Bound in Hatred: The role of group-based morality in acts of hate

Joe Hoover, Mohammad Atari*, Aida Mostafazadeh Davani*, Brendan Kennedy*,
Gwenyth Portillo-Wightman, Leigh Yeh, Drew Kogon & Morteza Dehghani
University of Southern California

Abstract

Acts of hate have been used to silence, terrorize, and erase marginalized social groups throughout history. The rising rates of these behaviors in recent years underscores the importance of developing a better understanding of when, why, and where they occur. In this work, we present a program of research that suggests that acts of hate may often be best understood not just as responses to threat, but also as morally motivated behaviors grounded in people's moral values and perceptions of moral violations. As evidence for this claim, we present findings from five studies that rely on a combination of natural language processing, spatial modeling, and experimental methods to investigate the relationship between moral values and acts of hate toward marginalized groups. Across these studies, we find consistent evidence that moral values oriented around ingroup preservation are disproportionately evoked in hate speech, predictive of the county-level prevalence of hate groups, and associated with the belief that acts of hate against marginalized groups are justified. Additional analyses suggest that the association between group-oriented moral values and hate acts against marginalized groups can be partly explained by the belief that these groups have done something morally wrong. By accounting for the role of moralization in acts of hate, this work provides a unified framework for understanding hateful behaviors and the events or dynamics that trigger them.

Bound in Hatred: The role of group-based morality in acts of hate

Throughout history, humans have discriminated against, persecuted, and murdered other humans because of their identities (Kiernan, 2007; Moore, 2000; Nirenberg, 2015). In addition to the horrors of death and extermination, such acts of hatred have tragic effects on survivors and survivors' communities. Victims of hate crime, for example, experience higher levels of depression and anxiety compared to victims of comparable crimes not motivated by bias (Hall, 2013) and they may ultimately reject or despise the part of their identity that was targeted (Cogan, 2002). Even for people who are not directly victimized, just sharing a trait targeted by a hate crime can cause clinical levels of post-traumatic stress (Government of Canada, Department of Justice, Research, & Statistics Division, 2011). Online racial discrimination has also been linked to elevated levels of depression and anxiety in victims (Tynes, Giang, Williams, & Thompson, 2008).

Tragically, the human tendency toward identity-based hatred and violence remains a major contributor to human suffering. In the 20th-21st century, genocide was one of the leading causes of preventable violent death (Blum, Stanton, Sagi, & Richter, 2008). In recent years, hate crime in the United States (Center for the Study of Hate & Extremsism, 2018; Eligon, 2018) and Europe (Engel et al., 2018) has systematically increased, with U.S. incidence reports reaching their highest levels since the September 11th attack on the World Trade Center. The number of hate groups operating in the U.S. has also recently reached a record high (SPLC, 2019), and concerns over the rising prevalence of online hate speech have led to shifts in social media content policies (Beckett, 2019; Conger, 2019; Frenkel, Isaac, & Conger, 2018).

These trends highlight the importance of developing a better understanding of why, when, and where acts of hate occur. Research addressing these questions has often focused on the roles of inter-group threat, resource based conflicts, and political ideology as focal mechanisms in the emergence of behaviors like hate crime (Hall, 2013; Stacey, Carbone-López, & Rosenfeld, 2011a), hate group activity (McCann, 2010; McVeigh, 2004; McVeigh & Sikkink, 2005; Medina, Nicolosi, Brewer, & Linke, 2018),

and hate speech (Piatkowska, Messner, & Yang, 2018). Echoing these findings, psychological investigations of attitudinal prejudice have consistently observed that perceptions of either realistic or symbolic outgroup threat (Stephan & Stephan, 1996, 2017) lead to increased prejudice toward outgroups and that this effect is positively mediated by attitudes associated with authoritarianism and social dominance (Asbrock, Sibley, & Duckitt, 2010; Charles-Toussaint & Crowson, 2010; Cohrs & Ibler, 2009; Duckitt & Sibley, 2009, 2017).

Together, this line of work suggests behaviors like hate crime, hate group activity, and hate speech can be at least partly understood as responses to perceived outgroup threats (Hall, 2013). However, this account begs an essential question: what is it about some threats — and the people who perceive them — that is sufficient for inducing such costly behaviors, behaviors that can lead to social exclusion, retaliation, and the heaviest of criminal penalties?

To answer this question, we propose that the *moralization* of a threat is a central factor in the motivational process underlying acts of hate such as hate speech, hate group activity, and hate crime — behaviors we refer to collectively as extreme behavioral expressions of prejudice (EBEPs). This view is grounded in a large body of research linking violence and extreme behavior to moral values, perceptions of moral violations, and feelings of moral obligation (Atran & Ginges, 2012; Darley, 2009; Fiske, Rai, & Pinker, 2014; Graham & Haidt, 2011; Mooijman, Hoover, Lin, Ji, & Dehghani, 2018; Rai, 2019; Skitka, Hanson, & Wisneski, 2017; Zaal, Van Laar, Ståhl, Ellemers, & Derks, 2011). Drawing on this work, we suggest that EBEPs are often motivated by the belief that an outgroup has done something morally wrong and, further, that a person's risk of perceiving such moral violations is partially dependent on their moral values — a hypothesis we refer to as the Moralized Threat Hypothesis.

By accounting for the role of moralization in EBEPs, this perspective provides a unified framework for understanding why certain events or dynamics trigger hateful behaviors. Under this hypothesis, EBEPs can be understood as a perpetrators' response to a perceived violation of their moral values. Thus, EBEP triggers that have largely

been studied in isolation, such terrorist attacks, (Byers & Jones, 2007; Hanes & Machin, 2014), immigration (Stacey, Carbone-López, & Rosenfeld, 2011b), same sex marriage (Levy & Levy, 2017; Valencia, Williams, & Pettis, 2019), interracial romantic relationships (Perry & Sutton, 2006, 2008) or the espousal of non-Western religious values (Green & Spry, 2014; Velasco González, Verkuyten, Weesie, & Poppe, 2008), are conceptualized as perceived moral violations. Accordingly, from the perpetrators' perspective, EBEPs function as a mechanism for regulating social relations (Fiske et al., 2014; Rai & Fiske, 2012) with the outgroup that is blamed for the moral violation.

To test the Moralized Threat Hypothesis, we relied on a diverse set of observational and experimental methodologies in order to investigate the role of moral values in both real-world EBEPs and beliefs about the justification of EBEPs. Given recent increases in EBEPs aligned with right-wing ideology (Eligon, 2018; Engel et al., 2018; Lowery, Kindy, & Tran, 2018) and concerns over the role of hate speech in violent crimes toward social identities often demonized by right-wing groups (Roose, 2018), we focused specifically on EBEPs that were aligned with right-wing ideologies. Accordingly, we expected that these EBEPs would be associated with moral values oriented around group preservation because such values have been linked to conservatism and right-wing ideologies in U.S. contexts (Frimer, Biesanz, Walker, & MacKinlay, 2013; Graham, Haidt, & Nosek, 2009). To operationalize these values, we rely on Moral Foundations Theory (MFT; Graham et al., 2013, 2011), which proposes a hierarchical model of moral values composed of two superordinate, bipolar categories: Individualizing values and Binding values. While the former is comprised of values focused on individuals' rights — caring for others and following principles of fairness — the latter is comprised of values considered to be associated with group preservation — maintaining ingroup solidarity, submitting to authority, and preserving the purity of the body and sacred objects.

Using this model of moral values, the Moralized Threat Hypothesis predicts that Binding values are associated with EBEPs toward groups marginalized by the ideological right. We examine this prediction across five studies. In Study 1, we use a

series of Long Short-Term Memory (Hochreiter & Schmidhuber, 1997) neural network models to study online hate speech and test the hypothesis that moralization and hate speech are concomitant. In Study 2, we move out of the digital world and focus on the geospatial relationship between U.S. county-level moral values and the county-level prevalence of hate groups. Then, in Studies 3, 4, and 5, we switch from naturally generated data to data collected via psychological surveys, which enable us to test our hypotheses with more precision and control. In Study 3, we investigate whether people see a range of EBEPs as more justified when they believe that the targeted outgroup has done something morally wrong. Then, in Study 4, we test for associations between American's group-oriented moral values and perceived justification of EBEPs against Mexican immigrants and investigate whether this association can be accounted for by their perceptions of outgroup moral wrongdoing. Finally, in Study 5, we investigate these effects for a different outgroup (Muslims) using a national U.S. sample stratified across participants' gender, age, and political ideology.

Together, these studies rely on a multi-methodological approach that uses observational and experimental research designs to test hypotheses against distinct operationalizations and measurements of EBEPs and group-oriented moral values. Relying on this approach enables us to directly study the phenomena of interest, EBEPs, while also maintaining the precision and control afforded by more traditional approaches to psychological research. Across all five studies, we found consistent evidence that extreme behavioral expressions of hatred and the belief that they are justified are associated with the Binding values. Further, data from the three surveys we conducted indicate that this association can be at least partly explained by the perception of outgroup moral wrongdoing. Notably, these estimated effects remain substantial even after adjusting for participants' political ideology.

Study 1: Hate Speech and Moral Rhetoric

Our first step toward testing the Moralized Threat Hypothesis is to investigate the relationship between expressions of hate speech in social media posts and the

concomitant reliance on moral rhetoric evoking the so-called Binding *vices*. Under MFT, each component of the Individualizing and Binding values is associated with two valences or poles: virtues (i.e. prescriptive moral concerns) and vices (i.e. proscriptive moral concerns). Conditional on the Moralized Threat Hypothesis, which holds that EBEPs are motivated by perceived moral violations, we hypothesized that online hate speech is most often articulated through the language of the Binding vices — language evoking concerns about violations of loyalty, authority, and purity.

To test this hypothesis, we focused on the social media platform Gab, which was recently embroiled in controversy following its role in the October 2018 terrorist attack on a synagogue in Pittsburgh, Pennsylvania (Roose, 2018). Gab purports to be a haven for free speech and has attracted a large membership of users who align themselves with far-right ideologies (Anthony, 2016; Benson, 2016). This emphasis on free speech entails the absence of any institutional oversight of content and, in contrast to mainstream social media platforms, Gab users are free to post anything, including hate speech and incitations of violence against marginalized groups.

Accordingly, Gab presents a valuable opportunity to investigate the rhetorical structure of hate speech because the combined effects of the ideological biases of its users and the absence of content moderation mitigate issues posed by the statistical rarity of hate speech on mainstream social media platforms. Indeed, a recent analysis of hate speech on Gab found that hate words (e.g. racial and group-oriented slurs) occur at 2.4 times the rate of hate words on Twitter (Zannettou et al., 2018). This relative prevalence of hate helps mitigate issues of sparsity that have been encountered in other computational studies of hate speech in social media discourse (Del Vigna, Cimino, Dell'Orletta, Petrocchi, & Tesconi, 2017; Saleem, Dillon, Benesch, & Ruths, 2017; Zhang, Robinson, & Tepper, 2018).

To investigate the relationship between hate speech and moral rhetoric evoking the Binding vices, we first annotated 7,692 messages posted by 800 Gab users for hate-based rhetoric (Kennedy et al., 2018) and moral sentiment (Hoover, Johnson-Grey,

Dehghani, & Graham, 2017)[1]. For examples of Gab posts labeled as hate-based rhetoric, which were also labeled positively for one of the moral vices, see Table 1. Using these annotated messages, we then trained two Long Short-Term Memory (LSTM) neural network models (Hochreiter & Schmidhuber, 1997) to detect hate speech and moral sentiment in Gab posts. LSTMs incorporate a recurrent structure that encodes long-term dependencies between words and their past context. This makes them particularly effective for quantifying the sequential and context-dependent nature of semantic structures in natural language.

The first model that we developed, a single-task LSTM model, was trained to predict whether posts contain hate speech. In contrast, the second model, a multi-task LSTM model (Collobert & Weston, 2008), was trained to predict the presence of each moral vice simultaneously. In both the single-task and multi-task models, posts are represented as matrices of pretrained GloVe word embeddings (Pennington, Socher, & Manning, 2014) corresponding to the words in the original post. This embedding matrix is then input to a 100 dimensional LSTM layer which is connected to a layer of fully connected units, with 0.33 dropout ratio (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). A softmax transformation is then applied to the output of the final layer in order to generate probabilistic predictions for the outcome (See Figure 1).

We then used these models to independently predict the presence of hate speech and each moral vice in the complete Gab corpus, which consists of 24,978,951 posts from 236,823 users, after removing posts with too few English tokens (Gaffney, 2018). The model trained to predict hate obtained an average F1 score of 0.62, precision of 0.65 and recall of 0.59 using 10 fold cross-validation. For the five jointly predicted moral vices in the multi-task model, F1 scores were 0.55 for “harm”, 0.46 for “cheating”, 0.55

[1] These posts were obtained by scraping 121,067 posts from 1,682 Gab users. Because there is no API available for Gab, we implemented a crawler that performs breadth-first graph traversal in order to collect network information for each user. With this network data, we then scraped each user’s page for their posts. From this data, we then identified the set of users with at least 10 posts ($N = 800$) and selected their 10 most recent posts for annotation. Of these, 308 were blank, yielding a final set of 7,692 posts.

Moral Vice	Gab Text
Harm	If you are right-wing and pay to send your daughter to college, you are retarded. Woman are submissive and much more prone to peer pressure. They are not equipped (NAWALT) to handle life in a re-education center. By sending her to college, you are destroying her future and increasing the risk she'll die in childless misery
Cheating	Britain is a country where a rich black man can set up a university scholarship for black people ONLY & if you question if it's 'racist' YOU are 'Racist'. But if a white man set one up for whites only, there would be no doubt it's 'racist' & he would likely be prosecuted for race hate.
Betrayal	Anyone surprised that another clueless May appointee hates the white working class while loving immigrants.
Subversion	Any honest Jewish person should be able to admit the truth, if they're honest. Anywhere across time and space in the Western world their people have gained a majority of power has rapidly descended into chaos and turned into a shithole country.
Degradation	FUCK LONDON ! Disgusting ! I WILL REMMEBER THIS DISRESPECT .... Everytime there is an Islamic terror attack ..... LONDON IS INFECTED .... I HOPE THE ENRICHMENT LEVELS BREAK THE SCALE IN THE COMING YEARS .....THATS ALL

Table 1

Examples of hate speech-labeled Gab posts. Each was also labeled with at least one moral vice, denoted in the "Moral Vice" column

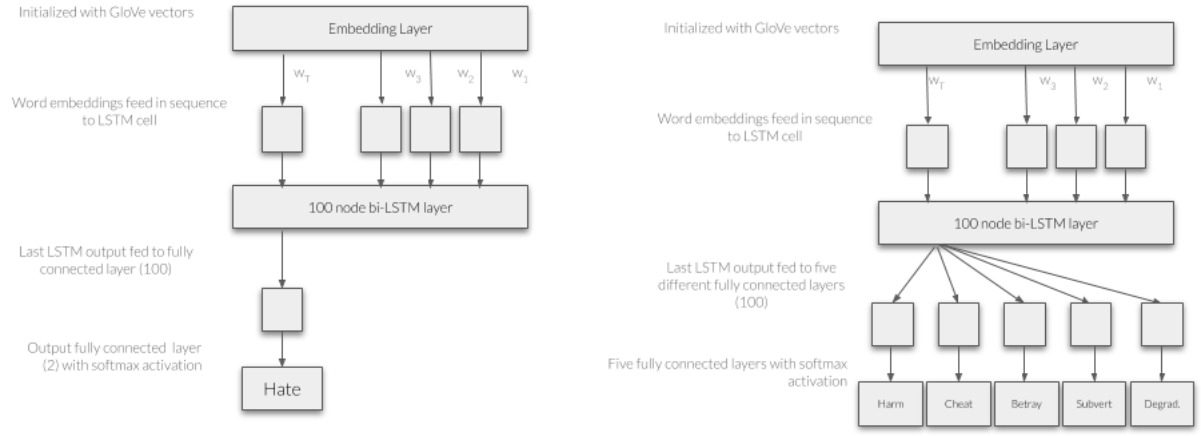

Figure 1. Visualizations of the classification models trained on Gab data and used to predict on entire Gab corpus.

for "betrayal", 0.44 for "subversion", and 0.44 for "degradation".

