the Republican candidate) to be more conservative than Clinton (the Democratic candidate). In the Congressional dataset, we quantify each legislator's ideology from liberal to conservative using DW-NOMINATE scores (Poole & Rosenthal, 1985), which are widely used in political science and related disciplines. DW-NOMINATE is a multidimensional scaling algorithm that places legislators in a two-dimensional space according to their Congressional voting records; legislators who are closer to each other in this space have more similar voting records than those who are farther apart. Historically, DW-NOMINATE dimension 1 (usually graphed on the *x*-axis) has reflected economic liberalism or conservatism, whereas dimension 2 has reflected other cross-cutting issues (e.g., racial politics in the 1960s civil rights era). However, in the last 30–40 years dimension, 1 has come to reflect overall left-right orientation, with dimension 2 becoming less and less important. Furthermore, over the same time period Democrats and Republicans have become much more polarized on dimension 1 (Hare & Poole, 2014). We therefore use dimension 1 scores as our measure of political ideology (these scores are coded such that higher numbers correspond to a more conservative voting record).

Analytic Strategy

In all models, we used log-transformed retweet counts as our dependent variables. Because retweets are almost always positively skewed, researchers generally log-transform them if linear models are used (e.g., Brady et al., 2019). The two data sets differ in the number of observations and the number of clusters (i.e., the number of accounts), so we used different approaches for each. For the Clinton-Trump dataset, we used OLS regression and dummy-coded tweet author (Clinton or Trump). For our analyses of tweets from Members of Congress, we used multilevel linear models, nesting tweets within accounts to model nonindependence between tweets from the same account.

When analyzing clustered data, the total regression coefficient for the relationship between X (in our case, moral language) and Y (in our case, retweets) is a blend of the within- and between-cluster coefficients (Raudenbush & Bryk, 2002, p. 138). In the current case, it is the within-cluster coefficient that tests the theoretically important question (i.e., when a tweet includes more moral language, is it retweeted more?). The between-cluster coefficient tests a different question (are accounts that use more moral language, on average, retweeted more?). Note that this could be for theoretically irrelevant reasons (as a hypothetical example, using more moral language might attract more politically-engaged audiences who are more likely to retweet political content in general). Therefore, in our analyses of the Congressional data we disaggregated within- and between-account effects by centering level 1 predictors (i.e., tweet-level moral language scores) within accounts, and then adding the account-level mean moral language scores as level 2 predictors. The "centered-within-cluster" (CWC) predictors test whether when a tweet uses more (or less) moral language than the average for the posting account, this is associated with more (or fewer) retweets of that tweet. The aggregate (account-level) predictors test whether accounts that use more moral language on average are retweeted more (for more on centering in multilevel modeling, see (Enders and Tofighi, 2007); and (Hamaker and Bengt, 2020). All models also included the account's number of followers (as of January 20, 2018; log-transformed) and legislator DW-NOMINATE scores (standardized) as level 2 covariates.

We first fit a model without interactions between moral language and legislator ideology, but that did allow random variation in the relationship between moral language and retweets across legislators (i.e., random slopes). Using the notation of Raudenbush and Bryk (2002) this model can be written as:

Level 1:

$$retweets_{ij} = \beta_{0j} + \beta_{1j}(positive_{CWC}) + {}_{2j}(negative_{CWC}) + r_{ij}$$

[1] In the following analyses, we use an n of 300, as is commonly done in word embedding studies (e.g., Mikolov et al., 2013).

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(positive_j) + \gamma_{02}(negative_j) + \gamma_{03}(followers_j) + \gamma_{04}(dw - nominate_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + u_{2j}$$

where $positive_{CWC}$ and $negative_{CWC}$ denote the positive and negative moral language scores for day i centered within account j, and $positive_j$ and $negative_j$ denote the mean positive and negative moral language scores for account j.

We then fit a model that added interactions between moral language and ideology. Because legislator ideology (dw-$nominate_j$) is a level 2 variable, the interactions between per-day centered-within-cluster moral language scores and ideology are cross-level. This model can be written as:

Level 1:

$$retweets_{ij} = \beta_{0j} + \beta_{1j}(positive_{CWC}) + \beta_{2j}(negative_{CWC}) + r_{ij}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(positive_j) + \gamma_{02}(negative_j) + \gamma_{03}(followers_j) + \gamma_{04}(dw - nominate_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(dw - nominate_j) + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}(dw - nominate_j) + u_{2j}$$

Note that for both data sets, we always tested the effects of positive moral language controlling for negative moral language and vice versa.[2]

All analyses were run in R (Version 3.6.3; R Core Team, 2020).

Results

Language Model Validation

We conducted a validation study comparing the performance of our language models to judgements by human coders (this study is described in detail in the online supplemental materials). Coders were trained research assistants who coded a subset of 850 tweets from Members of Congress on whether they contained positive moral language, negative moral language, or no moral language. Coders were instructed that the positive and negative moral categories were not exclusive, so that a tweet might contain language from both categories. Interrater reliability was measured with prevalence and bias adjusted kappa (PABAK; Byrt et al., 1993; Sim & Wright, 2005), with a mean PABAK of .80 (SD = .15) across categories. To classify tweets as containing positive and/or negative moral language based on their DDR scores, we used the scores as features in a logistic regression classifier with 10-fold cross-validation. This cross-validation splits the dataset into ten groups, fitting the classifier using each group as a training set and evaluating its performance on the other nine groups. This approach helps to measure out-of-sample performance (i.e., how well the model predicts labels for new data) more accurately. We used F1 scores (Powers, 2011), calculated as the harmonic mean of precision and recall with a maximum value of 1 and a minimum value of 0, to evaluate the performance of the DDR-based classifier (in this analysis, the coder's classifications were treated as the ground truth). This is a standard measure of performance in natural language processing (Jurafsky & Martin, 2020). Averaging across positive and negative moral language and coders, DDR achieved an F1 score of .93. We complete an additional validation using the Moral Foundations Twitter Corpus (Hoover et al., 2017), with an average F1 score across corpora and positive and negative moral language categories of .71 (see online supplemental materials for further details). The slightly lower F1 score here is reasonable given the greater heterogeneity of tweets in the Moral Foundations Twitter Corpus.

Clinton and Trump

We used the openly available data of Brady et al. (2019) downloaded from https://osf.io/reqx9/. We analyzed a set of tweets from both 2016 presidential candidates, using OLS regression for each dataset and log-transforming retweets. We find that negative moral language significantly predicts retweets, while positive moral language does not (see Table 1 and Figures 1–2). Testing for effects of source (i.e., whether the Tweet was authored by Clinton or Trump), we find a significant main effect of source, such that Trump tweets are shared more, $b = -1.14$, $t = 6.17$, $p < .001$, and a significant interaction between negative moral language and source, such that there is a stronger relationship between negative moral language and diffusion for Clinton than for Trump, $b = .73$, $t = 2.08$, $p = .037$. In a separate analysis of the effects of each moral foundation, we find that these effects seem to be largely driven by negative language pertaining to authority, fairness, and loyalty (see online supplemental materials for model details).

Senate and Congress (2016–2018)

We repeated these analyses in the dataset of all tweets from members of the U.S. Congress from January 1, 2016 to January 31, 2018 (see Wang & Inbar, 2021) using multilevel linear models implemented by the R package 'lme4' (Bates et al., 2015). These models are described in detail above in the "Analytic Strategy" section. All models include both "centered within cluster" (CWC) moral language scores at the tweet level and account-level aggregate scores; it is the coefficients for the CWC scores that test our primary hypotheses. Models included a measure of legislator ideology (DW-NOMINATE; see Measuring Ideology above for a description of these scores) and controlled for follower count (log-transformed to reduce positive skew).

We first fit a model that tested the relationship between moral language and retweets across all legislators (full results are shown in Table 2 (model 1). Overall, greater use of negative moral language was associated with more retweets, $b = 14.34$, $t(569.29)$ =

[2] This is because DDR scores for positive and moral language were correlated at the tweet level in both datasets (r = .83 in the Clinton-Trump dataset and r = .77; see Tables S3 and S4 in the online supplemental materials). These correlations are high both because positive and negative moralizing co-occurs, and because the natural language model (correctly) considers different types of moral language to be semantically related.

Table 1
Coefficients and Standard Errors From OLS Regressions Predicting Retweets for Clinton and Trump's Tweets

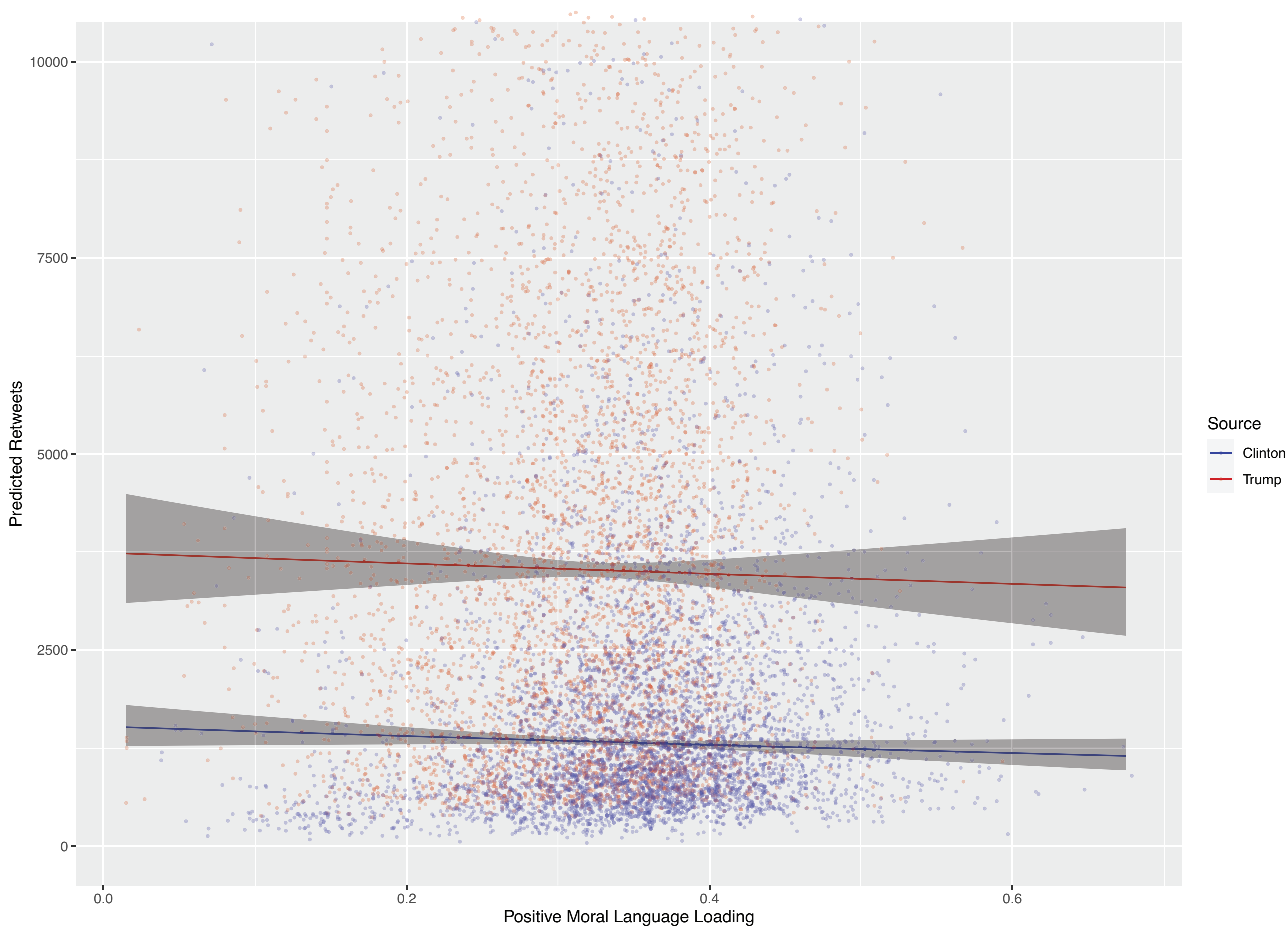
Predictor	Model 1	Model 2	Model 3
Positive moral language	−0.24 (0.20)	−0.19 (0.30)	−0.75 (0.32)*
Negative moral language	1.72 (0.17)***	1.46 (0.24)***	1.20 (0.28)***
Source (Clinton = 1; Trump = 0)	−0.97 (0.019)***	−1.14 (0.080)***	−1.32 (0.078)***
Positive Moral Language × Source		−0.23 (0.40)	0.37 (0.39)
Negative Moral Language × Source		0.73 (0.35)*	0.68 (0.35)·
Positive sentiment (VADER)			0.30 (065)***
Negative sentiment (VADER)			0.42 (0.10)***
Number of tweets from source that day			−0.027 (0.0,016)***
Mean negative moral language that day			2.78 (0.67)***
Mean positive moral language that day			0.36 (0.82)
Total tweets that day			0.016 (0.00,098)***