Finally, using these predicted labels, we tested the hypothesis that posts evoking either the Binding or Individualizing vices were more likely to contain hate speech. Using the union of the predicted labels for the Individualizing and Binding vices, we then estimated the probability that a given post contains hate speech as a function of whether or not it is labeled as evoking Individualizing or Binding vices. To account for the fact that this corpus contains multiple messages per user, we estimated this probability using a hierarchical logistic regression model in which varying intercepts were estimated for each user and the effects of Individualizing and Binding vices were permitted to vary across users. To minimize the possibility that model estimates were biased by messages generated by bots, we estimated two separate hierarchical logistic regression models. The first model was trained on a subset of the full corpus that only included posts by users who made fewer than 500 total posts (N_{posts} = 4,994,480; N_{users} = 229,538). The second model was trained on the entire Gab corpus.

Both models indicated a strong association between the Binding vices and the presence of hate speech. As model results were comparable, here we focus on the results from the first model (See 2 in Supplemental Material for estimates from the model trained on the full corpus). Specifically, after adjusting for the effect of Individualizing vices, posts labeled as evoking Binding vices had approximately 25 times the risk of

containing hate speech compared to posts that did not evoke the Binding vices, $b = 3.22$, $SE = 0.01$, $Z = 295$, Risk Ratio (RR) = 25.12. While the Individualizing vices were also positively associated with hate speech, their estimated effect was substantially smaller, such that a post that evokes the Individualizing vices approximately 6 times the risk of containing hate speech compared to posts that do not evoke the Individualizing vices, $b = 1.79$, $SE = 0.01$, $Z = 126$, RR = 5.92. Notably, while the effects of Binding and Individualizing vices showed variation across users ($SD_{b_{Binding}} = 0.79$, $SD_{b_{Individualizing}} = 0.86$), fixed effects estimates indicated that, on average, they were both positively associated with the presence of hate speech.

Together, these results are consistent with the hypothesis that hate speech tends to be articulated through the language of morality. They also indicate that the Binding values are particularly relevant to hate speech, as messages predicted as evoking the Binding vices were nearly 40% more likely to be predicted as hate speech, compared to messages predicted as evoking the Individualizing vices. This differential association suggests that articulations of hate speech rely much more strongly on language that evokes group-based moral values, compared to language that evokes individual-based moral values. Importantly, this is consistent with the Moralized Threat Hypothesis, which proposes that the perception of outgroup moral violations is a central risk factor for EBEPs. Indeed, overall, our results suggest that a sense of group-based moral violation is often encoded directly in articulations of hate speech.

Study 2: Hate Groups and County-level Moral Values

In Study 1, we found evidence that real-world hate speech often invokes Binding or group-based moral concerns, which is consistent with the hypothesized role of moral values in EBEPs. In Study 2, we extend this finding by investigating whether a comparable association exists between Binding moral values and a different EBEP, real-world hate group activity. That is, we test the hypothesis that county-level moral values — specifically the Individualizing and Binding values — are associated with the county-level prevalence of hate groups. Per our findings in Study 1, we expect to find a

positive association between Binding values and the prevalence of hate groups. Further, we expect that this effect will be stronger than the observed effect of the Individualizing values. Importantly, as in Study 1, this design enables a direct application of the Moralized Threat Hypothesis, which maintains that people's values should influence their perceptions of outgroup moral violations. If this is the case, there should be an association between the moral values held in a county and the prevalence of hate groups in that county, as certain configurations of county-level moral values should increase the local risk of hate group prevalence.

To estimate the county-level distribution of moral values, we use data collected via YourMorals.org, a website operated by the founders of MFT to collect measurements of voluntary respondents' moral values, from approximately 2012 to 2018 (N = 106,465). While this is a relatively large sample, it cannot be used to directly estimate county-level moral values because it is not randomly sampled or representative at the county level (Hoover & Dehghani, 2018). To account for these issues, we rely on Multilevel Regression and Synthetic Poststratification (MrsP; Leemann & Wasserfallen, 2017), a model-based approach to survey adjustment and sub-national estimation that extends Multilevel Regression and Poststratification (MrP; Park, Gelman, & Bafumi, 2004).

Both MrP and MrsP involve estimating regional outcomes on a target construct from individual-level data. This data is used to model the target construct as a multilevel function of demographic characteristics (e.g. gender, age, and level of education), regional indicators (e.g. county, state, or region), and regional factors (e.g. presidential vote proportion, median income, or educational attainment). This model is then used to generate predictions for each combination of demographic characteristics *within each region*. Finally, information about the population distribution of these demographic characteristics within each region are used to estimate a weighted mean based on the model predictions.

Here, we use MrsP, which follows the above approach, but also enables the inclusion of a more diverse set of demographic variables. Specifically, we model

individual-level responses to each moral foundation as a function of six demographic variables: gender (2 levels), age (3 levels), ethnicity (4 levels), level of education (3 levels), religious attendance (3 levels), and political ideology (3 levels). We also account for two levels of regional clustering, the county level and the region level and include the proportion of Democratic votes in the 2016 presidential election as a county-level factor. Finally, the multilevel model that we estimate also includes a hierarchical auto-regressive prior (Riebler, Sørbye, Simpson, & Rue, 2016) that, under the presence of spatial auto-correlation, induces local spatial smoothing (Hanretty, Lauderdale, & Vivyan, 2016; Hoover & Dehghani, 2018; Selb & Munzert, 2011) between proximate counties[2].

Using this approach, we estimated the county-level distribution of each Moral Foundation [3] and then used these estimates to calculate scores for the Individualizing (Care and Fairness) and Binding (Loyalty, Authority, and Purity) dimensions for each county (See Figure 2).

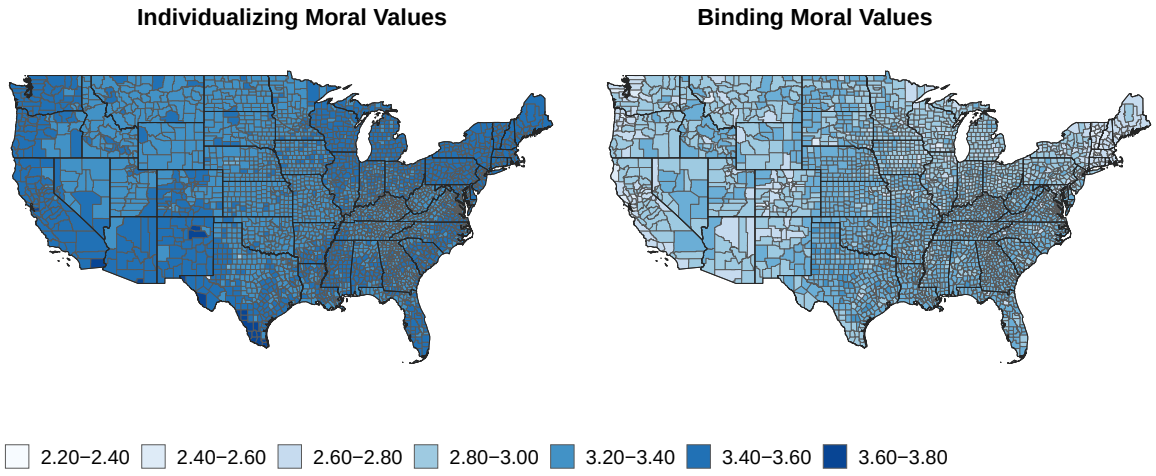

Figure 2. Estimates of county-level Individualizing and Binding Moral Foundations adjusted for representativeness via MrsP with spatial smoothing

Finally, we use these county-level estimates of Individualizing and Binding values

[2] For a detailed discussion of these methodologies and models, as well as an evaluation of the efficacy of using this approach with non-random, non-representative data see Hoover and Dehghani (2018).

[3] An interactive visualization of these estimates can be viewed at`https://mapyourmorals.usc.edu/`

to predict the county-level prevalence of hate groups. Estimates of this outcome were obtained from the Southern Poverty Law Center (SPLC; Southern Law Poverty Center, 2019), which maintains an ongoing hate group task force that monitors and documents hate group activity at the city level. We used this data to generate county-level counts of active hate groups by identifying the counties containing each city[4]. Finally, because the data used to estimate county-level moral values was collected from 2012-2018, we calculated the average county-level count of active hate groups from 2012 to 2017 (the latest available year in the SPLC data).

Using this measurement of hate-group prevalence as the outcome, we estimated the county-level rate of hate groups per 10,000 inhabitants using a Negative-binomial regression model with a proper conditional auto-regressive prior (Leroux, Lei, & Breslow, 2000; Lindgren & Rue, 2015) to account for spatial auto-correlation. A zero-inflated Negative Binomial model was also considered; however, comparisons of leave-one-out cross-validation estimates of model fit (Held, Schrödle, & Rue, 2010; Vehtari, Gelman, & Gabry, 2017; Vehtari, Mononen, Tolvanen, Sivula, & others, 2016) suggested worse fit for the zero-inflated model ($elpd = -4674.90, SE = 340.77$), compared to the Negative Binomial model ($elpd = -4296.17, SE = 304.07$; $elpd_{difference} = -378.72$, $SE_{difference} = 112.26$). As predictors in this model, we included standardized estimates of Individualizing and Binding values as well as the proportion of people below the poverty line, the proportion of people with four-year degrees, and the county-level proportion of White inhabitants.

Comparisons of model predictions of the county-level rate of hate groups were largely consistent with the observed rates (See Figure 3), RMSE = 0.11. Consistent with our hypotheses, our results indicate a relationship between the county-level rate of hate groups and county-level Binding values (See Figure 4). Even after attempting to adjust for ethnic composition, educational attainment, and the proportion of county population below the poverty line, the rate of hate groups is expected to increase by

[4] For cities located in multiple counties, we selected the county containing the largest proportion of the cities population in order to avoid over counting hate groups.

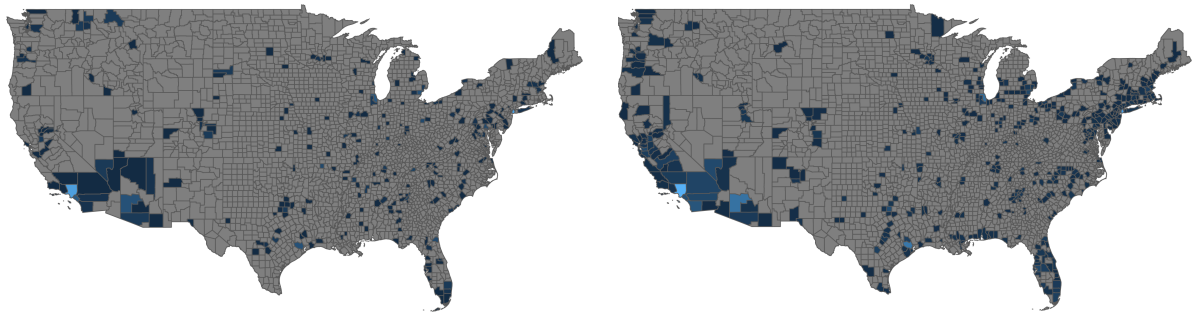

Figure 3. Observed (Left) and Predicted (Right) County-level Rate of Hate Groups

26% (posterior SD = 12%) with a standard deviation increase in Binding values. Notably, no such effect was observed for Individualizing values, which suggests that, after accounting for the effects of the other variables in the model, variations in Individualizing values are not associated with the prevalence of hate groups.

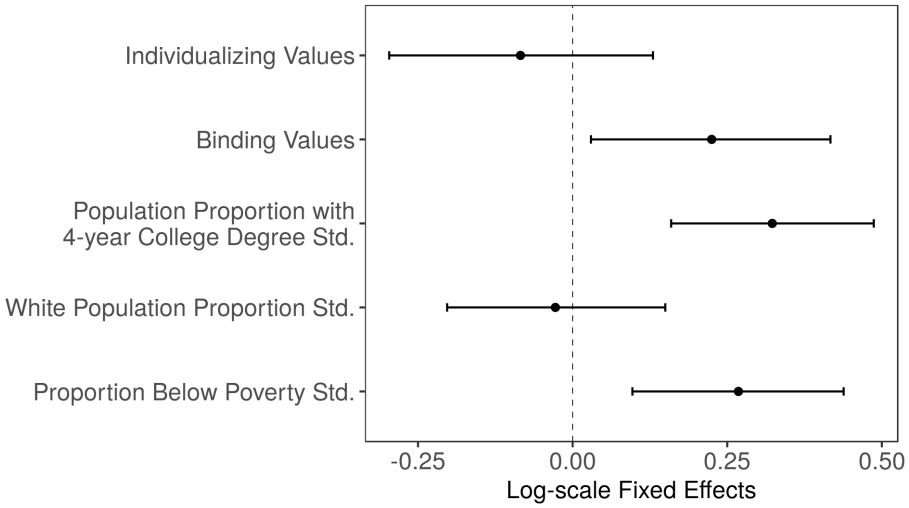

Figure 4. Estimated conditional association between Binding Values and the county-level prevalence of hate groups

Overall, our results suggest that, consistent with the Moralized Threat Hypothesis, the prevalence of hate groups in a particular region is linked to the Binding values held in that region. Thus, while there are, of course, many factors that likely drive participation in a hate group (Hall, 2013; Simi & Futrell, 2015), it appears that a region's moral milieu may play a systematic role in facilitating the emergence and

maintenance of hate groups. In our view, this association is likely multiply determined. It may be the case that Binding values function as risk factors at the individual level, such that people who prioritize Binding values are more likely, on average, to be susceptible to recruitment into hate groups. At the same time, the existence of hate groups may also be facilitated, for example, by a degree of tacit community acceptance (or an absence of community condemnation) that is linked to people's moral values.

Importantly, these results are theoretically consistent with our Study 1 finding that hate speech and the language of group-based morality overlap. Even though these studies relied on different measurement methodologies and operated on different levels of analysis, they both indicate that Binding values play an important role in real-world manifestations of EBEPs. However, neither of these studies directly investigate the association between people's moral values and EBEPs at the individual-level. Accordingly, we address this issue next through a series of survey-based studies that allow us to more precisely investigate the association between moral values and EBEPs.

Perceived Moral Violations and the Justification of EBEPs

In the previous studies, our results indicated an association between EBEPs and moral values in the context of real-world hate speech and the spatial distribution of hate groups. In the next series of studies, we extend these results by investigating the relationship between people's moral values and the degree to which they believe EBEPs are *justified*. While the self-reported justification of EBEPs is, of course, a different construct than actual EBEPs, this approach allows us to assess how public approval of EBEPs varies as a function of people's moral values. Further, self-reported justification of EBEPs may reflect tacit community acceptance. Using this approach, we also investigate the hypothesis that approval of EBEPs toward an outgroup is higher when people believe that the outgroup has done something immoral. Specifically, building on our findings in Studies 1 and 2, we use experimental and observational survey studies to test three primary hypotheses from the Moralized Threat Hypothesis:

- **Hypothesis 1.** An EBEP toward an outgroup is seen as more justified when that

outgroup is perceived as having done something morally wrong.

- **Hypothesis 2.** EBEPs should be seen as more justified by people who prioritize Binding values.

- **Hypothesis 3.** The association between Binding values and EBEPs toward a given outgroup is at least partially mediated by the perception that the outgroup has done something morally wrong.

Study 3: Experimental Manipulation of Perceived Moral Wrongness

First, we tested whether the perception that an outgroup has done something morally wrong predicts support for EBEPs against that outgroup (Hypothesis 1). To do this, we randomly assigned participants recruited via Amazon Mechanical Turk and paid $1.00 for their participation to one of two conditions — a 'high moral threat' condition and a 'low moral threat' condition — that manipulated the moral valence of a fictional outgroup. We focus on a fictional outgroup in order to limit the influence of participants' prior beliefs on their responses (Crandall & Schaller, 2005).

In both conditions, participants (N = 321; Mean Age = 33.92, SD = 10.88; 62% Female) were asked to read a fictional news story about 'Sandirian' (Crandall & Schaller, 2005) immigrants *taking* jobs in Webster Springs, Illinois, a fictional town. In the low moral threat condition, the Sandirians' actions were framed as stimulating the local economy. In contrast, in the high moral threat condition, the Sandirians were described as undermining the local economy and, thus, harming "native" citizens. We then asked participants to indicate on a 7-point scale (1 = 'Not at all morally wrong', 7 = 'Extremely morally wrong') the degree to which they believed it was morally wrong for the Sandirians to take jobs in Webster Springs ($M = 3.25$, $SD = 1.81$).

Finally, to assess participants' approval of EBEPs, we asked them to imagine a male Webster Springs resident, Dave, who believed that Sandirian immigrants were hurting his community. Participants then indicated how *justified* ('Not at all justified' = 1 to 'Extremely justified' = 7) Dave would be in committing four different EBEPs: posting hate speech to Facebook ($M = 2.80$, *Median* = 2, $SD = 1.88$), distributing hate

speech flyers in Webster Springs ($M = 2.73$, *Median* = 2, $SD = 1.86$), yelling slurs at a Sandirian resident of Webster Springs ($M = 1.90$, *Median* = 1, $SD = 1.47$), and physically assaulting a Sandirian resident of Webster Springs ($M = 1.36$, *Median* = 1, $SD = 1.04$). These exemplar EBEPs were selected to represent a variety of potential EBEPs that are characterized by different magnitudes of social norm violation.

Results. Twelve participants skipped the item measuring moral wrongness and an additional 15 participants spent less than 10 seconds reading the experimental manipulation, which was our a priori cutoff to ensure data quality. Accordingly, 294 participants were retained for analysis, though, robustness checks verified that retaining these participants had no substantive effect on our results. To test the hypothesis that perceived moral wrongdoing is associated with EBEP justification, we first estimated a Bayesian linear regression (Model 1) in which Z-scored perceived moral wrongdoing was regressed on experimental condition. Estimates from this model (See Table 3 for complete model estimates) show that participants in the high moral threat condition reported substantially higher levels of perceived moral wrongdoing, $b = 0.50$, *posterior* $SD = 0.11$, *95% CI* = [0.28, 0.73].

Next, we assessed the effect of experimental condition on EBEP justification. To do this, we estimated a second model (Model 2) in which EBEP justification was regressed on experimental condition. In this model, we treated responses to the EBEP items (Chronbach's $\alpha = 0.84$, 95% CI = [0.82, 0.86]) as a repeated measure of EBEP justification, yielding four measurements for each participant. To account for this, we used a Bayesian hierarchical modeling framework to allow for varying intercepts (i.e. random effects) for both participants and EBEP items. This approach enabled our model to address the facts that (1) different participants should be more or less likely to see EBEPs as justified in general and (2) that each EBEP item, on average, should been seen as more or less justified. Further, to account for the fact that the effects of experimental condition on EBEP justification may vary depending on the EBEP, we also allowed the effect of condition to vary across EBEP items (i.e. by estimating random slopes). Finally, because the distribution of responses to each EBEP item was

heavily skewed, we modeled EBEP justification using a cumulative logistic regression model (Bürkner & Vuorre, 2019). All together, this yielded a hierarchical Bayesian cumulative logistic regression model in which (1) EBEP justification was regressed on experimental condition and (2) varying intercepts for participant and EBEP item and a varying slope for both condition were estimated.

Results from Model 2 (See Table 3 for complete model estimates) indicate that, even after attempting to account for the random effects of participant (intercept SD = 3.79) and EBEP item (intercept SD = 4.04, b_{threat} SD = 0.62), participants in the high moral threat condition were substantially more likely to see EBEPs against Sandirians as more justified, compared to participants in the low moral threat condition, b = 1.44, *posterior SD* = 0.68, *CI 95%* = [0.16, 2.80], OR = 4.21, *CI 95%* = [1.18, 16.45]. In other words, for participants in the high moral threat condition, the odds of seeing EBEPs, on average, as extremely justified, versus less than extremely morally justified, was 4.21 times higher than for participants in the low moral threat condition.

Next, we directly investigated the role of perceived moral wrongdoing in participants' EBEP justification responses. To do this, we extended Model 2 by including standardized perceived moral wrongdoing (PMW) — the degree to which participants believed it was morally wrong for Sandirians to take jobs in Webster Springs — as an independent variable with varying slopes across EBEP items (Model 3). Estimates from this model indicated a strong positive association between believing it was morally wrong for Sandirians to take jobs and seeing EBEPs against Sandirians as more justified, b = 2.21, *posterior SD* = 0.42, *CI 95%* = [1.48, 2.94], OR = 9.20, *CI 95%* = [4.38, 18.92]. Notably, adjusting for PMW also led to a dramatic attenuation of the estimated effect of experimental condition, such that a clear positive effect was no longer supported, b = 0.28, *posterior SD* = 0.63, *CI 95%* = [-0.86, 1.57], OR = 1.29, *CI 95%* = [0.42, 4.79].

Finally, we tested the hypothesis that perceived moral wrongdoing mediated the effect of experimental condition on EBEP justification. Relying on Bayesian posterior simulation to estimate average mediation effects (AME) and average direct effects

(ADE) (Imai, Keele, & Tingley, 2010; Steen, Loeys, Moerkerke, & Vansteelandt, 2017; VanderWeele & Vansteelandt, 2014; VanderWeele, Zhang, & Lim, 2016), we found that perceived moral wrongdoing statistically mediates the effect of experimental condition on the probability of indicating that an EBEP was at least slightly justified (See Supplemental Material).

Consistent with Hypothesis 1, these results indicate that participants who were led to believe that a fictional immigrant group — the Sandirians — had done something immoral also believed that EBEPs against this group were more justified, compared to participants in the control condition. Importantly, adjusting for degree to which participants believed Sandirians had done something morally wrong also completely accounted for the effect of experimental condition. Consistent with this, mediation analyses also indicated that the effect of experimental manipulation was mediated by the degree to which participants believed that the Sandirians had done something immoral. Importantly, a secondary set of analyses in which participant political ideology was adjusted revealed no substantive changes in any of the reported findings (See in Supplemental Material for details). That is, these effects are hold even after adjusting for the degree to which participants identify as conservative.

Study 4: Justification of EBEPs against Mexicans

Next, we expand upon the experimental findings of Study 3 and investigate whether participants' Binding values are associated with the degree to which they see EBEPs as justified. In this study, rather than focusing on a fictional immigrant group, we directly address the perceived justification of EBEPs against Mexican immigrants. We focus on Mexican immigrants due to their cultural salience as a social group and consistent increases hate crimes targeting Mexicans in recent years (United States Department of Justice, Federal Bureau of Investigation, 2018). To this end, we asked participants ($N = 355$, Mean age = 33, 54% identifying as female) who were recruited from Amazon Mechanical Turk for $1.00 to read the same fictional news article shown in the high moral threat condition of the previous study. However, in this study, each

participant read a version of the article that replaced the Sandirians with Mexican immigrants. As in Study 3, participants were then asked to indicate their perceptions of moral wrongdoing via the same 6-point Likert scale. Prior to reading the news article, participants were also asked to complete the Moral Foundations Questionnaire (Graham et al., 2011), a 30 item scale designed to measure the degree to which people prioritize the five moral domains proposed by MFT. Responses to this scale were aggregated and standardized to construct Binding ($\alpha = 0.83$, 95% CI = [0.81, 0.86]) Individualizing ($\alpha = 0.78$, 95% CI = [0.75, 0.78]) scores for each participant. Finally, participants responded to the same four items measuring the to degree to which they thought EBEPs against Mexicans in Webster Springs were justified ($\alpha = 0.80$, 95% CI = [0.78, 0.83]).

Results

Three participants did not complete the MFQ and an additional 28 participants spent less than 10 seconds reading the experimental manipulation, spent less than eight minutes on the entire survey, or failed one of the MFQ manipulation checks, which were our a priori criteria to ensure data quality. Accordingly, 324 participants were retained for analysis, though, as in Study 3, robustness checks verified that retaining these participants had no substantive effect on our results.

To test Hypothesis 2, we modeled (Model 1) participants' responses to the EBEP items using a hierarchical Bayesian ordered logistic regression model. As in Study 3, we treated the EBEP responses as repeated measures and estimated varying intercepts for both participants and each EBEP item. In this model, we also estimated fixed and varying effects for participants' standardized Binding and Individualizing scores.

Consistent with Studies 1 and 2, model estimates indicated a strong association between participants' Binding values and the degree to which they believed EBEPs against Mexican immigrants in Webster Springs were justified (See Figure 5). Specifically, after attempting to account for the effect of Individualizing scores, the odds of selecting a higher, versus lower, response option were estimated to be 4.95 times higher given a standard deviation increase in Binding values, $b = 1.60$,

$posteriorSD = 0.27$, $95\%CI$ =[1.12, 2.14]. In contrast, this model indicated that Individualizing values were negatively associated with EBEPs, such that the odds of selecting a higher, versus lower, response option were estimated to be 0.31 times lower given a standard deviation increase in Individualizing values, $b = -1.15$, $posteriorSD = 0.41$, $95\%CI$ =[-1.88, -0.35].

Next, to test Hypotheses 1 and 3, we estimated two additional regression models. As in Study 3, we first modeled (Model 2) participants' standardized responses to the perceived moral wrongdoing item. However, in this model, we included participants' standardized Binding and Individualizing scores as predictors. We then estimated a third model (Model 3) that followed the same specification as Model 1 with the exception that fixed and random effects were also estimated for standardized perceived moral wrongdoing.

As expected, estimates from Model 2 indicated a positive association between participants' Binding values and the degree to which believed it was morally wrong for Mexican immigrants to "take jobs" in Webster Springs, $b = 0.47$, $posteriorSD = 0.05$, $95\%CI$ =[0.37, 0.56]. That is, a standard deviation increase in Binding values was associated with an estimated 0.47 standard deviation increase in perceived moral wrongness. In contrast, Individualizing values were estimated to be negatively associated with the perception of moral wrongdoing, $b = 0.47$, $posteriorSD = 0.05$, $95\%CI$ =[0.37, 0.56].

Further, as hypothesized, estimates from Model 3 indicated that even after attempting to adjust for the effects of standardized Individualizing and Binding values, standardized perceived moral wrongdoing was estimated to be positively associated with perceived EBEP justification, $b = 1.63$, $posteriorSD = 0.46$, $95\%CI$ =[0.70, 2.49]. Thus, the odds of seeing EBEPs as more justified than a given response level versus less justified or equal to that level are 5.10 times higher given a standard deviation increase in perceived moral wrongdoing. Notably, adjusting for the effect of perceived moral wrongdoing also substantially decreased the estimated effects of Binding ($b = 0.73$, $posteriorSD = 0.26$, $95\%CI$ =[0.24, 1.23]) and Individualizing ($b = -0.87$,

$posterior SD = 0.46$, 95%CI =[-1.65, 0.04]) values.

Finally, similar to Study 3, we relied on posterior simulation to test the hypothesis that perceived moral wrongdoing statistically mediates the association between Binding values and EBEP justification. Results from this analysis indicated that perceived moral wrongdoing partially mediated the association between Binding values and perceived EBEP justification (See Supplemental Material). Importantly, as in Study 3, we also repeated these analyses while attempting to adjust for participant political ideology (See in Supplemental Material for details). Again, we found that including political ideology in these analyses did not lead to substantive changes in any of our results.

Together, these results suggest that the degree to which people believe EBEPs are justified is positively associated with the degree to which they prioritize the Binding values. Further, our results are also consistent with the hypothesis that the association between the Binding values and EBEP justification is at least partially mediated by the perceived moral wrongness of outgroup behavior. However, as the current study focused on EBEPs against Mexican immigrants, it is not necessarily the case that these effects generalize to other real-world groups. To account for this, we conducted a final study with a stratified national sample investigating the perceived justification of EBEPs against Muslims motivated by the moral wrongness of Muslims spreading "Islamic culture".