Note. Retweet counts are log-transformed. Model 1 includes positive moral language, negative moral language, and source as predictors. Model 2 adds interactions between moral language and source. Model 3 adds robustness-check controls.
$\cdot p = .05$. $^{*} p < .05$. $^{**} p < .01$. $^{***} p < .001$.

8.91, $p < .001$. Greater use of positive moral language was associated with *fewer* retweets, $b = -5.28$, $t(570.59) = -2.99$, $p = .0029$.

We next tested for interactions between moral language and the ideology of the posting account. Interaction effects show the same pattern as in the Clinton/Trump dataset: Tweets with more negative moral language were retweeted more regardless of legislator ideology, but this effect was stronger for more liberal legislators, interaction $b = -1.53$, $t(570.37) = -13.48$, $p < .001$. There was also a significant

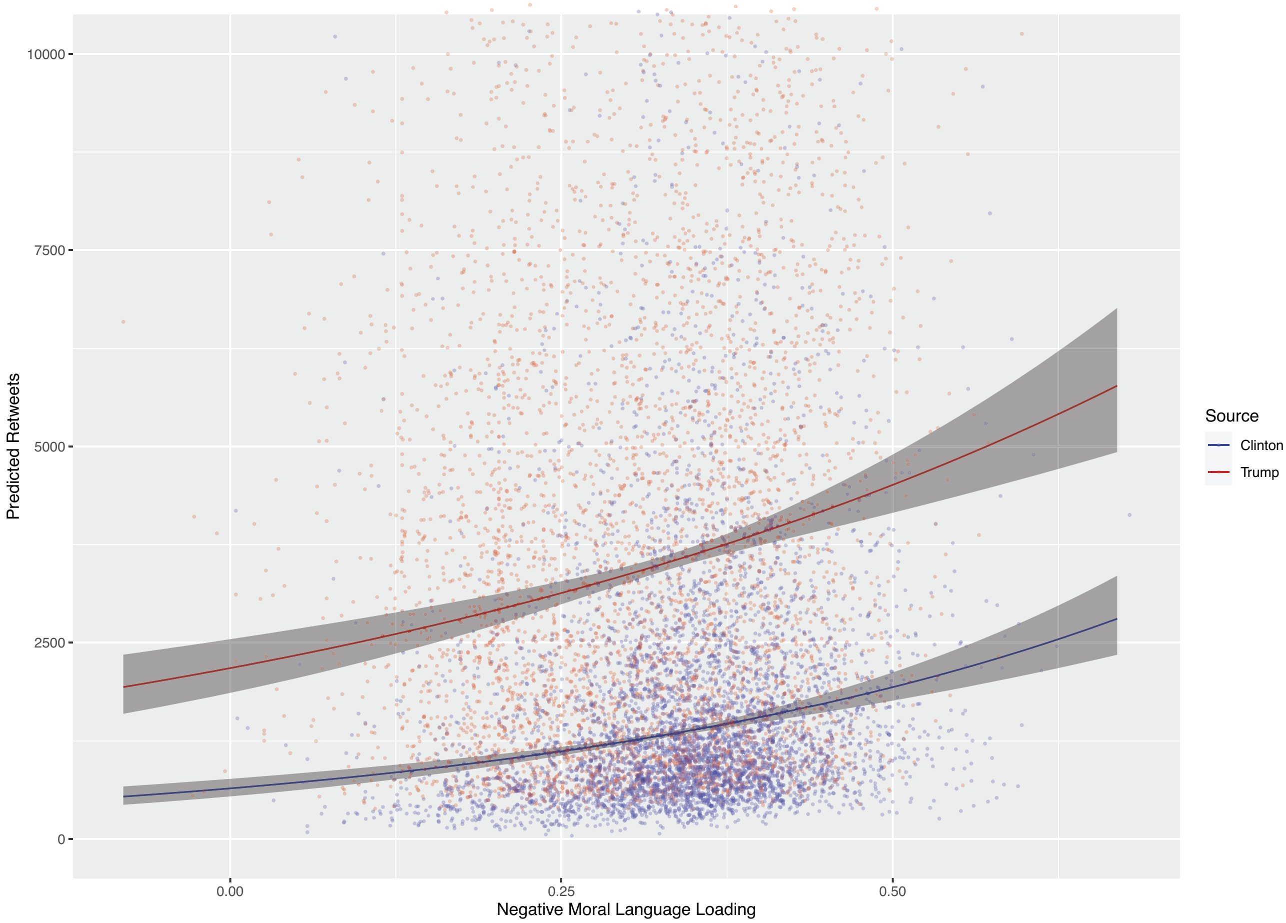
Figure 1
Association Between Positive Moral Language and Predicted Retweets for the 2016 Democratic and Republican Presidential Candidates



Note. See the online article for the color version of this figure.