Study 5: Justification of EBEPs against Muslims

Similar to Study 4, in this study we investigate the relationship between Binding values and the perceived justification of EBEPs against a real-world outgroup. However, here, we focus on a different outgroup, Muslims, which enables us to evaluate the degree to which the results from our previous studies generalize to other outgroups. Here, we focus specifically on the moral wrongness of Muslims' spreading Islamic values given the cultural salience of Muslims as a social outgroup, recent increases in hate crimes against Muslims (Pew Research Center, n.d.), and because focusing on the spread of Islamic values enabled us to evaluate the Moralized Threat Hypothesis under

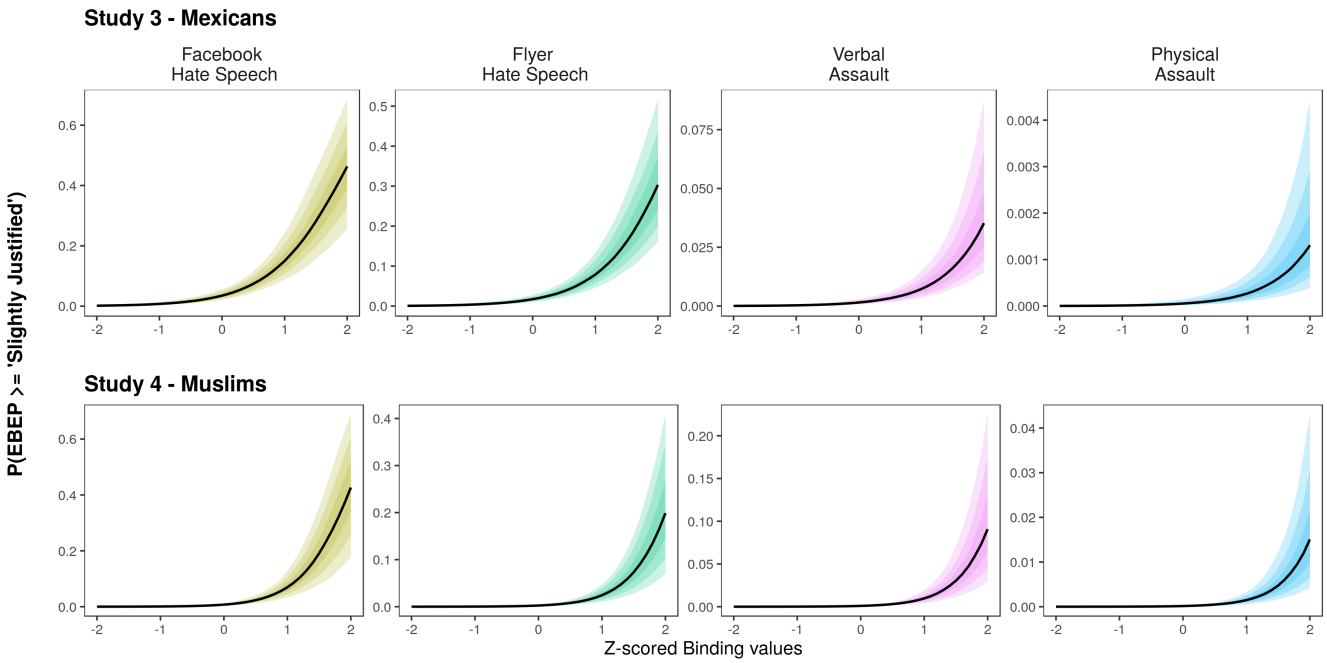

Figure 5. Estimated association between Binding values and perceived justification of EBEPs against Mexican immigrants (Study 4) and Muslims (Study 5). Participants who prioritized Binding values tended, on average, to see EBEPs against both groups as more justified.

conditions of symbolic threat.

To conduct this study, a sample of participants (N = 511) stratified by sex (51% female), age (10% to 20% per each 5-year bracket ranging from 18 to 65 or older), ethnicity (62% non-Hispanic White, 17% Hispanic, 13% Black, 7% other), and political affiliation (51% Democrat) was recruited by Qualtrics. After presenting a series of demographic questions, we measured participants' Individualizing (α = 0.80, 95% CI = [0.77, 0.82]) and Binding (α = 0.85, 95% CI = [0.83, 0.87]) values using the MFQ. We then measured perceived moral wrongness using a 6-point item that asked participants to indicate, "How morally wrong is it for Muslims to spread Islamic values or laws (e.g. Sharia law) in the US instead of assimilating into American culture?" To measure the perceived justification of EBEPs against Muslims, we asked participants to imagine a man named 'Dave' who "believes Muslims are hurting his community." As in the previous studies, we then asked participants to indicate how "justified" Dave would be in committing each of the exemplar EBEPs (α = 0.92, 95% CI = [0.90, 0.93]).

Results

To test our hypotheses, we used the same modeling procedure followed in Study 4. First, we modeled (Model 1) the association between participants' standardized Individualizing and Binding values and the degree to which they saw EBEPs against Muslims as justified. We then estimated two additional regression models, one (Model 2) in which the perceived moral wrongness of Muslims spreading Islamic values was regressed on participants' Individualizing and Binding values and another (Model 3) in which the perceived justification of EBEPs against Muslims was regressed on participants' Individualizing and Binding values as well as perceived moral wrongness. Finally, we used posterior simulations to estimate the degree to which perceived moral wrongness statistically mediates the effect of between Binding values on the perceived justification of EBEPs against Muslims.

Consistent with our previous studies, estimates from Model 1 indicated a strong association between participants' Binding values and the degree to which they believed EBEPs against Muslims were justified, $b = 2.29$, $posteriorSD = 0.35$, $95\%CI$ =[1.67, 2.96] (See Figure 5). As in Study 4, we also again observed a negative effect of Individualizing values, $b = -1.70$, $posteriorSD = 0.32$, $95\%CI$ =[-2.32, -1.10]. Further, estimates from Model 2 showed a positive association between perceived moral wrongdoing and EBEP justification, $b = 1.72$, $posteriorSD = 0.51$, $95\%CI$ =[0.75, 2.77]. Adjusting for perceived moral wrongdoing, in Model 2, also lead to a substantial reduction, relative to Model 1, in the magnitude of the effects of Individualizing values, $b = -1.29$, $posteriorSD = 0.34$, $95\%CI$ =[-1.89, -0.68], and Binding values, $b = 1.48$, $posteriorSD = 0.35$, $95\%CI$ =[0.82, 2.18]. Finally, our mediation analysis indicated that perceived moral wrongness statistically mediates the association between Binding values and EBEP justification (See Supplemental Material)

Importantly, these results completely replicated the patterns of effects observed in Study 4. We observed a strong positive association between participants' Binding values and the degree to which they believed EBEPs against Muslims were justified. We also found that the degree to which participants thought it was morally wrong for Muslims

to "spread Islamic values" was not positively associated with EBEP justification, but it also partially mediated the association between Binding values and EBEP justification. Finally, as in the previous studies, adjusting for participant political ideology did not lead to substantive changes in any of our results (See in Supplemental Material for details).

Combined with Studies 3 and 4, these results suggest that, at least in our current social context, people who prioritize group-oriented moral values are more likely to see EBEPs against an outgroup as justified. Further, it appears that this association may at least partly depend on the perception that the outgroup has done something morally wrong.

Conclusions

Taken together, our analyses of the entire Gab corpus (approximately 24 million posts), 3108 U.S. counties, and experimental and observational survey data collected from over 1,200 participants converge on the finding that extreme behavioral expressions of prejudice — behaviors like hate speech, hate group activity, and hate crime — are tied to people's group-oriented moral values. Specifically, we found that hate speech tends to be articulated through the language of Binding values; that hate groups are more prevalent in regions that prioritize Binding values, and that individuals who place a greater emphasis on Binding values are more likely to believe that hate speech, harassment, and even assault against outgroup members are more justified. Crucially, it also appears that this association between Binding values and EBEP justification can be at least partly explained by the belief that an outgroup has done something morally wrong.

While it may seem counter-intuitive to think of a class of behaviors — many of which that have received their own special legal designation as *particularly* heinous crimes (Hall, 2013) — as moral phenomena, this view is well-grounded in current understandings of the relationship between morality and acts of extremism or violence (Atran & Ginges, 2012, 2015; Darley, 2009; Dehghani et al., 2010; Mooijman et al.,

2018; Skitka et al., 2017; Zaal et al., 2011). Just as suicide attacks (Ginges & Atran, 2009; Ginges, Hansen, & Norenzayan, 2009) and other acts of violence (Fiske et al., 2014; Rai & Fiske, 2012) can often be better understood as moralized or sacred conflicts, our findings suggest that acts of hate may often be morally motivated. Notably, in this work, we focused on acts of hate perpetrated against outgroups that are more often demonized by people who subscribe to conservative or right-wing ideologies. This raises important questions about the motivations underlying acts of hate perpetrated by people who subscribe to liberal or left-wing ideologies. The Moralized Threat Hypothesis predicts that these acts, too, are motivated by perceived moral violations. However, the specific moral values that these violations are grounded in may differ from those that ground the moral violations perceived by conservatives. For instance, people who subscribe to more ideologically liberal belief systems may be more sensitive to violations of principals of Care or Fairness than they are to violations of the Binding values. At the same time, it may also be the case that, even among liberals, stronger subscription to group-oriented moral values is associated with EBEPs toward outgroups.

Given today's digital media environment and its potential for stoking moral outrage (Crockett, 2017) and uniting isolated individuals who share fringe ideologies, understanding these effects is particularly important. While much research on EBEPs has highlighted the role of specific, concrete threats (Green & Spry, 2014; Piatkowska et al., 2018), the Moralized Threat Hypothesis offers a framework for understanding when, where, and why people may engage in EBEPs even in the absence of an ostensible material threat. This hypothesis suggests that a person does not necessarily need to fear for their job or safety to engage in or approve of EBEPs; instead, it may be sufficient for them to simply feel a sense of moral outrage. Importantly, however, understanding EBEPs as morally motivated responses to perceived violations does not justify or excuse them; rather, it provides a psychological framework (Atran & Ginges, 2012; Ginges & Atran, 2009; Ginges et al., 2009; Skitka et al., 2017) for understanding why people engage in such extreme behaviors.

References

Anthony, A. (2016). Inside the hate-filled echo chamber of racism and conspiracy theories. *theguardian.com*. Retrieved from https://www.theguardian.com/media/2016/dec/18/gab-the-social-network-for-the-alt-right

Asbrock, F., Sibley, C. G., & Duckitt, J. (2010). Right-wing authoritarianism and social dominance orientation and the dimensions of generalized prejudice: A longitudinal test. *European Journal of Personality: Published for the European Association of Personality Psychology*, *24*(4), 324–340.

Atran, S., & Ginges, J. (2012, May). Religious and sacred imperatives in human conflict. *Science*, *336*(6083), 855–857.

Atran, S., & Ginges, J. (2015). Devoted actors and the moral foundations of intractable inter-group conflict. *The moral brain*, 69–85.

Beckett, L. (2019, March). Facebook to ban white nationalism and separatism content. *The Guardian*.

Benson, T. (2016). Inside the "twitter for racists": Gab — the site where milo yiannopoulos goes to troll now. *Salon.com*. Retrieved from https://www.salon.com/2016/11/05/inside-the-twitter-for-racists-gab-the-site-where-milo-yiannopoulos-goes-to-troll-now/

Blum, R., Stanton, G. H., Sagi, S., & Richter, E. D. (2008, April). 'ethnic cleansing' bleaches the atrocities of genocide. *European journal of public health*, *18*(2), 204–209.

Bürkner, P.-C., & Vuorre, M. (2019, February). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918823199.

Byers, B. D., & Jones, J. A. (2007, February). The impact of the terrorist attacks of 9/11 on Anti-Islamic hate crime. *Journal of Ethnicity in Criminal Justice*, *5*(1), 43–56.

Center for the Study of Hate & Extremsism. (2018). *Hate crimes rise in U.S. cities and counties in time of division & foreign interference.*

Charles-Toussaint, G. C., & Crowson, H. M. (2010, September). Prejudice against international students: the role of threat perceptions and authoritarian dispositions in U.S. students. *The Journal of psychology*, *144*(5), 413–428.

Cogan, J. C. (2002, September). Hate crime as a crime category worthy of policy attention. *The American behavioral scientist*, *46*(1), 173–185.

Cohrs, J. C., & Ibler, S. (2009, February). Authoritarianism, threat, and prejudice: An analysis of mediation and moderation. *Basic and applied social psychology*, *31*(1), 81–94.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Conger, K. (2019, May). Facebook says it is more aggressively enforcing content rules. *The New York Times*.

Crandall, C. S., & Schaller, M. (2005). *Social psychology of prejudice: Historical and contemporary issues*. Lewinian Press.

Crockett, M. J. (2017, November). Moral outrage in the digital age. *Nature human behaviour*, *1*(11), 769–771.

Darley, J. M. (2009). Morality in the law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science*, *5*, 1–23.

Dehghani, M., Atran, S., Iliev, R., Sachdeva, S., Medin, D., & Ginges, J. (2010). Sacred values and conflict over iran's nuclear program. *Judgment and Decision Making*, *5*(7), 540.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first italian conference on cybersecurity, venice, italy* (pp. 86–95).

Duckitt, J., & Sibley, C. G. (2009, August). A Dual-Process motivational model of ideology, politics, and prejudice. *Psychological inquiry*, *20*(2-3), 98–109.

Duckitt, J., & Sibley, C. G. (2017). The dual process motivational model of ideology

and prejudice. *The Cambridge handbook of the psychology of prejudice*, 188–221.

Eligon, J. (2018, November). Hate crimes increase for the third consecutive year, F.B.I. reports. *The New York Times*.

Engel, V., Camus, J.-Y., Feldman, M., Allchorn, W., Castriota, A., Barna, I., ... Semenov, M. (2018, December). *Xenophobia, radicalism, and hate crime in europe annual report* (Tech. Rep.).

Fiske, A. P., Rai, T. S., & Pinker, S. (2014). *Virtuous violence: Hurting and killing to create, sustain, end, and honor social relationships*. Cambridge University Press.

Frenkel, S., Isaac, M., & Conger, K. (2018, October). On instagram, 11,696 examples of how hate thrives on social media. *The New York Times*.

Frimer, J. A., Biesanz, J. C., Walker, L. J., & MacKinlay, C. W. (2013, June). Liberals and conservatives rely on common moral foundations when making moral judgments about influential people. *Journal of personality and social psychology*, *104*(6), 1040–1059.

Gaffney, G. (2018). *Pushshift gab corpus.* `https://files.pushshift.io/gab/`. (Accessed: 2019-5-23)

Ginges, J., & Atran, S. (2009). What motivates participation in violent political action. *Annals of the New York Academy of Sciences*, *1167*(1), 115–123.

Ginges, J., Hansen, I., & Norenzayan, A. (2009, February). Religion and support for suicide attacks. *Psychological science*, *20*(2), 224–230.

Government of Canada, Department of Justice, Research, & Statistics Division. (2011, June). *Understanding the community impact of hate crimes: A case study - victims of crime research digest, issue no. 4.* `https://www.justice.gc.ca/eng/rp-pr/cj-jp/victim/rd4-rr4/p4.html`. (Accessed: 2019-6-5)

Graham, J., & Haidt, J. (2011). Sacred values and evil adversaries: A moral foundations approach. *The social psychology of morality: Exploring the causes of good and evil.*, 1–18.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H.

(2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Elsevier.

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.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Green, D. P., & Spry, A. D. (2014, August). Hate crime research: Design and measurement strategies for improving causal inference. *Journal of contemporary criminal justice*, *30*(3), 228–246.

Hall, N. (2013). *Hate crime*. Routledge.

Hanes, E., & Machin, S. (2014, August). Hate crime in the wake of terror attacks: Evidence from 7/7 and 9/11. *Journal of contemporary criminal justice*, *30*(3), 247–267.

Hanretty, C., Lauderdale, B. E., & Vivyan, N. (2016). Comparing strategies for estimating constituency opinion from national survey samples. *Political Science Research and Methods*, 1–21.

Held, L., Schrödle, B., & Rue, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib & G. Tutz (Eds.), *Statistical modelling and regression structures: Festschrift in honour of ludwig fahrmeir* (pp. 91–110). Heidelberg: Physica-Verlag HD.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoover, J., & Dehghani, M. (2018). The big, the bad, and the ugly: Geographic estimation with flawed psychological data. *PsyArXiv. October*.

Hoover, J., Johnson-Grey, K., Dehghani, M., & Graham, J. (2017). *Moral values coding guide.*

Imai, K., Keele, L., & Tingley, D. (2010, December). A general approach to causal

mediation analysis. *Psychological methods*, *15*(4), 309–334.

Kennedy, B., Kogon, D., Coombs, K., Hoover, Park, Portillo-Wightman, G., ... Dehghani, M. (2018). *A typology and coding manual for the study of hate-based rhetoric.*

Kiernan, B. (2007). *Blood and soil: A world history of genocide and extermination from sparta to darfur*. Yale University Press.

Leemann, L., & Wasserfallen, F. (2017, October). Extending the use and prediction precision of subnational public opinion estimation: EXTENDING USE AND PRECISION OF MrP. *American journal of political science*, *61*(4), 1003–1022.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials* (pp. 179–191). Springer New York.

Levy, B. L., & Levy, D. L. (2017, January). When love meets hate: The relationship between state policies on gay and lesbian rights and hate crime incidence. *Social science research*, *61*, 142–159.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of statistical software*, *63*(19).

Lowery, W., Kindy, K., & Tran, A. B. (2018, November). In the united states, right-wing violence is on the rise. *The Washington Post*.

McCann, S. J. H. (2010, January). Authoritarianism, conservatism, racial diversity threat, and the state distribution of hate groups. *The Journal of psychology*, *144*(1), 37–60.

McVeigh, R. (2004, March). Structured ignorance and organized racism in the united states. *Social forces; a scientific medium of social study and interpretation*, *82*(3), 895–936.

McVeigh, R., & Sikkink, D. (2005, December). Organized racism and the stranger. *Sociological Forum*, *20*(4), 497–522.

Medina, R. M., Nicolosi, E., Brewer, S., & Linke, A. M. (2018, July). Geographies of organized hate in america: A regional analysis. *Annals of the Association of*

American Geographers. Association of American Geographers, *108*(4), 1006–1021.

Menke, J., & Martinez, T. R. (2004). Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons. In *Neural networks, 2004. proceedings. 2004 ieee international joint conference on* (Vol. 2, pp. 1331–1335).

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*, 389–396.

Moore, B. (2000). *Moral purity and persecution in history*. Princeton University Press.

Nirenberg, D. (2015). *Communities of violence: Persecution of minorities in the middle ages - updated edition*. Princeton University Press.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-Level estimates from national polls. *Political analysis: an annual publication of the Methodology Section of the American Political Science Association*, *12*(4), 375–385.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, *12*, 1532–1543.

Perry, B., & Sutton, M. (2006, October). Seeing red over black and white: Popular and media representations of inter-racial relationships as precursors to racial violence. *Canadian Journal of Criminology and Criminal Justice*, *48*(6), 887–904.

Perry, B., & Sutton, M. (2008). Policing the colour line violence against those in intimate interracial relationships. *Race, Gender & Class*, 240–261.

Pew Research Center. (n.d.). *Assaults against muslims in U.S. surpass 2001 level.* `https://www.pewresearch.org/fact-tank/2017/11/15/assaults-against-muslims-in-u-s-surpass-2001-level/`. (Accessed: 2019-6-5)

Piatkowska, S. J., Messner, S. F., & Yang, T.-C. (2018, November). Xenophobic and racially motivated crime in belgium: exploratory spatial data analysis and spatial

regressions of structural covariates. *Deviant behavior*, *39*(11), 1398–1418.

Rai, T. S. (2019, October). Higher self-control predicts engagement in undesirable moralistic aggression. *Personality and individual differences*, *149*, 152–156.

Rai, T. S., & Fiske, A. P. (2012). Beyond harm, intention, and dyads: Relationship regulation, virtuous violence, and metarelational morality. *Psychological inquiry*.

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016, August). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145–1165.

Roose, K. (2018, October). On gab, an Extremist-Friendly site, pittsburgh shooting suspect aired his hatred in full. *The New York Times*.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Selb, P., & Munzert, S. (2011). Estimating constituency preferences from sparse survey data using auxiliary geographic information. *Political analysis: an annual publication of the Methodology Section of the American Political Science Association*.

Simi, P., & Futrell, R. (2015). *American swastika: Inside the white power movement's hidden spaces of hate*. Rowman & Littlefield.

Skitka, L. J., Hanson, B. E., & Wisneski, D. C. (2017, February). Utopian hopes or dystopian fears? exploring the motivational underpinnings of moralized political engagement. *Personality & social psychology bulletin*, *43*(2), 177–190.

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management* (pp. 623–632). New York, NY, USA: ACM.

Southern Law Poverty Center. (2019). *Splc hate map.* `https://www.splcenter.org/hate-map`. (Accessed: 2019-6-24)

SPLC. (2019). *Hate groups reach record high.* `https://www.splcenter.org/news/2019/02/19/hate-groups-reach-record-high`. (Accessed: 2019-6-5)

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stacey, M., Carbone-López, K., & Rosenfeld, R. (2011a, August). Demographic change and ethnically motivated crime: The impact of immigration on Anti-Hispanic hate crime in the united states. *Journal of contemporary criminal justice*, *27*(3), 278–298.

Stacey, M., Carbone-López, K., & Rosenfeld, R. (2011b, August). Demographic change and ethnically motivated crime: The impact of immigration on Anti-Hispanic hate crime in the united states. *Journal of contemporary criminal justice*, *27*(3), 278–298.

Steen, J., Loeys, T., Moerkerke, B., & Vansteelandt, S. (2017, July). Flexible mediation analysis with multiple mediators. *American journal of epidemiology*, *186*(2), 184–193.

Stephan, W. G., & Stephan, C. W. (1996, June). Predicting prejudice. *International journal of intercultural relations: IJIR*, *20*(3), 409–426.

Stephan, W. G., & Stephan, C. W. (2017, November). Intergroup threat theory. In Y. Y. Kim (Ed.), *The international encyclopedia of intercultural communication* (Vol. 39, pp. 1–12). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008, December). Online racial discrimination and psychological adjustment among adolescents. *The Journal of adolescent health: official publication of the Society for Adolescent Medicine*, *43*(6), 565–569.

United States Department of Justice, Federal Bureau of Investigation. (2018, November). *Hate crime statistics, 2017.* (`https://ucr.fbi.gov/hate-crime/2017/topic-pages/jurisdiction`)

Valencia, Z., Williams, B., & Pettis, R. (2019, March). *Pride and prejudice: Same-Sex marriage legalization announcements and LGBT Hate-Crimes.*

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual review*

of public health, *37*, 17–32.

VanderWeele, T. J., & Vansteelandt, S. (2014, January). Mediation analysis with multiple mediators. *Epidemiologic methods*, *2*(1), 95–115.

VanderWeele, T. J., Zhang, Y., & Lim, P. (2016, September). Brief report: Mediation analysis with an ordinal outcome. *Epidemiology*, *27*(5), 651–655.

Vehtari, A., Gelman, A., & Gabry, J. (2017, September). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413–1432.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & others. (2016). Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine*.

Velasco González, K., Verkuyten, M., Weesie, J., & Poppe, E. (2008, December). Prejudice towards muslims in the netherlands: testing integrated threat theory. *The British journal of social psychology / the British Psychological Society*, *47*(Pt 4), 667–685.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing*. Academic press, New York.

Zaal, M. P., Van Laar, C., Ståhl, T., Ellemers, N., & Derks, B. (2011, December). By any means necessary: the effects of regulatory focus and moral conviction on hostile and benevolent forms of collective action. *The British journal of social psychology / the British Psychological Society*, *50*(4), 670–689.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Sirivianos, M., Stringhini, G., Kwak, H., & Blackburn, J. (2018). What is gab? a bastion of free speech or an alt-right echo chamber? *arXiv preprint arXiv:1802.05287*.

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference* (pp. 745–760).

Supplemental Material

Study 1

Analysis A

Results. In Study 1, our primary analysis was conducted with the sub sample of messages posted by Gab users wither fewer than 500 posts. Specifically, we used a hierarchical logistic regression model to estimate the probability of a message being labeled as hate speech conditional on whether it was also labeled as evoking the Individualizing vices or Binding vices. In this model, the intercept and the effects of the moral vices were permitted to vary across users. Below, we report results from the same model that was estimated using the entire Gab corpus (N_posts = 24,978,951; N_users; = 236,823).

Table 2

Effect of the presence of Individualizing and Binding vices on the log-probability of Message-level hate speech

Fixed Effects		
	Intercept	-5.926^{***}
		(0.012)
	Individualizing Rhetoric	1.713^{***}
		(0.009)
	Binding Rhetoric	3.125^{***}
		(0.007)
Random Effects		
	Intercept_{User}	1.49
	$\text{Individualizing Rhetoric}_{User}$	0.66
	$\text{Binding Rhetoric}_{User}$	0.65

Note: Estimates reported on log-scale. $^{***}p < 0.01$

Analysis B

To supplement the primary analysis for Study 1, we also evaluated the association between hate speech and the Binding and Individualizing foundations using an alternative analytic strategy. Specifically, rather than using independent Long Short-term Memory (LSTM) neural network models models (Hochreiter & Schmidhuber, 1997) to detect hate speech and moral rhetoric, we trained two additional LSTMs that included indicators for either the Binding or Individualizing vices as additional predictive features. This allowed us to directly address the questions of whether (1) accounting for moral rhetoric improves hate speech detection and (2) whether accounting for rhetoric evoking the Binding vices improves performance more than accounting for the Individualizing vices.

To train these models, we relied on the same data used for our first analysis. Specifically, as in our first analysis, the first model that we developed (See left panel of Figure 6) was a standard LSTM, trained only to predict whether or not a given post was labeled as hate speech. In contrast, the second and third models were designed to incorporate labels indicating the presence of Binding or Individualizing vices. These labels were represented as a vector of features and concatenated to the LSTM's output vector to predict the hate label.

Because this architecture directly incorporates contextual information about the moral content of a message, it allowed us to test the hypothesis that the semantic spaces of moral concerns and hate overlap. If our hypothesis is wrong (i.e. if hate speech does not rely on the language of the moral vices), the feature-based models should perform worse or no better than the vanilla LSTM model because adding irrelevant features should add noise to the information extracted from a sentence. On the other hand, if our hypothesis is correct, the feature-based models should perform better than the basic model and this would suggest an intersection between hate speech and moral language.

Similar to our other models, the vanilla and feature-based models both represent posts as matrices of pretrained GloVe word embeddings (Pennington et al., 2014) corresponding to the words in the original post. This embedding matrix is then input to

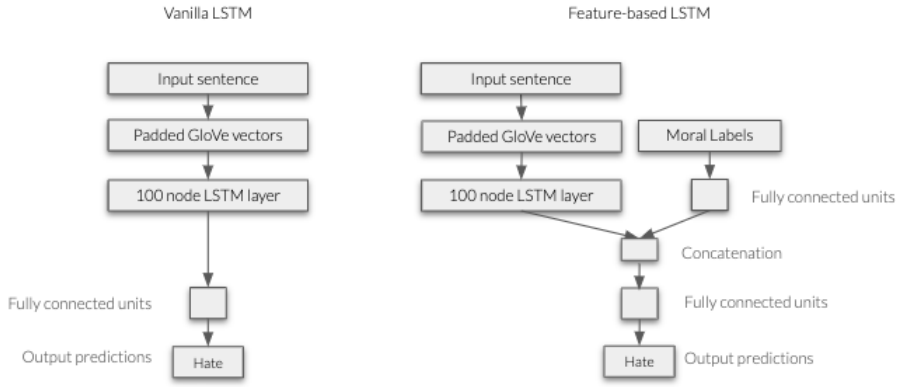

Figure 6. Vanilla LSTM Model Structure (left) and feeature-based LSTM (right). The feature-based LSTM exploits the input features along with the LSTM output to predict the label.

a 100 dimensional LSTM layer which is connected to a layer of fully connected units, with 0.33 dropout ratio. A softmax transformation is then applied to the output of the final layer in order to generate probabilistic predictions for the outcome. Further, in the feature-base model, the feature set is transformed to vector and concatenated to the output of the LSTM before generating the predictions. Specifically,the feature-based LSTMs were trained with Binding and Individualizing moral vices independently. The accuracy of the models were then calculated by training and testing the models in 10 iterations of 10-fold cross validations.

The comparative performance of feature-based and vanilla LSTM models was consistent with our hypotheses. Specifically, the feature-based model trained on the *Binding* vices ($M_{F1} = 0.671, SD_{F1} = 0.032$) substantially outperformed the vanilla LSTM model, which was trained to predict only hate ($M_{F1} = 0.64, SD_{F1} = 0.031$), Yuen's $t(117.9) = 6.47$, $SE = 0.005$, 95% CI = [-0.04 -0.02], $\xi = 0.64$, 95% CI [0.44, 0.77] [5]. In contrast, the model trained with the Individualizing vices

[5] Here, we report Yuen's t because F1 scores are not normally distributed (Menke & Martinez, 2004; Smucker, Allan, & Carterette, 2007) and Yuen's t is robust to deviations from normality (Wilcox, 2017). Effect sizes for Yuen's t, ξ, are also reported and can be interpreted as similar to Cohen's d, such that xi of 0.15, 0.35, 0.50 is comparable to $d = 0.2, 0.5, 0.8$.

($M_{F1} = 0.645, SD_{F1} = 0.034$) did not substantively improve performance, Yuen's $t(117.4) = 1.90$, $SE = 0.005$, 95% CI = [-0.02, 0.0004], $\xi = 0.20$, 95% CI = [0, 0.4].

Consistent with the results from our primary analysis, these results indicate that hate speech is often articulated via language that evokes the Binding concerns. Specifically, by incorporating information about the presence or absence of language evoking the Binding vices into the representations of Gab posts, we were able to substantively improve our models' capacity for detecting hate speech.

Study 3

Regression Models

Here we report model estimates for each of the three regression models estimated for Study 3 (See main text for discussion). Specifically, we report mean posterior estimates with Highest Posterior Density Intervals for the fixed and random effects estimated via these models (See Table 3). We also report estimates from models that adjust for individual-level political ideology (M = 2.67, SD = 1.78; See Table 4). Note, for these models, N = 278 as 10 participants indicated that they either did not know or have a political ideology.

Mediation Analysis

To investigate whether perceived moral wrongdoing statistically mediated the effect of condition, we relied on Bayesian posterior simulation to estimate average mediation effects and average direct effects (Imai et al., 2010; Steen et al., 2017; VanderWeele & Vansteelandt, 2014; VanderWeele et al., 2016). We used this framework because standard approaches to mediation analysis are not appropriate for ordered logistic regression outcome models (VanderWeele et al., 2016) due to their fixed error variance (Imai et al., 2010) and because it offers a holistic approach to mediation estimation for hierarchical models (VanderWeele, 2016).