Figure 2

Association Between Negative Moral Language and Predicted Retweets for the 2016 Democratic and Republican Presidential Candidates



Note. See the online article for the color version of this figure.

interaction between legislator ideology and positive moral language, $b = 1.00$, $t(562.97) = 14.06$, $p < .001$. Probing this interaction at +1/−1 standard deviations of DW-NOMINATE showed that for more conservative legislators (+1 *SD*), positive moral language was not associated with retweets, $b = .10$, $t(615.16) = 1.13$, $p = .26$. For more liberal legislators (−1 *SD*), positive moral language was *negatively* associated with retweets, $b = -1.89$, $t(522.36) = -17.26$, $p < .001$. These results are pictured in Figures 3 and 4; model results are shown in Table 2 (model 2).

These results are robust after accounting for overall positive and negative sentiment expressed in tweets, number of tweets per day, number of tweets by each author per day, and the mean positive and negative moral language by day (see Table 2 [model 3] and online supplemental materials for full details on robustness checks).

Discussion

Our primary aim in this research was to test the association between moral language and message spread online using a measure of moral language that avoids the limitations of word-counting techniques used in previous research (for example, Brady et al., 2017, 2019). Replicating previous research, we find that messages using more moral language spread more widely (that is, they are retweeted more). However, this main effect of moral language is moderated both by speaker ideology and by whether the language was positive (that is, praising moral virtues) or negative (that is, condemning moral vices).

Previous research on whether positive or negative language is more associated with message spread is mixed, with some research finding an advantage for positive moral language (Ferrera & Yang, 2015) and other research finding the opposite (Brady et al., 2019; Schöne et al., 2021). We find that regardless of dataset or speaker ideology, negative moral language is more strongly associated with message spread than positive moral language is. In fact, among liberals in the Congressional dataset, positive moral language is *negatively* associated with message spread. In combination with previous research, the current results give a clear picture: At least when it comes to contemporary U.S. politics, negative moral language spreads more readily than positive.

We also find consistent effects of ideology across data sets: the relationship between negative moral language and message spread

Table 2
Coefficients and Standard Errors From Multilevel Models on Predicting Retweets for Tweets From Members of Congress

Predictor	Model 1	Model 2	Model 3
Positive moral language (CWC)	−0.70 (0.083)***	−0.89 (0.072)***	−0.33 (0.071)***
Negative moral language (CWC)	4.32 (0.13)***	4.63 (0.12)***	3.33 (0.11)***
Positive moral language (account mean)	−5.28 (1.76)**	−5.44 (1.77)**	−0.23 (2.31)
Negative moral language (account mean)	14.34 (1.61)***	15.15 (1.58)***	8.53 (2.54)***
DW-NOMINATE	−0.097 (0.025)***	0.70 (0.071)*	0.48 (0.29)
Followers	0.54 (0.018)***	0.54 (0.018)***	0.57 (0.017)***
Positive Moral Language (CWC) × DW-NOMINATE		1.00 (0.071)***	0.94 (0.068)***
Negative Moral Language (CWC) × DW-NOMINATE		−1.53 (0.11)***	−1.42 (0.11)***
Positive Moral Language (account mean) × DW-NOMINATE		3.98 (1.64)*	4.26 (1.57)**
Negative Moral Language (account mean) × DW-NOMINATE		−7.39 (1.43)***	−6.86 (1.36)***
Mean negative moral language by day			17.35 (0.14)***
Mean positive moral language by day			−1.45 (0.13)***
Number of tweets by that account that day			−0.0056 (0.00012)***
Total tweets that day			−0.016 (0.0,029)***
Positive sentiment (VADER) (CWC)			−0.54 (0.0,098)***
Negative sentiment (VADER) (CWC)			−0.16 (0.015)***
Positive language (account mean)			−3.35 (0.88)***
Negative language (account mean)			0.93 (1.90)

Note. Model 1 Includes positive moral language, negative moral language, legislator ideology (DW-NOMINATE), and account follower count as predictors. Model 2 adds interactions between moral language and legislator ideology. Model 3 adds robustness-check controls. Retweet counts and follower counts are log-transformed. Count of total tweets made that day is log-transformed. DW-NOMINATE scores are scaled.
* $p < .05$. ** $p < .01$. *** $p < .001$.

is stronger for liberal speakers than conservatives. Negative moral language use is more strongly related to retweets for Hillary Clinton than for Donald Trump (although this difference is small and barely below the $p < .05$ significance threshold). It is also more strongly related to retweets for more liberal Members of Congress than more conservative ones (this difference is larger and significant at $p < .001$). Politicians tend to be followed disproportionately by members of their party (Barberá, 2015), suggesting that, at least in the time period covered by our data (2016–2018), negative moral language is more motivating for members of the public on the left than the right. In the same time period, Democratic legislators used more moral language of most kinds on Twitter than did Republicans (Wang & Inbar, 2021), which may be because Democrats' audiences were, on average, more responsive to it.

Integration With Previous Research

Previous findings of a "moral contagion" effect—that is, a relationship between certain kinds of moral language and message diffusion (Brady et al., 2017, 2019)—have recently been questioned by researchers showing that the word-counting approach used in these studies often does not correspond to human coders, and that it may produce spurious positive results (Burton et al., 2021). It is reassuring for the moral contagion hypothesis that we reproduced many of the key findings of this research here while using a natural language analysis technique that is not susceptible to the limitations of word-counting (and that we show has good agreement with human coders).

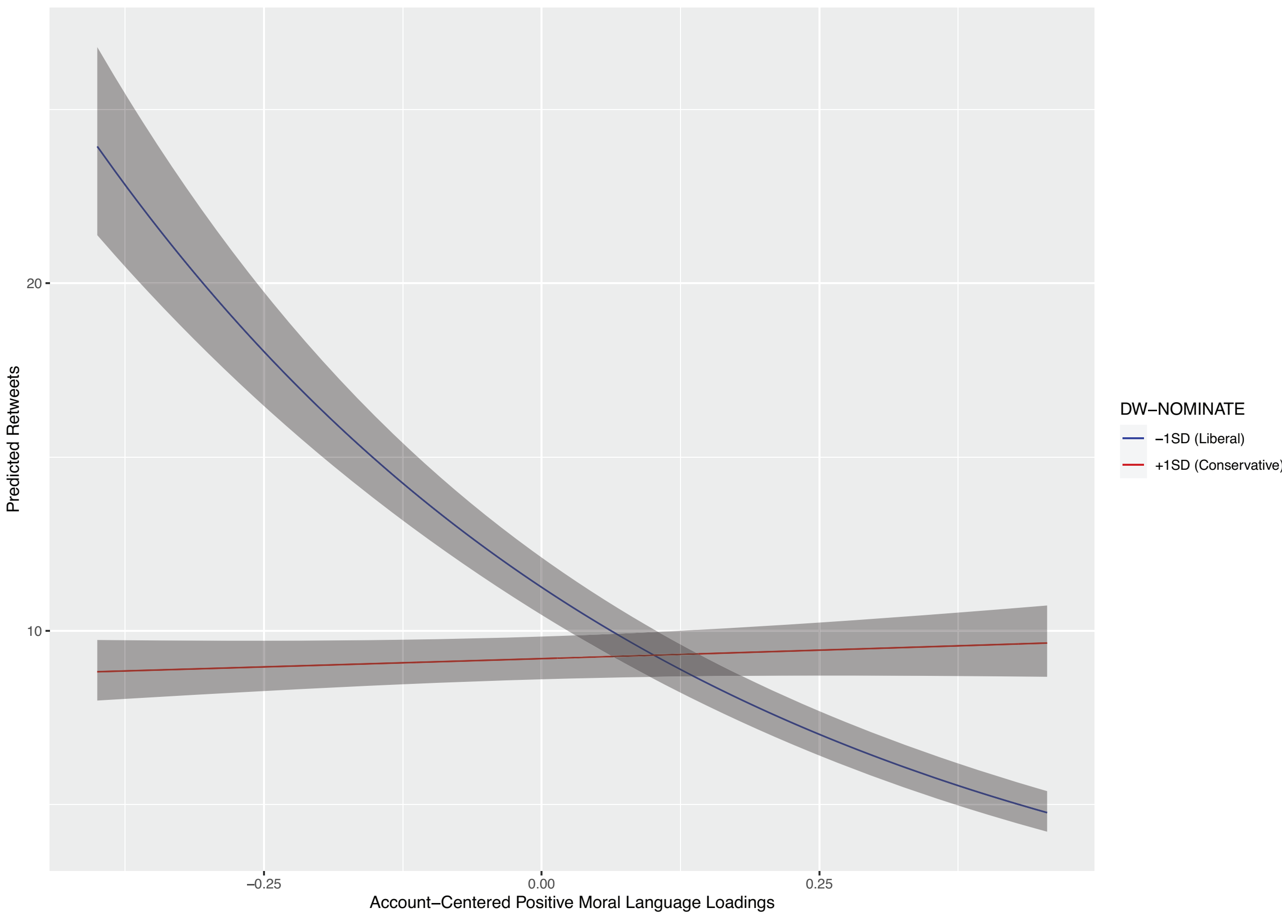
The most important difference between our results and those found in previous research is that we find a robustly stronger association between negative moral language and retweets for more liberal elites. For positive moral language, we find either null or negative associations with retweets for both liberal and conservative elites. In contrast, previous research has found that negative "moral-emotional" language is equally associated with retweets for liberal and conservative elites, whereas positive moral-emotional language is more strongly associated with retweets for conservative elites than liberals (Brady et al., 2019). This might be attributable to differences in the theoretical conceptualization of morality: Brady et al. (2019) differentiate between "moral-emotional" language (that is, language that expresses both morality and emotion) and purely "moral" language, whereas we do not.

However, we think it is more likely that methodological limitations of word-counting—specifically, its dependence on specific and inclusive dictionaries, and thus its sensitivity to researcher choices—are responsible for the divergent results. We examined this possibility in two analyses reported in detail in the online supplemental materials. First, in our Congressional dataset we compared the 200 tweets with the highest "moral-emotional" word count (using the dictionaries from Brady et al., 2019) with the 200 tweets that loaded most highly on moral language using DDR. The discrepancy between language analysis methods appears to be attributable to false positives (that is, low precision) when using word-counting. Of the 200 tweets that contain the most moral words according to word-counting, 49 do not have high DDR scores (that is, within the top 10% of scores). In contrast, of the 200 tweets that load most highly on moral language using DDR, only six contain no moral-emotional words according to word-counting.