Method. Under this approach, the average mediation effect (AME; e.g. indirect effect) for a mediator M is estimated via posterior simulations from two Bayesian

Table 3

Study 3 Model Estimates.

Fixed Effects		Model 1 PMW Std.	Model 2 EBEP J	Model 3 EBEP J
	Intercept	−0.24*		
		[−0.40; −0.09]		
	High Moral Threat	0.50*	1.44*	0.28
		[0.27; 0.72]	[0.15; 2.78]	[−0.87; 1.55]
	PMW Std.			2.21*
				[1.48; 2.94]
	Intercept[1]		0.71	0.17
			[−3.55; 4.43]	[−3.63; 3.76]
	Intercept[2]		2.65	2.10
			[−1.46; 6.56]	[−1.64; 5.74]
	Intercept[3]		3.65	3.10
			[−0.66; 7.34]	[−0.76; 6.67]
	Intercept[4]		5.19*	4.68*
			[0.89; 8.93]	[0.74; 8.19]
	Intercept[5]		6.70*	6.25*
			[2.28; 10.34]	[2.23; 9.69]
	Intercept[6]		9.14*	8.83*
			[4.74; 12.87]	[4.98; 12.48]
SD of Random Effects				
	$\text{Intercept}_{\text{Subject}}$		3.79*	2.99*
			[3.27; 4.37]	[2.55; 3.45]
	$\text{Intercept}_{\text{EBEP}}$		4.04*	3.78*
			[1.36; 8.12]	[1.28; 7.48]
	$\text{High Moral Threat}_{\text{EBEP}}$		0.62*	0.71*
			[0.00; 1.76]	[0.00; 2.06]
	$\text{PMW Std}_{\text{EBEP}}$			0.47*
				[0.00; 1.29]

* 0 outside 95% highest posterior density interval. Estimates for Models 2 and 3 reported on log scale.

Table 4

Study 3 Model Estimates Adjusted for Ideology

Fixed Effects		Model 1	Model 2	Model 3
		PMW Std.	EBEP J	EBEP J
	Intercept	−0.23*		
		[−0.38; −0.08]		
	High Moral Threat	0.47*	1.48*	0.51
		[0.25; 0.68]	[0.27; 2.68]	[−0.72; 1.70]
	Ideology Std.	0.32*	1.88*	1.22*
		[0.21; 0.43]	[1.15; 2.64]	[0.38; 1.96]
	PMW Std.			1.83*
				[1.03; 2.63]
	Intercept[1]		0.67	0.24
			[−3.68; 4.32]	[−3.55; 3.67]
	Intercept[2]		2.65	2.21
			[−1.68; 6.31]	[−1.55; 5.65]
	Intercept[3]		3.65	3.22
			[−0.65; 7.34]	[−0.56; 6.65]
	Intercept[4]		5.26*	4.86*
			[1.00; 9.03]	[1.15; 8.39]
	Intercept[5]		6.80*	6.46*
			[2.51; 10.57]	[2.62; 9.94]
	Intercept[6]		9.23*	9.00*
			[4.83; 12.97]	[5.13; 12.56]
SD of Random Effects				
	$\text{Intercept}_{Subject}$		3.40*	2.85*
			[2.89; 3.93]	[2.41; 3.29]
	Intercept_{EBEP}		4.12*	3.85*
			[1.33; 8.06]	[1.21; 7.52]
	$\text{High Moral Threat}_{EBEP}$		0.60*	0.66*
			[0.00; 1.73]	[0.00; 1.97]
	$\text{Ideology Std}_{EBEP}$		0.46*	0.51*
			[0.00; 1.26]	[0.00; 1.41]
	PMW Std_{EBEP}			0.50*
				[0.00; 1.40]

* 0 outside 95% highest posterior density interval. Estimates for Models 2 and 3 reported on log scale. Ideology is coded such that higher values indicate stronger associations with Conservative ideology.

regression models. In the first of these models, M is regressed on the independent or treatment variable T. In the second of these models, the endogenous dependent variable Y is regressed on both M and T.

Next, the model posteriors are used to generate a set of counterfactual predictions that are used to estimate the AME and average direct effect (ADE). Specifically, the first model is used to simulate N predicted values of $M|T = T$. For example, this might involve simulating two sets of values for M by drawing 500 values for $M|T = control$ and 500 values for $M|T = experimental$ from the model posterior. Per convention, we represent these sets of simulated values as $M_{T=0}$ and $M_{T=1}$, respectively, where $M_{T=0}$ represents the conditional posterior distribution of M when treatment equals zero.

Then, $M_{T=0}$ and $M_{T=1}$, the simulated values generated in the previous step, are used to simulate the expected values of $Y|T = t, M = m$. Specifically, three sets of values are simulated for Y: $Y|T = 0, M = M_{T=1}$, $Y|T = 1, M = M_{T=0}$, $Y|T = 0, M = M_{T=0}$. These simulation sets approximate the conditional distribution of Y given treatment equals t and the M equals a plausible value under a given treatment condition. For example, the first set, which we represent as $Y_{T=0,M_{T=1}}$, approximates the expected distribution of Y where $T = 0$ but M is set as if $T = 1$. That is, $Y_{T=0,M_{T=1}}$ estimates the posterior distribution of Y for values of M expected under the treatment condition while setting T to zero. Similarly, the second and third sets, $Y_{T=1,M_{T=0}}$ and $Y_{T=0,M_{T=0}}$, approximate the posterior distributions for Y when $T = 1$ but M is set as if $T = 0$ and for Y when $T = 0$ and M is set as if $T = 0$. Thus, in contrast to $Y_{T=0,M_{T=1}}$, $Y_{T=1,M_{T=0}}$ estimates the posterior distribution of Y under the treatment condition while effectively blocking the effect of M and $Y_{T=0,M_{T=0}}$ estimates the posterior distribution of Y under the control condition.

Finally, the simulation sets $Y_{T=0,M_{T=1}}$, $Y_{T=1,M_{T=0}}$, and $Y_{T=0,M_{T=0}}$ are used to estimate the AME and ADE. Specifically, the posterior distribution of the AME is calculated as $Y_{T=0,M_{T=1}} - Y_{T=0,M_{T=0}}$ and the ADE is calculated as $Y_{T=1,M_{T=0}} - Y_{T=0,M_{T=0}}$. Accordingly, the AME estimates the counterfactual change in Y that is expected given that the treatment is held constant at control but the mediator is

changed as if the treatment was administered. Similarly, the ADE estimates the counterfactual change in Y that is expected given that the treatment is administered, but the mediator is restricted to values consistent with the control condition.

Because this approach to mediation estimation relies on model predictions, concerns about coefficient comparisons across models are irrelevant. Importantly, even for generalized linear models, it also enables direct estimates of mediation effects on the scale of the dependent variable. Thus, combining this approach with ordered logistic regression allows us to estimate AMEs and ADEs for the probabilities of selecting each response level of the dependent variable. Finally, this approach also facilitates estimating AMEs and ADEs while adjusting for covariates. To do this, covariates are included in the regression models and then conditioned on at specific levels during the simulation step.

Using this approach, we estimated AMEs and ADEs for perceived moral wrongdoing (M), experimental condition (T) and EBEP justification (Y) using the two-step process outlined above. Further, we estimated these effects both without and with adjustments for political ideology.

First, for both the low and high moral threat conditions, we used Model 1 to simulate two sets of 500 perceived moral wrongdoing scores, which we represent as $M_{T=0}$ and $M_{T=1}$, respectively. Then, for each EBEP item j, we used Model 3 to simulate three sets, each consisting of 500 draws, of EBEP justification scores conditional on experimental condition and the simulated moral wrongness scores, $Y^j_{T=0,M_{T=1}}$, $Y^j_{T=1,M_{T=0}}$, and $Y^j_{T=0,M_{T=0}}$. For all posterior draws from Model 3, we conditioned on the random effects of EBEP item and marginalized over the random effects of subject. Thus, for each of the four EBEP items, this process yielded posterior approximations of $Y^j_{T=0,M_{T=1}}$, $Y^j_{T=1,M_{T=0}}$, and $Y^j_{T=0,M_{T=0}}$, such that each simulation set consisted of 500 (simulated moral wrongness scores) $\times$500 (posterior draws) = $250,000$ values.

Finally, we calculated AMEs and ADEs for each $j \in j = 1, ..., 4$ EBEP items as well as marginal AME and ADE across all EBEP items as

$$AME^j = \frac{1}{N_j}(\sum_{i=1}^{i=N_j} Y_{ij,T=0,M_{T=1}} - Y_{ij,T=0,M_{T=0}})$$

$$ADE^j = \frac{1}{N_j}(\sum_{i=1}^{i=N_j} Y_{ij,T=1,M_{T=0}} - Y_{ij,T=0,M_{T=0}})$$

$$AME = \frac{1}{\sum_{j=1}^{j=4} N_j}(\sum_{j=1}^{j=4}\sum_{i=1}^{i=N} Y_{ij,T=0,M_{T=1}} - Y_{ij,T=0,M_{T=0}})$$

$$ADE = \frac{1}{\sum_{j=1}^{j=4} N_j}(\sum_{j=1}^{j=4}\sum_{i=1}^{i=N} Y_{ij,T=1,M_{T=0}} - Y_{ij,T=0,M_{T=0}}),$$

where $N = 250{,}000$.

Results. As noted above, our mediation procedure yielded AME and ADE estimates for the probability of selecting a given EBEP response level for each EBEP item and a given EBEP response level marginalized across EBEP items. Here, we present all of these estimates, as well as summary estimates that represent the AMEs and ADEs for responses ≥ "slightly justified".

Table 5

Study 3 Mediation Estimates for PMW for EBEP ≥ "Slightly justified"

EBEP	AME	ADE	Total
Facebook	**0.03** [0, 0.24]	0.01 [-0.03, 0.08]	**0.04** [0, 0.28]
Flyer	**0.03** [0, 0.2]	0.01 [-0.02, 0.11]	**0.04** [0, 0.26]
Yell	**0.03** [0, 0.19]	0.01 [-0.02, 0.07]	**0.03** [0, 0.23]
Assault	**0.03** [0, 0.21]	0.01 [-0.02, 0.11]	**0.04** [0, 0.25]
Marginal	**0.03** [0, 0.2]	0.01 [-0.02, 0.08]	**0.04** [0, 0.24]

NOTE: Cell values represent posterior means and 95% CIs. Bold entries indicate that the CI does not overlap with zero. Because EBEP is ordinal, mediation is estimated at each level of the variable. Here, however, results are summarized to reflect the indirect, direct, and total effects on the probability of an EBEP being rated as at least 'Slightly Justified' marginalized across EBEP items.

Table 6

Study 3 Mediation Estimates for PMW for EBEP ≥ "Slightly justified" Adjusted for Ideology

EBEP	AME	ADE	Total
Facebook	**0.02** [0, 0.19]	0.02 [-0.01, 0.15]	**0.04** [0, 0.32]
Flyer	**0.03** [0, 0.2]	0.01 [-0.02, 0.13]	**0.04** [0, 0.32]
Yell	**0.02** [0, 0.17]	0.01 [-0.01, 0.11]	**0.03** [0, 0.23]
Assault	**0.02** [0, 0.15]	0.01 [-0.01, 0.12]	**0.03** [0, 0.23]
Marginal	**0.02** [0, 0.18]	0.01 [-0.01, 0.11]	**0.03** [0, 0.26]

NOTE: Cell values represent posterior means and 95% CIs. Bold entries indicate that the CI does not overlap with zero. Because EBEP is ordinal, mediation is estimated at each level of the variable. Here, however, results are summarized to reflect the indirect, direct, and total effects on the probability of an EBEP being rated as at least 'Slightly Justified' marginalized across EBEP items. All effects were estimated with standardized political ideology set to its mean.

Study 4

Here we report model estimates for each of the three regression models estimated for Study 4 (See main text for discussion). Specifically, we report mean posterior estimates with Highest Posterior Density Intervals for the fixed and random effects estimated via these models (See Table 9). We also report estimates from models that adjust for individual-level political ideology (M = 2.28, SD = 1.66; See Table 10). Note, for these models, N = 313 as 11 participants indicated that they either did not know or have a political ideology.

Mediation Results

To investigate whether perceived moral wrongdoing statistically mediated the effect of the Binding values, we used the same approach as in Study 3. Specifically, we

Table 7

Study 3 Mediation Estimates for PMW at Each Response Level

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.185** [-0.37, -0.01]	-0.045 [-0.29, 0.15]	-0.23 [-0.56, 0]
	2	0.072 [-0.19, 0.23]	0.021 [-0.1, 0.18]	0.092 [-0.24, 0.34]
	3	0.041 [-0.08, 0.13]	0.009 [-0.06, 0.09]	0.049 [-0.1, 0.18]
	4	0.041 [-0.05, 0.17]	0.008 [-0.05, 0.09]	0.049 [-0.07, 0.22]
	5	**0.019** [0, 0.14]	0.003 [-0.02, 0.04]	**0.023** [0, 0.17]
	6	**0.01** [0, 0.11]	0.003 [-0.01, 0.02]	**0.014** [0, 0.13]
	7	**0.002** [0, 0.01]	0.001 [0, 0]	**0.003** [0, 0.02]
Flyer	1	**-0.185** [-0.37, -0.01]	-0.047 [-0.29, 0.12]	**-0.232** [-0.55, 0]
	2	0.07 [-0.19, 0.23]	0.018 [-0.13, 0.16]	0.088 [-0.25, 0.33]
	3	0.042 [-0.07, 0.12]	0.01 [-0.05, 0.09]	0.052 [-0.09, 0.18]
	4	0.044 [-0.01, 0.17]	0.01 [-0.05, 0.12]	0.054 [-0.01, 0.24]
	5	**0.018** [0, 0.13]	0.004 [-0.02, 0.06]	**0.022** [0, 0.16]
	6	**0.008** [0, 0.07]	0.004 [0, 0.04]	**0.011** [0, 0.1]
	7	**0.003** [0, 0.01]	0.001 [0, 0]	**0.004** [0, 0.02]
Yell	1	**-0.18** [-0.36, -0.01]	-0.046 [-0.29, 0.15]	**-0.225** [-0.54, 0]
	2	0.067 [-0.19, 0.22]	0.016 [-0.12, 0.15]	0.082 [-0.25, 0.31]
	3	0.042 [-0.07, 0.13]	0.011 [-0.04, 0.1]	0.053 [-0.08, 0.18]
	4	**0.044** [0, 0.17]	0.012 [-0.04, 0.11]	0.056 [-0.01, 0.24]
	5	**0.018** [0, 0.12]	0.005 [-0.01, 0.05]	**0.023** [0, 0.15]
	6	**0.007** [0, 0.06]	0.002 [0, 0.02]	**0.01** [0, 0.07]
	7	**0.002** [0, 0.01]	0 [0, 0]	**0.002** [0, 0.01]
Assault	1	**-0.185** [-0.37, -0.01]	-0.049 [-0.31, 0.16]	**-0.234** [-0.58, 0]
	2	0.063 [-0.18, 0.23]	0.013 [-0.17, 0.16]	0.076 [-0.25, 0.34]
	3	0.046 [-0.07, 0.13]	0.013 [-0.06, 0.09]	0.059 [-0.08, 0.19]
	4	0.046 [-0.04, 0.17]	0.013 [-0.04, 0.11]	0.06 [-0.03, 0.24]
	5	**0.019** [0, 0.11]	0.006 [-0.01, 0.08]	**0.025** [0, 0.15]
	6	**0.009** [0, 0.09]	0.003 [0, 0.03]	**0.012** [0, 0.11]
	7	**0.002** [0, 0.01]	0 [0, 0]	**0.002** [0, 0.02]