Second, we tested how sensitive both DDR and word-counting are to the inclusion of specific terms. We conducted a bootstrapping analysis where we randomly sampled half of the dictionary terms for word counting or half of the seed words for DDR over 30 iterations. We then examined how well estimates using a randomly selected half of the dictionary (or seed words) agreed with the original estimates for the effects of moral-emotional language/negative moral language in the Clinton-Trump dataset. We found that the results for DDR negative moral language were much more consistent across iterations than were results from word-counting

Figure 3

Association Between Positive Moral Language and Predicted Retweets for Members of Congress From January 1, 2016 to January 31, 2018



Note. See the online article for the color version of this figure.

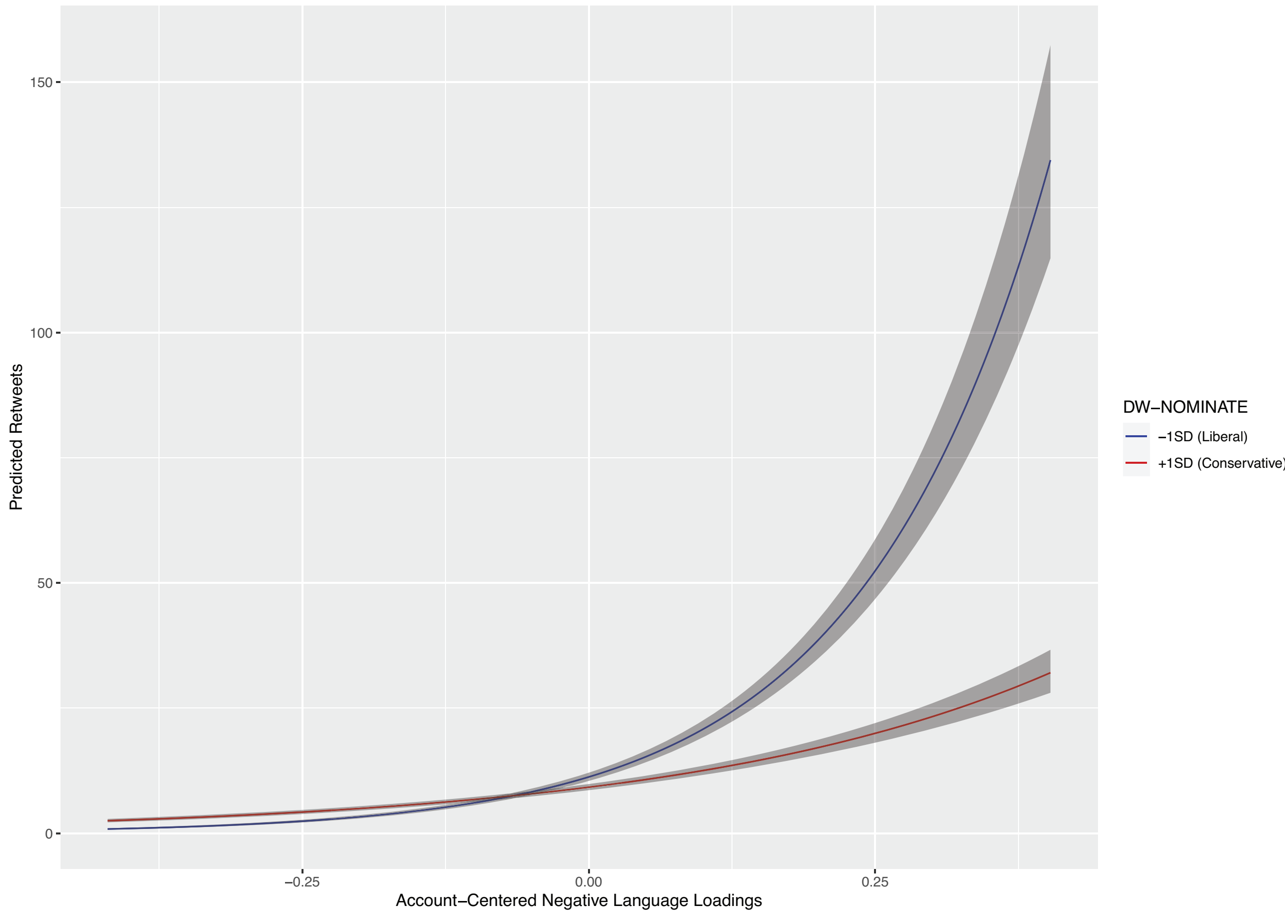
with the moral-emotional dictionary used by Brady et al. (2019). Using this word counting with the moral-emotional dictionary, only 53% of bootstrapped estimates were significant in the same direction as the original, and 7% were significant in the opposite direction. In contrast, for DDR negative moral language estimates were directionally consistent in all cases and significant in 97% of bootstrapped estimates. Combined the low agreement with human coders shown by word-counting here and in other research, these analyses raise doubts about whether word-counting can produce valid results when measuring morality in short texts such as tweets, at least as currently implemented (that is, given the moral-emotional word lists used in Brady et al., 2017, 2019).

Our results also highlight one other problem in interpreting past research, and a cautionary note for future research. In the Congressional dataset, we found that *both* within-account and between-account variation in moral language use was associated with retweets. This means that analyses that do not disaggregate within- and between-account effects may erroneously conclude that there is a causal relationship between moral language use and retweets, when in fact the overall association is due to account-level confounding variables. As an illustration, neither the analyses reported in Brady et al. (2019) nor those in the Supplemental Material of Wang and Inbar (2021) disaggregated within- and between- account effects. We recommend that future researchers disaggregate within- and between-account relationships when possible.

Negative Moral Rhetoric and Message Spread: Is Bad Always Stronger Than Good?

As a rule, negative stimuli have a greater psychological impact than positive stimuli across psychological domains. For example, in financial decision-making, people respond more strongly to losses than to equivalently-sized gains (Kahneman & Tversky, 1979). In impression formation, negative information consistently has a stronger effect on overall impressions than positive information (Baumeister et al., 2001; Fiske, 1980; Skowronski & Carlston, 1989). The ubiquitous phenomenon of contamination is another example: A single cockroach is enough to spoil a kettle of soup, but a kettle of roaches is just as revolting with a small amount of a positive substance added (Rozin & Royzman, 2001). This very general phenomenon has been called

Figure 4
Association Between Negative Moral Language and Predicted Retweets for Members of Congress From January 1, 2016 to January 31, 2018



Note. See the online article for the color version of this figure.

"negativity bias" (Rozin and Royzman, 2001) or "bad is stronger than good" (Baumeister et al., 2001).

The very breadth of negativity bias strongly suggests that it has multiple causes, but one cause seems to be that people generally pay more attention to negative than positive information. For example, people automatically attend more to negative than positive personality traits (Pratto & John, 1991), and when assessing their own performance, they are more aware of barriers that held them back than to benefits that have helped them succeed (Davidai & Gilovich, 2016). It may be that there is a similar asymmetry for moral information, such that negative moral information more readily captures attention than does positive moral information (see Brady et al., 2020) for a similar argument regarding "attentional capture" by moral rhetoric, although they do not distinguish between negative and positive moralizing). Although this seems plausible, it would not explain why the link between negative moral rhetoric and retweets was stronger for more liberal legislators. And, more broadly, the notion that people attend more to negative information is inconsistent with research finding that *positive* sentiment in a message is associated with greater spread on Twitter (Ferrera & Yang, 2015).

We propose that threat, and particularly group-based threat, may be a key moderator. Brady et al. (2020) argued that online discussions of politics naturally make partisan group identity salient for many people, leading to a greater tendency to moralize. They claimed that this should be the case both for negative moral rhetoric (outgroup derogation) and for positive moral rhetoric (ingroup praise). We propose that although this may be true, in the case of perceptions of intergroup threat, *negative* moral rhetoric becomes relatively more appealing than positive—that is, moral rhetoric condemning the perfidy of one's opponents may become relatively more appealing than rhetoric praising the virtues of one's allies. The logic of this is straightforward: Almost by definition, intergroup threat makes actual or possible bad outcomes salient (Stephan & Stephan, 2000). These bad outcomes then then take on primary motivational relevance. As Baumeister et al. (2001, p. 325) put it, "Survival requires urgent attention to possible bad outcomes, but it is less urgent with regard to good ones." In most of the time period covered by our Twitter data, Democrats had suffered a shocking loss in the presidential election and held neither the Senate nor the House of Representatives. Plausibly, therefore, they were experiencing more threat than Republicans, making negative moral rhetoric particularly appealing to Democrats during this time period. Admittedly, this explanation is somewhat speculative and does not as readily explain the difference in response to negative moral rhetoric from Clinton versus Trump (although this difference was much weaker). We therefore see it not as a settled fact but rather as an avenue for future research. For example, Wang and Inbar (2021) found that moral rhetoric of all kinds from legislators increases when their party is out of power. If the logic outlined here is correct, this should particularly be the case for negative moral rhetoric.

Implications for Moral Foundations Theory

Moral foundations theory (Haidt & Graham, 2007) posits five moral domains ("foundations") thought to constitute the basic building blocks of morality across cultures. The "individualizing" foundations (harm and fairness) concern individual rights and well-being, whereas the "binding" foundations of loyalty, authority, and purity concern adherence to norms that maintain group cohesion. In the United States and other countries, liberals and conservatives differ in how much they consider these foundations to be morally relevant. Both liberal and conservative members of the public rate the individualizing foundations as morally relevant (although liberals endorse them somewhat more strongly). The binding foundations, however, are endorsed much more strongly by conservatives than liberals (Graham et al., 2009). Based on these data, a natural prediction is that moral rhetoric relevant to the binding foundations should appeal more to conservatives, and moral rhetoric relevant to the individualizing foundations should apply more to liberals.

Because our seed word dictionaries contained words tapping virtue and vice aspects of each of the five moral foundations, we could test whether the relationship between moral rhetoric, political ideology, and message spread is moderated by a match between moral foundation and legislator ideology. Is moral language referencing loyalty, authority, and purity more associated with retweets for conservative legislators, and moral language referencing harm and fairness more associated with retweets for liberal legislators? In general, the answer is "no." As Table S2 in the online supplemental materials shows, for every moral foundation taken individually, negative moral rhetoric was associated with retweets more strongly overall, and more so for liberals than conservatives. Although it is inconsistent with the self-report differences between liberals and conservatives described above, this finding is consistent with other research finding that Democratic and Republican politicians do not differ in their use of moral rhetoric in the way Moral Foundations Theory predicts (Sagi and Dehghani, 2014; Sterling and Jost, 2018; Wang & Inbar, 2021). It may yet turn out to be the case that the predictions of Moral Foundations Theory are supported when moral language use is analyzed differently (e.g., topic modeling might reveal that conservative legislators are more likely to discuss topics relevant to the binding foundations). However, currently the track record of Moral Foundations theory is not particularly good when the predictions concern the moral rhetoric used by elites. Future research could explore what is responsible for this. Is it that elites differ from the public in how they use moral language? Are the liberal-conservative differences in moral foundation relevance evident in self-report measures but not in observed behavior? Ultimately, these questions will need to be answered by future research.