Table 8

Study 3 Mediation Estimates for PMW at Each Response Level Adjusted for Ideology

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.137** [-0.3, 0]	-0.084 [-0.38, 0.1]	-0.221 [-0.57, 0]
	2	0.057 [-0.15, 0.19]	0.033 [-0.17, 0.22]	0.09 [-0.26, 0.34]
	3	0.028 [-0.07, 0.1]	0.017 [-0.05, 0.11]	0.045 [-0.11, 0.18]
	4	0.029 [-0.06, 0.13]	0.017 [-0.06, 0.14]	0.046 [-0.1, 0.25]
	5	**0.014** [0, 0.11]	0.009 [-0.02, 0.09]	**0.022** [0, 0.19]
	6	**0.008** [0, 0.09]	0.007 [0, 0.06]	**0.015** [0, 0.17]
	7	**0.001** [0, 0.01]	0.001 [0, 0.01]	**0.002** [0, 0.03]
Flyer	1	**-0.139** [-0.3, 0]	-0.073 [-0.36, 0.16]	-0.212 [-0.54, 0]
	2	0.051 [-0.15, 0.19]	0.028 [-0.13, 0.19]	0.08 [-0.24, 0.33]
	3	0.031 [-0.07, 0.11]	0.016 [-0.06, 0.11]	0.047 [-0.11, 0.19]
	4	0.031 [-0.08, 0.14]	0.016 [-0.07, 0.12]	0.047 [-0.11, 0.22]
	5	**0.014** [0, 0.11]	0.007 [-0.02, 0.08]	**0.021** [0, 0.17]
	6	**0.009** [0, 0.11]	0.004 [0, 0.05]	**0.013** [0, 0.15]
	7	**0.002** [0, 0.02]	0.002 [0, 0.01]	**0.004** [0, 0.02]
Yell	1	**-0.145** [-0.3, -0.01]	-0.079 [-0.32, 0.1]	**-0.224** [-0.52, 0]
	2	0.057 [-0.15, 0.19]	0.029 [-0.14, 0.19]	0.086 [-0.26, 0.34]
	3	0.033 [-0.06, 0.1]	0.018 [-0.04, 0.11]	0.051 [-0.08, 0.18]
	4	0.033 [-0.03, 0.14]	0.018 [-0.03, 0.13]	0.051 [-0.05, 0.23]
	5	**0.014** [0, 0.1]	0.008 [-0.01, 0.07]	**0.022** [0, 0.14]
	6	**0.007** [0, 0.06]	0.004 [0, 0.03]	**0.011** [0, 0.09]
	7	**0.001** [0, 0.01]	0.001 [0, 0]	**0.002** [0, 0.01]
Assault	1	**-0.14** [-0.3, -0.01]	-0.077 [-0.33, 0.09]	**-0.216** [-0.54, 0]
	2	0.054 [-0.15, 0.18]	0.032 [-0.11, 0.18]	0.086 [-0.22, 0.32]
	3	0.034 [-0.05, 0.1]	0.017 [-0.05, 0.11]	0.051 [-0.08, 0.18]
	4	0.033 [-0.01, 0.14]	0.017 [-0.03, 0.13]	0.05 [-0.01, 0.22]
	5	**0.013** [0, 0.09]	0.007 [-0.01, 0.08]	**0.02** [0, 0.15]
	6	**0.005** [0, 0.05]	0.003 [0, 0.04]	**0.008** [0, 0.08]
	7	**0.001** [0, 0.01]	0.001 [0, 0]	**0.002** [0, 0.01]

NOTE: Cell values represent posterior means and 95% CIs. Bold entries indicate that the CI does not overlap with zero. All effects were estimated with standardized political ideology set to its mean.

estimated AMEs and ADEs for PMW (M) using standardized Binding values as the exogenous treatment variable and perceived EBEP justification as the outcome variable. To evaluate the effects of standardized Binding values, we focus on expected changes in PMW and EBEP justification given a change from 0 to 1 in standardized Binding values (i.e. the difference between being at the mean of Binding values vs being one standard deviation above the mean). All procedural steps were otherwise identical to those used for the mediation analysis reported for Study 3.

Study 5

Here we report model estimates for each of the three regression models estimated for Study 5 (See main text for discussion). Specifically, we report mean posterior estimates with Highest Posterior Density Intervals for the fixed and random effects estimated via these models (See Table 15). We also report estimates from models that adjust for individual-level political ideology (M = 2.97, SD = 2.01; See Table 16). Note, for these models, N = 508 as 3 participants indicated that they either did not know or have a political ideology.

Mediation Results

To investigate whether perceived moral wrongdoing statistically mediated the effect of the Binding values, we used the same approach as in Study 3. Specifically, we estimated AMEs and ADEs for PMW (M) using standardized Binding values as the exogenous treatment variable and perceived EBEP justification as the outcome variable. To evaluate the effects of standardized Binding values, we focus on expected changes in PMW and EBEP justification given a change from 0 to 1 in standardized Binding values (i.e. the difference between being at the mean of Binding values vs being one standard deviation above the mean). All procedural steps were otherwise identical to those used for the mediation analysis reported for Study 3.

Table 9

Study 4 Model Estimates

Fixed Effects		Model 1	Model 2	Model 3
		PMW Std.	EBEP	EBEP
	Intercept	0.00		
		[−0.09; 0.09]		
	I Values Std.	−0.16*	−1.15*	−0.87*
		[−0.26; −0.06]	[−1.91; −0.40]	[−1.70; −0.03]
	B Values Std.	0.47*	1.60*	0.73*
		[0.37; 0.56]	[1.10; 2.11]	[0.21; 1.19]
	PMW Std.			1.63*
				[0.74; 2.51]
	Intercept[1]		0.20	0.19
			[−3.89; 3.93]	[−3.69; 3.98]
	Intercept[2]		2.34	2.35
			[−1.82; 6.03]	[−1.79; 5.89]
	Intercept[3]		3.45	3.47
			[−0.59; 7.26]	[−0.73; 6.96]
	Intercept[4]		4.70*	4.73*
			[0.60; 8.51]	[0.54; 8.22]
	Intercept[5]		6.42*	6.51*
			[2.16; 10.12]	[2.40; 10.13]
	Intercept[6]		8.13*	8.28*
			[3.95; 11.96]	[4.17; 11.97]
SD of Random Effects				
	$\text{Intercept}_{Subject}$		2.83*	2.39*
			[2.43; 3.24]	[2.01; 2.75]
	Intercept_{EBEP}		4.12*	3.94*
			[1.43; 8.01]	[1.41; 7.78]
	$\text{I Values Std}_{EBEP}$		0.57*	0.63*
			[0.03; 1.45]	[0.01; 1.61]
	$\text{B Values Std}_{EBEP}$		0.21*	0.23*
			[0.00; 0.69]	[0.00; 0.75]
	PMW Std_{EBEP}			0.61*
				[0.00; 1.70]

* 0 outside 95% highest posterior density interval. Estimates for Models 2 and 3 reported on log scale.

Table 10

Study 4 Model Estimates Adjusted for Ideology

Fixed Effects		Model 1 PMW Std.	Model 2 EBEP	Model 3 EBEP
	Intercept	−0.01		
		[−0.10; 0.09]		
	I Values Std.	−0.07	−0.68	−0.59
		[−0.18; 0.04]	[−1.61; 0.33]	[−1.55; 0.39]
	B Values Std.	0.35*	1.08*	0.47
		[0.23; 0.46]	[0.47; 1.69]	[−0.15; 1.03]
	PMW Std.			1.53*
				[0.44; 2.57]
	Ideology Std.	0.22*	1.13*	0.78*
		[0.10; 0.35]	[0.53; 1.73]	[0.18; 1.42]
	Intercept[1]		0.13	0.19
			[−4.13; 3.73]	[−4.12; 3.65]
	Intercept[2]		2.33	2.41
			[−2.04; 5.84]	[−1.66; 6.09]
	Intercept[3]		3.50	3.59
			[−0.73; 7.17]	[−0.44; 7.33]
	Intercept[4]		4.76*	4.87*
			[0.44; 8.36]	[0.78; 8.57]
	Intercept[5]		6.54*	6.72*
			[2.26; 10.25]	[2.48; 10.33]
	Intercept[6]		8.47*	8.74*
			[4.22; 12.25]	[4.50; 12.41]
SD of Random Effects				
	Intercept_{EBEP}		4.25*	4.03*
			[1.48; 8.18]	[1.53; 7.93]
	$\text{Intercept}_{Subject}$		2.83*	2.44*
			[2.42; 3.26]	[2.06; 2.81]
	$\text{I Values Std}_{EBEP}$		0.69*	0.72*
			[0.01; 1.76]	[0.03; 1.83]
	$\text{B Values Std}_{EBEP}$		0.27*	0.28*
			[0.00; 0.84]	[0.00; 0.89]
	PMW Std_{EBEP}			0.80*
				[0.00; 2.10]
	$\text{Ideology Std}_{EBEP}$		0.28*	0.34*
			[0.00; 0.87]	[0.00; 1.03]

* 0 outside 95% highest posterior density interval. Estimates for Models 2 and 3 reported on log scale. Ideology is coded such that higher values indicate stronger associations with Conservative ideology.

Table 11

Study 4 Mediation Estimates for PMW for EBEP ≥ "Slightly justified"

EBEP	AME	ADE	Total
Facebook	**0.03** [0, 0.19]	**0.02** [0, 0.18]	**0.05** [0, 0.37]
Flyer	**0.03** [0, 0.18]	**0.02** [0, 0.18]	**0.05** [0, 0.34]
Yell	**0.03** [0, 0.19]	**0.03** [0, 0.17]	**0.06** [0, 0.36]
Assault	**0.03** [0, 0.18]	**0.03** [0, 0.15]	**0.05** [0, 0.31]
Assault	**0.03** [0, 0.18]	**0.03** [0, 0.17]	**0.05** [0, 0.34]

Table 12

Study 4 Mediation Estimates for PMW for EBEP ≥ "Slightly justified" Adjusted for Ideology

EBEP	AME	ADE	Total
Facebook	**0.02** [0, 0.13]	0.02 [0, 0.14]	**0.03** [0, 0.26]
Flyer	**0.02** [0, 0.13]	0.02 [0, 0.12]	**0.03** [0, 0.24]
Yell	**0.02** [0, 0.11]	0.01 [0, 0.09]	**0.03** [0, 0.17]
Assault	**0.01** [0, 0.1]	0.01 [0, 0.09]	**0.03** [0, 0.18]
Marginal	**0.02** [0, 0.11]	0.01 [0, 0.1]	**0.03** [0, 0.19]

Table 13

Study 4 Mediation Estimates for PMW at Each Response Level

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.122** [-0.25, 0]	**-0.114** [-0.26, 0]	**-0.237** [-0.45, -0.01]
	2	0.046 [-0.15, 0.16]	0.044 [-0.13, 0.17]	0.089 [-0.27, 0.3]
	3	0.028 [-0.07, 0.1]	0.026 [-0.06, 0.09]	0.055 [-0.13, 0.17]
	4	0.022 [-0.04, 0.1]	0.02 [-0.04, 0.09]	0.042 [-0.1, 0.17]
	5	**0.017** [0, 0.11]	**0.015** [0, 0.1]	**0.032** [0, 0.21]
	6	**0.006** [0, 0.06]	**0.006** [0, 0.07]	**0.013** [0, 0.13]
	7	**0.003** [0, 0.03]	**0.003** [0, 0.04]	**0.007** [0, 0.08]
Flyer	1	**-0.123** [-0.25, 0]	**-0.12** [-0.25, 0]	**-0.243** [-0.45, -0.01]
	2	0.043 [-0.16, 0.17]	0.043 [-0.14, 0.17]	0.086 [-0.28, 0.31]
	3	0.029 [-0.07, 0.09]	0.029 [-0.06, 0.1]	0.058 [-0.13, 0.18]
	4	0.024 [-0.03, 0.11]	0.023 [-0.03, 0.1]	0.048 [-0.07, 0.19]
	5	**0.018** [0, 0.11]	**0.017** [0, 0.11]	**0.035** [0, 0.21]
	6	**0.006** [0, 0.05]	**0.005** [0, 0.05]	**0.011** [0, 0.11]
	7	**0.003** [0, 0.02]	**0.002** [0, 0.02]	**0.005** [0, 0.04]
Yell	1	**-0.121** [-0.25, 0]	**-0.113** [-0.25, 0]	**-0.233** [-0.44, 0]
	2	0.037 [-0.15, 0.17]	0.036 [-0.16, 0.16]	0.073 [-0.29, 0.3]
	3	0.029 [-0.07, 0.1]	0.027 [-0.06, 0.1]	0.056 [-0.13, 0.18]
	4	0.024 [-0.06, 0.1]	0.022 [-0.06, 0.1]	0.046 [-0.12, 0.2]
	5	0.018 [0, 0.12]	0.017 [-0.01, 0.12]	**0.035** [0, 0.22]
	6	**0.007** [0, 0.08]	**0.006** [0, 0.08]	**0.014** [0, 0.15]
	7	**0.004** [0, 0.05]	**0.004** [0, 0.04]	**0.009** [0, 0.08]
Assault	1	**-0.119** [-0.24, 0]	**-0.115** [-0.25, 0]	**-0.235** [-0.45, -0.01]
	2	0.037 [-0.16, 0.16]	0.036 [-0.17, 0.17]	0.073 [-0.3, 0.3]
	3	0.029 [-0.07, 0.1]	0.029 [-0.05, 0.1]	0.058 [-0.11, 0.19]
	4	0.025 [-0.05, 0.11]	0.025 [-0.04, 0.11]	0.05 [-0.08, 0.2]
	5	**0.018** [0, 0.12]	**0.017** [0, 0.1]	**0.035** [0, 0.2]
	6	**0.007** [0, 0.06]	**0.006** [0, 0.05]	**0.013** [0, 0.11]
	7	**0.003** [0, 0.02]	**0.003** [0, 0.02]	**0.006** [0, 0.06]