Constraints on Generality

The results reported here are from a specific time (2016–2018) and political system (all political elites included in the sample were American, although the Twitter users retweeting them might not have been). Previous research has shown that partisan differences in moral language use change over time, with legislators using more moral language when their party is out of power (Wang & Inbar, 2021), so the partisan differences we observe here may not generalize to other time periods. Even the diffusion advantage of negative over positive moral language might not generalize to countries or eras characterized by lower polarization or political conflict, as some segments of the public may find negative moral rhetoric aversive (e.g., political independents, Klar & Krupnikov, 2016).

As with past work, the current research focuses on retweets as a measure of moral contagion. Although Twitter is an important platform for political discussion, and one that has been widely adopted by political elites, future work could assess moral contagion across other platforms both online and offline. Other forms of engagement on Twitter, like the content of replies, could also be analyzed.

Conclusion

We examine the association between moral language from political elites on Twitter and message spread using distributed dictionary representations, a method that we demonstrate has excellent agreement with trained human coders. We find that tweets containing more moral language are retweeted more, that this is specific to negative moral language, and that this relationship is strongest for more liberal elites.

References

Aral, S. (2020). *The hype machine: How social media disrupts our elections, our economy, and our health—and how we must adapt. Currency.*

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD Model of Moral Contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Gantman, A. P., & Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, *149*(4), 746–756. https://doi.org/10.1037/xge0000673

Brady, W. J., Wills, J. A., Burkart, D., Jost, J. T., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*, *148*(10), 1802–1813. https://doi.org/10.1037/xge0000532

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks.

Proceedings of the National Academy of Sciences of the United States of America, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635. https://doi.org/10.1038/s41562-021-01133-5

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, *75*(3), 659–671. https://doi.org/10.1017/S0022381613000492

Davidai, S., & Gilovich, T. (2016). The headwinds/tailwinds asymmetry: An availability bias in assessments of barriers and blessings. *Journal of Personality and Social Psychology*, *111*(6), 835–851. https://doi.org/10.1037/pspa0000066

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old iss. *Psychological Methods*, *12*(2), 121–138.

Ferrera, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PLoS ONE*, *11*(10), 14.

Fiske, A. P., & Rai, T. S. (2014). *Virtuous violence*. Cambridge University Press. https://doi.org/10.1017/CBO9781316104668

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037/0022-3514.38.6.889

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods*, *50*(1), 344–361. https://doi.org/10.3758/s13428-017-0875-9

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), 1029–1046. https://doi.org/10.1037/a0015141

Haidt, J., & Graham, J. (2007). When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Hamaker, E. L., & Bengt, M. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379.

Hare, C., & Poole, K. T. (2014). The polarization of contemporary American politics. *Polity*, *46*(3), 411–429. https://doi.org/10.1057/pol.2014.10

Hoover, J., Johnson-Grey, K. M., Dehghani, M., & Graham, J. (2017). *Moral values coding guide*. https://doi.org/10.31234/osf.io/5dmgj

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T. E., Chin, J., Leong, C., Leung, J. Y., Mirinjian, A., & Dehghani, M. (2020). Moral foundations Twitter corpus: A Collection of 35k tweets annotated for moral sentiment. *Social Psychological & Personality Science*, *11*(8), 1057–1071. https://doi.org/10.1177/1948550619876629

Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Klar, S., & Krupnikov, Y. (2016). *Independent politics: How American disdain for parties leads to political inaction*. Cambridge University Press.

Lakoff, G. (2004). *Don't think of an elephant!: know your values and frame the debate: the essential guide for progressives*. Chelsea Green Pub. Co.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Marantz, A. (2019). *Antisocial: Online extremists, techno-utopians, and the hijacking of the American conversation*. Viking.

Marietta, M. (2008). From my cold, dead hands: Democratic consequences of sacred rhetoric. *The Journal of Politics*, *70*(3), 767–779. https://doi.org/10.1017/S0022381608080742

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*. https://arxiv.org/pdf/1301.3781v1

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396. https://doi.org/10.1038/s41562-018-0353-0

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Poole, K. T., & Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, *29*(2), 357–384. https://doi.org/10.2307/2111172

Powers, D. M. W. (2011). Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391. https://doi.org/10.1037/0022-3514.61.3.380

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). SAGE Publications.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Ryan, T. (2017). No compromise: Political consequences of moralized attitudes. *American Journal of Political Science*, *61*(2), 409–423. https://doi.org/10.1111/ajps.12248

Sagi, E., & Dehghani, M. (2014). Measuring Moral Rhetoric in Text. *Social Science Computer Review*, *32*(2), 132–144. https://doi.org/10.1177/0894439313506837

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity Spreads More than Positivity on Twitter After Both Positive and Negative Political Situations. *Affective Science*, *2*, 379–390.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257–268. https://doi.org/10.1093/ptj/85.3.257

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267–281. https://doi.org/10.1111/j.1751-9004.2010.00254.x

Skitka, L. J., & Mullen, E. (2002). Understanding judgments of fairness in a real-world political context: A test of the value protection model of justice reasoning. *Personality and Social Psychology Bulletin*, *28*(10), 1419–1429. https://doi.org/10.1177/014616702236873

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In S. Oskamp (Ed.), *Reducing prejudice and discrimination* (pp. 23–45). Erlbaum Publishers.

Sterling, J., & Jost, J. T. (2018). Moral discourse in the Twitterverse. *Journal of Language and Politics*, *17*(2), 195–221. https://doi.org/10.1075/jlp.17034.ste

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tufekci, Z. (2017). *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

Wang, S. N., & Inbar, Y. (2021). Moral-language use by U.S. political elites. *Psychological Science*, *32*(1), 14–26. https://doi.org/10.1177/0956797620960397

Received April 28, 2021
Revision received March 28, 2022
Accepted March 31, 2022 ■