Table 14

Study 4 Mediation Estimates for PMW at Each Response Level Adjusted for Ideology

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.085** [-0.2, 0]	-0.077 [-0.22, 0.01]	**-0.162** [-0.36, -0.01]
	2	0.033 [-0.11, 0.13]	0.03 [-0.11, 0.14]	0.063 [-0.19, 0.25]
	3	0.021 [-0.05, 0.08]	0.018 [-0.05, 0.08]	0.039 [-0.1, 0.14]
	4	0.015 [-0.03, 0.07]	0.013 [-0.02, 0.07]	0.028 [-0.06, 0.13]
	5	0.011 [0, 0.08]	0.01 [0, 0.09]	**0.021** [0, 0.16]
	6	**0.004** [0, 0.05]	0.004 [0, 0.04]	**0.008** [0, 0.1]
	7	**0.002** [0, 0.01]	0.001 [0, 0.01]	**0.003** [0, 0.02]
Flyer	1	**-0.084** [-0.2, 0]	-0.076 [-0.21, 0.01]	**-0.161** [-0.36, -0.01]
	2	0.032 [-0.11, 0.13]	0.029 [-0.11, 0.14]	0.062 [-0.2, 0.24]
	3	0.02 [-0.05, 0.08]	0.018 [-0.05, 0.08]	0.038 [-0.1, 0.14]
	4	0.015 [-0.04, 0.07]	0.014 [-0.02, 0.07]	0.028 [-0.05, 0.14]
	5	**0.011** [0, 0.08]	0.01 [0, 0.08]	**0.021** [0, 0.14]
	6	**0.004** [0, 0.05]	0.004 [0, 0.04]	**0.008** [0, 0.08]
	7	**0.002** [0, 0.02]	0.001 [0, 0.01]	**0.003** [0, 0.03]
Yell	1	**-0.082** [-0.19, 0]	-0.075 [-0.21, 0.02]	**-0.157** [-0.34, 0]
	2	0.029 [-0.12, 0.13]	0.028 [-0.1, 0.14]	0.057 [-0.19, 0.23]
	3	0.021 [-0.04, 0.07]	0.02 [-0.03, 0.09]	0.042 [-0.07, 0.14]
	4	0.017 [-0.01, 0.08]	0.015 [-0.01, 0.07]	0.031 [-0.01, 0.13]
	5	0.01 [0, 0.08]	0.008 [-0.01, 0.06]	**0.019** [0, 0.12]
	6	**0.003** [0, 0.03]	0.003 [0, 0.02]	**0.006** [0, 0.04]
	7	**0.002** [0, 0.01]	0.001 [0, 0]	**0.003** [0, 0.01]
Assault	1	**-0.084** [-0.19, 0]	-0.074 [-0.2, 0]	**-0.159** [-0.34, -0.01]
	2	0.03 [-0.11, 0.13]	0.026 [-0.12, 0.13]	0.056 [-0.2, 0.23]
	3	0.023 [-0.04, 0.08]	0.02 [-0.04, 0.08]	0.043 [-0.07, 0.14]
	4	0.017 [-0.02, 0.08]	0.016 [-0.01, 0.08]	0.032 [-0.01, 0.14]
	5	0.01 [0, 0.07]	0.009 [0, 0.07]	**0.019** [0, 0.12]
	6	**0.003** [0, 0.03]	0.002 [0, 0.02]	**0.006** [0, 0.04]
	7	**0.001** [0, 0.01]	0.001 [0, 0.01]	**0.002** [0, 0.01]

Table 15

Study 5 Model Estimates

Fixed Effects		Model 1	Model 2	Model 3
		PMW Std.	EBEP	EBEP
	Intercept	0.00		
		[−0.08; 0.08]		
	I Values Std.	−0.25*	−1.70*	−1.29*
		[−0.34; −0.17]	[−2.33; −1.11]	[−1.86; −0.66]
	B Values Std.	0.44*	2.29*	1.48*
		[0.35; 0.52]	[1.65; 2.94]	[0.81; 2.16]
	PMW Std.			1.72*
				[0.68; 2.67]
	Intercept[1]		1.40	1.57
			[−2.18; 4.31]	[−1.62; 4.10]
	Intercept[2]		3.26	3.44*
			[−0.46; 6.09]	[0.27; 6.00]
	Intercept[3]		4.21*	4.42*
			[0.61; 7.18]	[1.32; 7.06]
	Intercept[4]		5.80*	6.07*
			[2.06; 8.64]	[2.90; 8.65]
	Intercept[5]		6.86*	7.16*
			[3.25; 9.82]	[3.96; 9.77]
	Intercept[6]		8.62*	8.97*
			[4.87; 11.58]	[5.88; 11.74]
SD of Random Effects				
	$\text{Intercept}_{Subject}$		4.41*	4.16*
			[3.90; 4.96]	[3.66; 4.65]
	Intercept_{EBEP}		3.00*	2.51*
			[0.84; 6.36]	[0.76; 5.39]
	$\text{I Values Std}_{EBEP}$		0.27*	0.31*
			[0.00; 0.77]	[0.00; 0.89]
	$\text{B Values Std}_{EBEP}$		0.20*	0.27*
			[0.00; 0.63]	[0.00; 0.84]
	PMW Std_{EBEP}			0.70*
				[0.10; 1.77]

* 0 outside 95% highest posterior density interval

Table 16

Study 5 Model Estimates Adjusted for Ideology

Fixed Effects		Model 1 PMW Std.	Model 2 EBEP	Model 3 EBEP
	Intercept	0.00		
		[−0.07; 0.07]		
	I Values Std.	−0.09*	−1.22*	−1.08*
		[−0.17; −0.00]	[−1.95; −0.43]	[−1.85; −0.29]
	B Values Std.	0.25*	1.70*	1.31*
		[0.16; 0.34]	[1.00; 2.50]	[0.58; 2.09]
	PMW Std.			1.48*
				[0.60; 2.44]
	Ideology Std.	0.41*	1.29*	0.67
		[0.33; 0.50]	[0.31; 2.20]	[−0.14; 1.43]
	Intercept[1]		1.61	1.62
			[−1.75; 4.18]	[−1.43; 4.36]
	Intercept[2]		3.49*	3.51*
			[0.30; 6.26]	[0.40; 6.20]
	Intercept[3]		4.47*	4.50*
			[1.12; 7.11]	[1.27; 7.11]
	Intercept[4]		6.11*	6.17*
			[2.81; 8.79]	[2.90; 8.74]
	Intercept[5]		7.19*	7.27*
			[3.90; 9.88]	[4.07; 9.96]
	Intercept[6]		8.96*	9.08*
			[5.73; 11.77]	[5.99; 11.92]
SD of Random Effects				
	$\text{Intercept}_{Subject}$		4.38*	4.19*
			[3.85; 4.91]	[3.70; 4.70]
	Intercept_{EBEP}		2.65*	2.52*
			[0.79; 5.63]	[0.70; 5.32]
	$\text{I Values Std}_{EBEP}$		0.43*	0.43*
			[0.00; 1.15]	[0.00; 1.22]
	$\text{B Values Std}_{EBEP}$		0.31*	0.36*
			[0.00; 0.96]	[0.00; 1.08]
	PMW Std_{EBEP}			0.59*
				[0.00; 1.61]
	$\text{Ideology Std}_{EBEP}$		0.63*	0.48*
			[0.09; 1.62]	[0.00; 1.27]

* 0 outside 95% highest posterior density interval

Table 17

Study 5 Mediation Estimates for PMW for EBEP ≥ "Slightly justified"

EBEP	AME	ADE	Total
Facebook	**0.01** [0, 0.07]	**0.02** [0, 0.13]	**0.03** [0, 0.2]
Flyer	**0.01** [0, 0.07]	**0.02** [0, 0.15]	**0.03** [0, 0.25]
Yell	**0.01** [0, 0.05]	**0.02** [0, 0.15]	**0.03** [0, 0.22]
Assault	**0.01** [0, 0.06]	**0.02** [0, 0.13]	**0.03** [0, 0.21]
Marginal	**0.01** [0, 0.06]	**0.02** [0, 0.15]	**0.03** [0, 0.23]

Table 18

Study 5 Mediation Estimates for PMW for EBEP ≥ "Slightly justified" Adjusted for Ideology

EBEP	AME	ADE	Total
Facebook	**0.003** [0, 0.026]	**0.018** [0, 0.159]	**0.021** [0.001, 0.174]
Flyer	**0.003** [0, 0.029]	**0.017** [0, 0.137]	**0.02** [0.001, 0.161]
Yell	**0.003** [0, 0.02]	**0.015** [0, 0.122]	**0.018** [0, 0.146]
Assault	**0.003** [0, 0.015]	**0.014** [0, 0.115]	**0.016** [0, 0.126]
Marginal	**0.003** [0, 0.02]	**0.015** [0, 0.123]	**0.018** [0, 0.147]

Table 19

Study 5 Mediation Estimates for PMW at Each Response Level

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.112** [-0.26, -0.01]	**-0.246** [-0.45, -0.04]	**-0.358** [-0.64, -0.05]
	2	0.066 [-0.09, 0.16]	0.133 [-0.18, 0.27]	0.2 [-0.27, 0.38]
	3	**0.022** [0, 0.08]	**0.051** [0, 0.15]	**0.073** [0, 0.2]
	4	**0.017** [0, 0.1]	**0.042** [0, 0.19]	**0.059** [0, 0.27]
	5	**0.004** [0, 0.03]	**0.01** [0, 0.07]	**0.014** [0, 0.11]
	6	**0.002** [0, 0.02]	**0.006** [0, 0.04]	**0.008** [0, 0.06]
	7	**0.001** [0, 0]	**0.003** [0, 0.01]	**0.005** [0, 0.01]
Flyer	1	**-0.116** [-0.25, -0.01]	**-0.252** [-0.44, -0.04]	**-0.368** [-0.63, -0.08]
	2	0.064 [-0.12, 0.16]	0.133 [-0.19, 0.27]	0.198 [-0.29, 0.39]
	3	0.023 [0, 0.08]	0.052 [-0.01, 0.13]	**0.075** [0, 0.2]
	4	**0.02** [0, 0.11]	**0.045** [0, 0.21]	**0.065** [0, 0.3]
	5	**0.005** [0, 0.04]	**0.012** [0, 0.09]	**0.017** [0, 0.14]
	6	**0.003** [0, 0.02]	**0.007** [0, 0.05]	**0.01** [0, 0.08]
	7	**0.001** [0, 0.01]	**0.002** [0, 0.01]	**0.003** [0, 0.02]
Yell	1	**-0.111** [-0.26, -0.01]	**-0.241** [-0.45, -0.04]	**-0.353** [-0.62, -0.06]
	2	0.064 [-0.08, 0.16]	0.131 [-0.19, 0.28]	0.195 [-0.27, 0.39]
	3	**0.022** [0, 0.08]	**0.049** [0, 0.13]	**0.071** [0, 0.2]
	4	**0.018** [0, 0.1]	**0.042** [0, 0.19]	**0.06** [0, 0.28]
	5	**0.005** [0, 0.03]	**0.011** [0, 0.08]	**0.016** [0, 0.13]
	6	**0.002** [0, 0.01]	**0.007** [0, 0.05]	**0.009** [0, 0.06]
	7	**0** [0, 0]	**0.002** [0, 0.01]	**0.002** [0, 0.01]
Assault	1	**-0.115** [-0.25, -0.01]	**-0.252** [-0.46, -0.04]	**-0.366** [-0.63, -0.07]
	2	0.067 [-0.1, 0.15]	0.138 [-0.16, 0.27]	0.205 [-0.26, 0.38]
	3	**0.022** [0, 0.08]	0.051 [0, 0.14]	**0.073** [0, 0.2]
	4	**0.018** [0, 0.1]	**0.043** [0, 0.2]	**0.061** [0, 0.28]
	5	**0.004** [0, 0.04]	**0.01** [0, 0.07]	**0.015** [0, 0.11]
	6	**0.002** [0, 0.02]	**0.006** [0, 0.04]	**0.009** [0, 0.07]
	7	**0.001** [0, 0]	**0.002** [0, 0.01]	**0.003** [0, 0.02]

Table 20

Study 5 Mediation Estimates for PMW at Each Response Level Adjusted for Ideology

EBEP	Response Level	AME	ADE	Total
Facebook	1	**-0.051** [-0.13, 0]	**-0.209** [-0.42, -0.02]	**-0.259** [-0.49, -0.03]
	2	0.032 [-0.04, 0.08]	0.117 [-0.16, 0.27]	0.149 [-0.21, 0.31]
	3	0.009 [0, 0.04]	0.041 [-0.03, 0.13]	0.05 [-0.01, 0.16]
	4	**0.007** [0, 0.04]	**0.033** [0, 0.17]	**0.04** [0, 0.21]
	5	**0.002** [0, 0.01]	**0.009** [0, 0.08]	**0.011** [0, 0.09]
	6	**0.001** [0, 0.01]	**0.006** [0, 0.06]	**0.007** [0, 0.06]
	7	**0** [0, 0]	**0.002** [0, 0.01]	**0.003** [0, 0.02]
Flyer	1	**-0.051** [-0.13, 0]	**-0.206** [-0.41, -0.03]	**-0.258** [-0.5, -0.04]
	2	0.031 [-0.04, 0.08]	0.116 [-0.18, 0.25]	0.147 [-0.2, 0.3]
	3	0.009 [0, 0.04]	0.04 [-0.02, 0.13]	0.049 [-0.01, 0.16]
	4	**0.007** [0, 0.05]	**0.034** [0, 0.18]	**0.042** [0, 0.23]
	5	**0.002** [0, 0.02]	**0.009** [0, 0.08]	**0.011** [0, 0.1]
	6	**0.001** [0, 0.01]	**0.005** [0, 0.04]	**0.006** [0, 0.05]
	7	**0** [0, 0]	**0.002** [0, 0.01]	**0.002** [0, 0.01]
Yell	1	**-0.051** [-0.13, 0]	**-0.212** [-0.43, -0.03]	**-0.263** [-0.53, -0.04]
	2	0.031 [-0.04, 0.08]	0.115 [-0.18, 0.27]	0.145 [-0.21, 0.31]
	3	**0.01** [0, 0.04]	**0.044** [0, 0.14]	**0.054** [0, 0.16]
	4	**0.008** [0, 0.05]	**0.038** [0, 0.19]	**0.046** [0, 0.23]
	5	**0.002** [0, 0.01]	**0.009** [0, 0.07]	**0.011** [0, 0.09]
	6	**0.001** [0, 0.01]	**0.005** [0, 0.04]	**0.006** [0, 0.05]
	7	**0** [0, 0]	**0.001** [0, 0.01]	**0.001** [0, 0.01]
Assault	1	**-0.051** [-0.12, 0]	**-0.208** [-0.41, -0.02]	**-0.26** [-0.5, -0.03]
	2	0.034 [-0.02, 0.08]	0.123 [-0.15, 0.26]	0.156 [-0.15, 0.31]
	3	**0.009** [0, 0.04]	**0.041** [0, 0.13]	**0.05** [0, 0.15]
	4	**0.006** [0, 0.04]	**0.031** [0, 0.17]	**0.037** [0, 0.2]
	5	**0.001** [0, 0.01]	**0.007** [0, 0.06]	**0.008** [0, 0.07]
	6	**0.001** [0, 0]	**0.005** [0, 0.03]	**0.006** [0, 0.04]
	7	**0** [0, 0]	**0.002** [0, 0.01]	**0.002** [0, 0.01]